What's New in Econometrics **NBER, Summer 2007**

Lecture 1, Monday, July 30th, 9.00-10.30am

Estimation of Average Treatment Effects Under Unconfoundedness

1. Introduction

In this lecture we look at several methods for estimating average effects of a program, treatment, or regime, under unconfoundedness. The setting is one with a binary program. The traditional example in economics is that of a labor market program where some individuals receive training and others do not, and interest is in some measure of the effectiveness of the training. Unconfoundedness, a term coined by Rubin (1990), refers to the case where (non-parametrically) adjusting for differences in a fixed set of covariates removes biases in comparisons between treated and control units, thus allowing for a causal interpretation of those adjusted differences. This is perhaps the most important special case for estimating average treatment effects in practice. Alternatives typically involves strong assumptions linking unobservables to observables in specific ways in order to allow adjusting for the relevant differences in unobserved variables. An example of such a strategy is instrumental variables, which will be discussed in Lecture 3. A second example that does not involve additional assumptions is the bounds approach developed by Manski (1990, 2003).

Under the specific assumptions we make in this setting, the population average treatment effect can be estimated at the standard parametric $\sqrt{N}$ rate without functional form assumptions. A variety of estimators, at first sight quite different, have been proposed for implementing this. The estimators include regression estimators, propensity score based estimators and matching estimators. Many of these are used in practice, although rarely is this choice motivated by principled arguments. In practice the differences between the estimators are relatively minor when applied appropriately, although matching in combination with regression is generally more robust and is probably the recommended choice. More important than the choice of estimator are two other issues. Both involve analyses of the data without the outcome variable. First, one should carefully check the extent of the overlap

in covariate distributions between the treatment and control groups. Often there is a need for some trimming based on the covariate values if the original sample is not well balanced. Without this, estimates of average treatment effects can be very sensitive to the choice of, and small changes in the implementation of, the estimators. In this part of the analysis the propensity score plays an important role. Second, it is useful to do some assessment of the appropriateness of the unconfoundedness assumption. Although this assumption is not directly testable, its plausibility can often be assessed using lagged values of the outcome as pseudo outcomes. Another issue is variance estimation. For matching estimators bootstrapping, although widely used, has been shown to be invalid. We discuss general methods for estimating the conditional variance that do not involve resampling.

In these notes we first set up the basic framework and state the critical assumptions in Section 2. In Section 3 we describe the leading estimators. In Section 4 we discuss variance estimation. In Section 5 we discuss assessing one of the critical assumptions, unconfoundedness. In Section 6 we discuss dealing with a major problem in practice, lack of overlap in the covariate distributions among treated and controls. In Section 7 we illustrate some of the methods using a well known data set in this literature, originally put together by Lalonde (1986).

In these notes we focus on estimation and inference for treatment effects. We do not discuss here a recent literature that has taken the next logical step in the evaluation literature, namely the optimal assignment of individuals to treatments based on limited (sample) information regarding the efficacy of the treatments. See Manski (2004, 2005, Dehejia (2004), Hirano and Porter (2005).

2. Framework

The modern set up in this literature is based on the potential outcome approach developed by Rubin (1974, 1977, 1978), which view causal effects as comparisons of potential outcomes defined on the same unit. In this section we lay out the basic framework.

2.1 Definitions

We observe N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population. We postulate the existence for each unit of a pair of potential outcomes, $Y_i(0)$ for the outcome under the control treatment and $Y_i(1)$ for the outcome under the active treatment. In addition, each unit has a vector of characteristics, referred to as covariates, pretreatment variables or exogenous variables, and denoted by X_i.[1] It is important that these variables are not affected by the treatment. Often they take their values prior to the unit being exposed to the treatment, although this is not sufficient for the conditions they need to satisfy. Importantly, this vector of covariates can include lagged outcomes. Finally, each unit is exposed to a single treatment; $W_i = 0$ if unit i receives the control treatment and $W_i = 1$ if unit i receives the active treatment. We therefore observe for each unit the triple (W_i, Y_i, X_i), where Y_i is the realized outcome:

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Distributions of (W_i, Y_i, X_i) refer to the distribution induced by the random sampling from the population.

Several additional pieces of notation will be useful in the remainder of these notes. First, the propensity score (Rosenbaum and Rubin, 1983) is defined as the conditional probability of receiving the treatment,

$$e(x) = \Pr(W_i = 1 | X_i = x) = \mathbb{E}[W_i | X_i = x].$$

Also, define, for $w \in \{0, 1\}$, the two conditional regression and variance functions:

$$\mu_w(x) = \mathbb{E}[Y_i(w) | X_i = x], \qquad \sigma_w^2(x) = \mathbb{V}(Y_i(w) | X_i = x).$$

2.2 Estimands: Average Treatment Effects

[1] Calling such variables exogenous is somewhat at odds with several formal definitions of exogeneity (e.g., Engle, Hendry and Richard, 1974), as knowledge of their distribution can be informative about the average treatment effects. It does, however, agree with common usage. See for example, Manski, Sandefur, McLanahan, and Powers (1992, p. 28).

In this discussion we will primarily focus on a number of average treatment effects (ATEs). For a discussion of testing for the presence of any treatment effects under unconfoundedness see Crump, Hotz, Imbens and Mitnik (2007). Focusing on average effects is less limiting than it may seem, however, as this includes averages of arbitrary transformations of the original outcomes.[2] The first estimand, and the most commonly studied in the econometric literature, is the population average treatment effect (PATE):

$$\tau_P = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Alternatively we may be interested in the population average treatment effect for the treated (PATT, e.g., Rubin, 1977; Heckman and Robb, 1984):

$$\tau_{P,T} = \mathbb{E}[Y_i(1) - Y_i(0)|W = 1].$$

Most of the discussion in these notes will focus on τ_P, with extensions to $\tau_{P,T}$ available in the references.

We will also look at sample average versions of these two population measures. These estimands focus on the average of the treatment effect in the specific sample, rather than in the population at large. These include, the sample average treatment effect (SATE) and the sample average treatment effect for the treated (SATT):

$$\tau_S = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big), \qquad \text{and} \quad \tau_{S,T} = \frac{1}{N_T}\sum_{i:W_i=1}\Big(Y_i(1) - Y_i(0)\Big),$$

where $N_T = \sum_{i=1}^{N} W_i$ is the number of treated units. The sample average treatment effects have received little attention in the recent econometric literature, although it has a long tradition in the analysis of randomized experiments (e.g., Neyman, 1923). Without further assumptions, the sample contains no information about the population ATE beyond the

[2]Lehman (1974) and Doksum (1974) introduce quantile treatment effects as the difference in quantiles between the two marginal treated and control outcome distributions. Bitler, Gelbach and Hoynes (2002) estimate these in a randomized evaluation of a social program. Firpo (2003) develops an estimator for such quantiles under unconfoundedness.

sample ATE. To see this, consider the case where we observe the sample $(Y_i(0), Y_i(1), W_i, X_i)$, $i = 1, \ldots, N$; that is, we observe for each unit both potential outcomes. In that case the sample average treatment effect, $\tau_S = \sum_i (Y_i(1) - Y_i(0))/N$, can be estimated without error. Obviously the best estimator for the population average effect, τ_P, is τ_S. However, we cannot estimate τ_P without error even with a sample where all potential outcomes are observed, because we lack the potential outcomes for those population members not included in the sample. This simple argument has two implications. First, one can estimate the sample ATE at least as accurately as the population ATE, and typically more so. In fact, the difference between the two variances is the variance of the treatment effect, which is zero only when the treatment effect is constant. Second, a good estimator for one average treatment effect is automatically a good estimator for the other. One can therefore interpret many of the estimators for PATE or PATT as estimators for SATE or SATT, with lower implied standard errors.

The difference in asymptotic variances forces the researcher to take a stance on what the quantity of interest is. For example, in a specific application one can legitimately reach the conclusion that there is no evidence, at the 95% level, that the PATE is different from zero, whereas there may be compelling evidence that the SATE is positive. Typically researchers in econometrics have focused on the PATE, but one can argue that it is of interest, when one cannot ascertain the sign of the population-level effect, to know whether one can determine the sign of the effect for the sample. Especially in cases, which are all too common, where it is not clear whether the sample is representative of the population of interest, results for the sample at hand may be of considerable interest.

2.2 Identification

We make the following key assumption about the treatment assignment:

Assumption 1 (Unconfoundedness)

$$\Big(Y_i(0), Y_i(1)\Big) \perp\!\!\!\perp W_i \,\Big|\, X_i.$$

This assumption was first articulated in this form in Rosenbaum and Rubin (1983a). Lechner (1999, 2002) refers to this as the "conditional independence assumption," Following a parametric version of this in Heckman and Robb (1984) it is also referred to as "selection on observables." In the missing data literature the equivalent assumption is referred to as "missing at random."

To see the link with standard exogeneity assumptions, suppose that the treatment effect is constant: $\tau = Y_i(1) - Y_i(0)$ for all i. Suppose also that the control outcome is linear in X_i:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i,$$

with $\varepsilon_i \perp\!\!\!\perp X_i$. Then we can write

$$Y_i = \alpha + \tau \cdot W_i + X_i'\beta + \varepsilon_i.$$

Given the constant treatment effect assumption, unconfoundedness is equivalent to independence of W_i and ε_i conditional on X_i, which would also capture the idea that W_i is exogenous. Without this constant treatment effect assumption, however, unconfoundedness does not imply a linear relation with (mean-)independent errors.

Next, we make a second assumption regarding the joint distribution of treatments and covariates:

Assumption 2 (Overlap)

$$0 < \Pr(W_i = 1|X_i) < 1.$$

Rosenbaum and Rubin (1983a) refer to the combination of the two assumptions as "stongly ignorable treatment assignment." For many of the formal results one will also need smoothness assumptions on the conditional regression functions and the propensity score ($\mu_w(x)$ and $e(x)$), and moment conditions on $Y_i(w)$. I will not discuss these regularity conditions here. Details can be found in the references for the specific estimators given below.

There has been some controversy about the plausibility of Assumptions 1 and 2 in economic settings and thus the relevance of the econometric literature that focuses on estimation and inference under these conditions for empirical work. In this debate it has been argued that agents' optimizing behavior precludes their choices being independent of the potential outcomes, whether or not conditional on covariates. This seems an unduly narrow view. In response I will offer three arguments for considering these assumptions. The first is a statistical, data descriptive motivation. A natural starting point in the evaluation of any program is a comparison of average outcomes for treated and control units. A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment). Such an analysis may not lead to the final word on the efficacy of the treatment, but the absence of such an analysis would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment.

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not. The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated. Economic theory can help in classifying variables into those that need to be adjusted for versus those that do not, on the basis of their role in the decision process (e.g., whether they enter the utility function or the constraints). Given that, the unconfoundedness assumption merely asserts that all variables that need to be adjusted for are observed by the researcher. This is an empirical question, and not one that should be controversial as a general principle. It is clear that settings where some of these covariates are not observed will require strong assumptions to allow for identification. Such assumptions include instrumental variables settings where some covariates are assumed to be independent of the potential outcomes. Absent those assumptions, typically only bounds can be identified (e.g., Manski, 1990, 1995).

A third, related, argument is that even when agents optimally choose their treatment, two agents with the same values for observed characteristics may differ in their treatment choices

without invalidating the unconfoundedness assumption if the difference in their choices is driven by differencese in unobserved characteristics that are themselves unrelated to the outcomes of interest. The plausability of this will depend critically on the exact nature of the optimization process faced by the agents. In particular it may be important that the objective of the decision maker is distinct from the outcome that is of interest to the evaluator. For example, suppose we are interested in estimating the average effect of a binary input (e.g., a new technology) on a firm's output. Assume production is a stochastic function of this input because other inputs (e.g., weather) are not under the firm's control, or $Y_i = g(W, \varepsilon_i)$. Suppose that profits are output minus costs, $\pi_i(w) = g(w, \varepsilon_i) - c_i \cdot w$, and also that a firm chooses a production level to maximize expected profits, equal to output minus costs:

$$W_i = \arg\max_w \mathbb{E}[\pi_i(w)|c_i] = \arg\max_w \mathbb{E}[g(w, \varepsilon_i) - c_i \cdot w|c_i],$$

implying

$$W_i = 1\{\mathbb{E}[g(1, \varepsilon_i) - g(0, \varepsilon_i) \geq c_i|c_i]\} = h(c_i).$$

If unobserved marginal costs c_i differ between firms, and these marginal costs are independent of the errors ε_i in the firms' forecast of production given inputs, then unconfoundedness will hold as

$$(g(0, \varepsilon_i), g(1, \varepsilon_i)) \perp\!\!\!\perp c_i.$$

Note that under the same assumptions one cannot necessarily identify the effect of the input on profits since $(\pi_i(0), \pi_i(1))$ are not independent of c_i. See for a related discussion, in the context of instrumental variables, Athey and Stern (1998). Heckman, Lalonde and Smith (2000) discuss alternative models that justify unconfoundedness. In these models individuals do attempt to optimize the same outcome that is the variable of interest to the evaluator. They show that selection on observables assumptions can be justified by imposing restrictions

on the way individuals form their expectations about the unknown potential outcomes. In general, therefore, a researcher may wish to, either as a final analysis or as part of a larger investigation, consider estimates based on the unconfoundedness assumption.

Given strongly ignorable treatment assignment one can identify the population average treatment effect. The key insight is that given unconfoundedness, the following equalities holds:

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x] = \mathbb{E}[Y_i(w)|W_i = w, X_i = x] = \mathbb{E}[Y_i|W_i = w, X_i = x],$$

and $\mu_w(x)$ is identified. Thus one can estimate the average treatment effect τ by first estimating the average treatment effect for a subpopulation with covariates $X = x$:

$$\begin{aligned}\tau(x) &\equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x] \\ &= \mathbb{E}[Y_i(1)|X_i = x, W_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, W_i = 0] \\ &= \mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0].\end{aligned}$$

To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y_i|X_i = x, W_i = w]$ for all values of w and x in the support of these variables. This is where the second assumption enters. If the overlap assumption is violated at $X = x$, it would be infeasible to estimate both $\mathbb{E}[Y_i|X_i = x, W_i = 1]$ and $\mathbb{E}[Y_i|X_i = x, W_i = 0]$ because at those values of x there would be either only treated or only control units.

Some researchers use weaker versions of the unconfoundedness assumption (e.g., Heckman, Ichimura, and Todd, 1998). If the interest is in the population average treatment effect, it is in fact sufficient to assume that

$$\mathbb{E}[Y_i(w)|W_i, X_i] = \mathbb{E}[Y_i(w)|X_i],$$

for $w = 0, 1$. Although this assumption is unquestionably weaker, in practice it is rare that a convincing case is made for the weaker assumption without the case being equally strong

for the stronger Assumption 1. The reason is that the weaker assumption is intrinsically tied to functional form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (e.g., logarithms) without the strong assumption.

One can weaken the unconfoundedness assumption in a different direction if one is only interested in the average effect for the treated (e.g., Heckman, Ichimura and Todd, 1997). In that case one need only assume $Y_i(0) \perp\!\!\!\perp W_i \,\Big|\, X_i$. and the weaker overlap assumption $\Pr(W_i = 1|X_i) < 1$. These two assumptions are sufficient for identification of PATT because moments of the distribution of $Y(1)$ for the treated are directly estimable.

An important result building on the unconfoundedness assumption shows that one need not condition simultaneously on all covariates. The following result shows that all biases due to observable covariates can be removed by conditioning solely on the propensity score:

Result 1 *Suppose that Assumption 1 holds. Then:*

$$\Big(Y_i(0), Y_i(1)\Big) \perp\!\!\!\perp W_i \,\Big|\, e(X_i).$$

Proof: We will show that $\Pr(W_i = 1|Y_i(0), Y_i(1), e(X_i)) = \Pr(W_i = 1|e(X_i)) = e(X_i)$, implying independence of $(Y_i(0), Y_i(1))$ and W_i conditional on $e(X_i)$. First, note that

$$\begin{aligned}
\Pr(W_i = 1|Y_i(0), Y_i(1), e(X_i)) &= \mathbb{E}[W_i = 1|Y_i(0), Y_i(1), e(X_i)] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|Y_i(0), Y_i(1), e(X), X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|Y_i(0), Y_i(1), X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] = \mathbb{E}\left[e(X_i)|Y_i(0), Y_i(1), e(X_i)\right] = e(X_i),
\end{aligned}$$

where the last equality but one follows from unconfoundedness. The same argument shows that

$$\Pr(W_i = 1|e(X_i)) = \mathbb{E}[W_i = 1|e(X_i)] = \mathbb{E}\left[\mathbb{E}[W_i = 1|X_i]\Big|e(X_i)\right] = \mathbb{E}\left[e(X_i)|e(X_i)\right] = e(X_i).$$

□

Extensions of this result to the multivalued treatment case are given in Imbens (2000) and Lechner (2001).

To provide intuition for the Rosenbaum-Rubin result, recall the textbook formula for omitted variable bias in the linear regression model. Suppose we have a regression model with two regressors:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2' X_i + \varepsilon_i.$$

The bias of omitting X_i from the regression on the coefficient on W_i is equal to $\beta_2'\delta$, where δ is the vector of coefficients on W_i in regressions of the elements of X_i on W_i. By conditioning on the propensity score we remove the correlation between X_i and W_i because $X_i \perp\!\!\!\perp W_i | e(X_i)$. Hence omitting X_i no longer leads to any bias (although it may still lead to some efficiency loss).

2.4 Efficiency Bounds and Asymptotic Variances for Population Average Treatment Effects

Next we review some results on the effiency bound for estimators of the average treatment effects τ_P. This requires strong ignorability and some smoothness assumptions on the conditional expectations of potential outcomes and the treatment indicator (for details, see Hahn, 1998). Formally, Hahn (1998) shows that for any regular estimator for τ_P, denoted by $\hat{\tau}$, with

$$\sqrt{N} \cdot (\hat{\tau} - \tau_P) \xrightarrow{d} \mathcal{N}(0, V),$$

we can show that

$$V \geq \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)} + (\tau(X_i) - \tau_P)^2 \right]. \tag{1}$$

Knowing the propensity score does not affect this efficiency bound.

Hahn also shows that asymptotically linear estimators exist that achieve the efficiency bound, and hence such efficient estimators can be approximated as

$$\hat{\tau} = \tau_P + \frac{1}{N}\sum_{i=1}^{N} \psi(Y_i, W_i, X_i, \tau_P) + o_p(N^{-1/2}),$$

where $\psi(\cdot)$ is the efficient score:

$$\psi(y, w, x, \tau_P) = \left(\frac{wy}{e(x)} - \frac{(1-w)y}{1-e(x)}\right) - \tau_P - \left(\frac{\mu_1(x)}{e(x)} + \frac{\mu_0(x)}{1-e(x)}\right) \cdot (w - e(x)). \tag{2}$$

3. Estimating Average Treatment Effects

Here we discuss the leading estimators for average treatment effects under unconfoundedness. What is remarkable about this literature is the wide range of ostensibly quite different estimators, many of which are regularly used in empirical work. We first briefly describe a number of the estimators, and then discuss their relative merits.

3.1 Regression

The first class of estimators relies on consistent estimation of $\mu_w(x)$ for $w = 0, 1$. Given $\hat{\mu}_w(x)$ for these regression functions, the PATE and SATE are estimated by averaging their difference over the empirical distribution of the covariates:

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} \Big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\Big). \tag{3}$$

In most implementations the average of the predicted treated outcome for the treated is equal to the average observed outcome for the treated (so that $\sum_i W_i \cdot \hat{\mu}_1(X_i) = \sum_i W_i \cdot Y_i$), and similarly for the controls, implying that $\hat{\tau}_{\text{reg}}$ can also be written as

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} W_i \cdot \Big(Y_i - \hat{\mu}_0(X_i)\Big) + (1 - W_i) \cdot \Big(\hat{\mu}_1(X_i) - Y_i\Big).$$

Early estimators for $\mu_w(x)$ included parametric regression functions, for example linear regression (e.g., Rubin, 1977). Such parametric alternatives include least squares estimators

with the regression function specified as

$$\mu_w(x) = \beta' x + \tau \cdot w,$$

in which case the average treatment effect is equal to τ. In this case one can estimate τ simply by least squares estimation using the regression function

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i.$$

More generally, one can specify separate regression functions for the two regimes, $\mu_w(x) = \beta_w' x$. In that case one estimate the two regression functions separately on the two subsamples and then substitute the predicted values in (3).

These simple regression estimators can be sensitive to differences in the covariate distributions for treated and control units. The reason is that in that case the regression estimators rely heavily on extrapolation. To see this, note that the regression function for the controls, $\mu_0(x)$ is used to predict missing outcomes for the treated. Hence on average one wishes to use predict the control outcome at $\overline{X}_T = \sum_i W_i \cdot X_i / N_T$, the average covariate value for the treated. With a linear regression function, the average prediction can be written as $\bar{Y}_C + \hat{\beta}'(\overline{X}_T - \overline{X}_C)$. If $\overline{X}_T$ and the average covariate value for the controls, $\overline{X}_C$ are close, the precise specification of the regression function will not matter much for the average prediction. However, with the two averages very different, the prediction based on a linear regression function can be sensitive to changes in the specification.

More recently, nonparametric estimators have been proposed. Imbens, Newey and Ridder (2005) and Chen, Hong, and Tarozzi (2005) propose estimating $\mu_w(x)$ through series or sieve methods. A simple version of that with a scalar X would specify the regression function as

$$\mu_w(x) = \sum_{l=0}^{L_N} \beta_{w,l} \cdot x^k,$$

with L_N, the number of terms in the polynomial expansion, an increasing function of the sample size. They show that this estimator for τ_P achieves the semiparametric efficiency

bounds. Heckman, Ichimura and Todd (1997, 1998), and Heckman, Ichimura, Smith and Todd (1998) consider kernel methods for estimating $\mu_w(x)$, in particular focusing on local linear approaches. Given a kernel $K(\cdot)$, and a bandwidth h_N let

$$\left(\hat{\alpha}_{w,x}, \hat{\beta}_{w,x}\right) = \arg\min_{\alpha_{w,x},\beta_{w,x}} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h_N}\right) \cdot \left(Y_i - \alpha_{w,x} - \beta_{w,x} \cdot X_i\right)^2,$$

leading to the estimator

$$\hat{\mu}_w(x) = \hat{\alpha}_{w,x}.$$

3.2 Matching

Regression estimators impute the missing potential outcomes using the estimated regression function. Thus, if $W_i = 1$, $Y_i(1)$ is observed and $Y_i(0)$ is missing and imputed with a consistent estimator $\hat{\mu}_0(X_i)$ for the conditional expectation. Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of nearest neighbours of the opposite treatment group. In that sense matching is similar to nonparametric kernel regression methods, with the number of neighbors playing the role of the bandwidth in the kernel regression. In fact, matching can be inrrepreted as a limiting version of the standard kernel estimator where the bandwidth goes to zero. This minimizes the bias among nonnegative kernels, but potentially increases the variance relative to kernel estimators. A formal difference with kernel estimators is that the asymptotic distribution is derived conditional on the implicit bandwidth, that is, the number of neighbours, which is often fixed at one. Using such asymptotics, the implicit estimate $\hat{\mu}_w(x)$ is (close to) unbiased, but not consistent for $\mu_w(x)$. In contrast, the regression estimators discussed earlier relied on the consistency of $\mu_w(x)$.

Matching estimators have the attractive feature that given the matching metric, the researcher only has to choose the number of matches. In contrast, for the regression estimators discussed above, the researcher must choose smoothing parameters that are more difficult to interpret; either the number of terms in a series or the bandwidth in kernel regression.

Within the class of matching estimators, using only a single match leads to the most credible inference with the least bias, at most sacrificing some precision. This can make the matching estimator easier to use than those estimators that require more complex choices of smoothing parameters, and may explain some of its popularity.

Matching estimators have been widely studied in practice and theory (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1989, 1995, 2002; Rubin, 1973b, 1979; Heckman, Ichimura and Todd, 1998; Dehejia and Wahba, 1999; Abadie and Imbens, 2002, AI). Most often they have been applied in settings with the following two characteristics: (i) the interest is in the average treatment effect for the treated, and (ii), there is a large reservoir of potential controls. This allows the researcher to match each treated unit to one or more distinct controls (referred to as matching without replacement). Given the matched pairs, the treatment effect within a pair is then estimated as the difference in outcomes, with an estimator for the PATT obtained by averaging these within-pair differences. Since the estimator is essentially the difference in two sample means, the variance is calculated using standard methods for differences in means or methods for paired randomized experiments. The remaining bias is typically ignored in these studies. The literature has studied fast algorithms for matching the units, as fully efficient matching methods are computationally cumbersome (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1995). Note that in such matching schemes the order in which the units are matched is potentially important.

Here we focus on matching estimators for PATE and SATE. In order to estimate these targes we need to match both treated and controls, and allow for matching with replacement. Formally, given a sample, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $\ell_m(i)$ be the index l that satisfies $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1\Big\{\|X_j - X_i\| \leq \|X_l - X_i\|\Big\} = m,$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the m-th closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$.

In particular, $\ell_1(i)$ is the nearest match for unit i. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i: $\mathcal{J}_M(i) = \{\ell_1(i), \ldots, \ell_M(i)\}$. Define the imputed potential outcomes as:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

The simple matching estimator is then

$$\hat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{Y}_i(1) - \hat{Y}_i(0)\Big). \tag{4}$$

AI show that the bias of this estimator is of order $O(N^{-1/K})$, where K is the dimension of the covariates. Hence, if one studies the asymptotic distribution of the estimator by normalizing by $\sqrt{N}$ (as can be justified by the fact that the variance of the estimator is of order $O(1/N)$), the bias does not disappear if the dimension of the covariates is equal to two, and will dominate the large sample variance if K is at least three.

Let us make clear three caveats to the AI result. First, it is only the continuous covariates that should be counted in K. With discrete covariates the matching will be exact in large samples, therefore such covariates do not contribute to the order of the bias. Second, if one matches only the treated, and the number of potential controls is much larger than the number of treated units, one can justify ignoring the bias by appealling to an asymptotic sequence where the number of potential controls increases faster than the number of treated units. Specifically, if the number of controls, N_0, and the number of treated, N_1, satisfy $N_1/N_0^{4/K} \to 0$, then the bias disappears in large samples after normalization by $\sqrt{N}_1$. Third, even though the order of the bias may be high, the actual bias may still be small if the coefficients in the leading term are small. This is possible if the biases for different units are at least partially offsetting. For example, the leading term in the bias relies on the regression function being nonlinear, and the density of the covariates having a nonzero slope. If either the regression function is close to linear, or the density of the covariates close to constant, the resulting bias may be fairly limited. To remove the bias, AI suggest combining the matching process with a regression adjustment.

Another point made by AI is that matching estimators are generally not efficient. Even in the case where the bias is of low enough order to be dominated by the variance, the estimators are not efficient given a fixed number of matches. To reach efficiency one would need to increase the number of matches with the sample size, as done implicitly in kernel estimators. In practice the efficiency loss is limited though, with the gain of going from two matches to a large number of matches bounded as a fraction of the standard error by 0.16 (see AI).

In the above discussion the distance metric in choosing the optimal matches was the standard Euclidan metric $d_E(x,z) = (x-z)'(x-z)$. All of the distance metrics used in practice standardize the covariates in some manner. The most popular metrics are the Mahalanobis metric, where

$$d_M(x,z) = (x-z)'(\Sigma_X^{-1})(x-z),$$

where Σ is covariance matrix of the covairates, and the diagonal version of that

$$d_{AI}(x,z) = (x-z)'\text{diag}(\Sigma_X^{-1})(x-z).$$

Note that depending on the correlation structure, using the Mahalanobis metric can lead to situations where a unit with $X_i = (5,5)$ is a closer match for a unith with $X_i = (0,0)$ than a unit with $X_i = (1,4)$, despite being further away in terms of each covariate separately.

3.3 Propensity Score Methods

Since the work by Rosenbaum and Rubin (1983a) there has been considerable interest in methods that avoid adjusting directly for all covariates, and instead focus on adjusting for differences in the propensity score, the conditional probability of receiving the treatment. This can be implemented in a number of different ways. One can weight the observations in terms of the propensity score (and indirectly also in terms of the covariates) to create balance between treated and control units in the weighted sample. Hirano, Imbens and Ridder (2003) show how such estimators can achieve the semiparametric efficiency bound.

Alternatively one can divide the sample into subsamples with approximately the same value of the propensity score, a technique known as blocking. Finally, one can directly use the propensity score as a regressor in a regression approach or match on the propensity score.

If the researcher knows the propensity score all three of these methods are likely to be effective in eliminating bias. Even if the resulting estimator is not fully efficient, one can easily modify it by using a parametric estimate of the propensity score to capture most of the efficiency loss. Furthermore, since these estimators do not rely on high-dimensional nonparametric regression, this suggests that their finite sample properties would be attractive.

In practice the propensity score is rarely known, and in that case the advantages of the estimators discussed below are less clear. Although they avoid the high-dimensional nonparametric estimation of the two conditional expectations $\mu_w(x)$, they require instead the equally high-dimensional nonparametric estimation of the propensity score. In practice the relative merits of these estimators will depend on whether the propensity score is more or less smooth than the regression functions, or whether additional information is available about either the propensity score or the regression functions.

3.3.1 Weighting

The first set of "propensity score" estimators use the propensity score as weights to create a balanced sample of treated and control observations. Simply taking the difference in average outcomes for treated and controls,

$$\hat{\tau} = \frac{\sum W_i Y_i}{\sum W_i} - \frac{\sum (1-W_i) Y_i}{\sum 1 - W_i},$$

is not unbiased for $\tau^P = \mathbb{E}[Y_i(1) - Y_i(0)]$ because, conditional on the treatment indicator, the distributions of the covariates differ. By weighting the units by the inverse of the probability of receiving the treatment, one can undo this imbalance. Formally, weighting estimators rely on the equalities:

$$\mathbb{E}\left[\frac{WY}{e(X)}\right] = \mathbb{E}\left[\frac{WY_i(1)}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY_i(1)}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{e(X)Y_i(1)}{e(X)}\right]\right] = \mathbb{E}[Y_i(1)],$$

and similarly

$$\mathbb{E}\left[\frac{(1-W)Y}{1-e(X)}\right] = \mathbb{E}[Y_i(0)],$$

implying

$$\tau_P = \mathbb{E}\left[\frac{W\cdot Y}{e(X)} - \frac{(1-W)\cdot Y}{1-e(X)}\right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right). \tag{5}$$

In this particular form this is not necessarily an attractive estimator. The main reason is that, although the estimator can be written as the difference between a weighted average of the outcomes for the treated units and a weighted average of the outcomes for the controls, the weights do not necessarily add to one. Specifically, in (5), the weights for the treated units add up to $(\sum W_i/e(X_i))/N$. In expectation this is equal to one, but since its variance is positive, in any given sample some of the weights are likely to deviate from one. One approach for improving this estimator is simply to normalize the weights to unity. One can further normalize the weights to unity within subpopulations as defined by the covariates. In the limit this leads to the estimator proposed by Hirano, Imbens and Ridder (2003) who suggest using a nonparametric series estimator for $e(x)$. More precisely, they first specify a sequence of functions of the covariates, e.g., a power series, $h_l(x)$, $l = 1,\ldots,\infty$. Next, they choose a number of terms, $L(N)$, as a function of the sample size, and then estimate the L-dimensional vector γ_L in

$$\Pr(W=1|X=x) = \frac{\exp((h_1(x),\ldots,h_L(x))\gamma_L)}{1+\exp((h_1(x),\ldots,h_L(x))\gamma_L)},$$

by maximizing the associated likelihood function. Let $\hat{\gamma}_L$ be the maximum likelihood estimate. In the third step, the estimated propensity score is calculated as:

$$\hat{e}(x) = \frac{\exp((h_1(x),\ldots,h_L(x))\hat{\gamma}_L)}{1+\exp((h_1(x),\ldots,h_L(x))\hat{\gamma}_L)}.$$

Finally they estimate the average treatment effect as:

$$\hat{\tau}_{\text{weight}} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1-W_i) \cdot Y_i}{1-\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{(1-W_i)}{1-\hat{e}(X_i)}. \tag{6}$$

Hirano, Imbens and Ridder (2003) show that this estimator is efficient, whereas with the true propensity score the estimator would not be fully efficient (and in fact not very attractive).

This estimator highlights one of the interesting features of the problem of efficiently estimating average treatment effects. One solution is to estimate the two regression functions $\mu_w(x)$ nonparametrically; that solution completely ignores the propensity score. A second approach is to estimate the propensity score nonparametrically, ignoring entirely the two regression functions. If appropriately implemented, both approaches lead to fully efficient estimators, but clearly their finite sample properties may be very different, depending, for example, on the smoothness of the regression functions versus the smoothness of the propensity score. If there is only a single binary covariate, or more generally with only discrete covariates, the weighting approach with a fully nonparametric estimator for the propensity score is numerically identical to the regression approach with a fully nonparametric estimator for the two regression functions.

One difficulty with the weighting estimators that are based on the estimated propensity score is again the problem of choosing the smoothing parameters. Hirano, Imbens and Ridder (2003) use series estimators, which requires choosing the number of terms in the series. Ichimura and Linton (2001) consider a kernel version, which involves choosing a bandwidth. Theirs is currently one of the few studies considering optimal choices for smoothing parameters that focuses specifically on estimating average treatment effects. A departure from standard problems in choosing smoothing parameters is that here one wants to use nonparametric regression methods even if the propensity score is known. For example, if the probability of treatment is constant, standard optimality results would suggest using a high degree of smoothing, as this would lead to the most accurate estimator for the propensity score. However, this would not necessarily lead to an efficient estimator for the average treatment effect of interest.

3.3.2 Blocking on the Propensity Score

In their original propensity score paper Rosenbaum and Rubin (1983a) suggest the following "blocking propensity score" estimator. Using the (estimated) propensity score, divide the sample into M blocks of units of approximately equal probability of treatment, letting J_{im} be an indicator for unit i being in block m. One way of implementing this is by dividing the unit interval into M blocks with boundary values equal to m/M for $m = 1, \ldots, M-1$, so that

$$J_{im} = 1\{(m-1)/M < e(X_i) \leq m/M\},$$

for $m = 1, \ldots, M$. Within each block there are N_{wm} observations with treatment equal to w, $N_{wm} = \sum_i 1\{W_i = w, J_{im} = 1\}$. Given these subgroups, estimate within each block the average treatment effect as if random assignment holds,

$$\hat{\tau}_m = \frac{1}{N_{1m}} \sum_{i=1}^{N} J_{im} W_i Y_i - \frac{1}{N_{0m}} \sum_{i=1}^{N} J_{im}(1 - W_i) Y_i.$$

Then estimate the overall average treatment effect as:

$$\hat{\tau}_{\text{block}} = \sum_{m=1}^{M} \hat{\tau}_m \cdot \frac{N_{1m} + N_{0m}}{N}.$$

Blocking can be interpreted as a crude form of nonparametric regression where the unknown function is approximated by a step function with fixed jump points. To establish asymptotic properties for this estimator would require establishing conditions on the rate at which the number of blocks increases with the sample size. With the propensity score known, these are easy to determine; no formal results have been established for the unknown case.

The question arises how many blocks to use in practice. Cochran (1968) analyses a case with a single covariate, and, assuming normality, shows that using five blocks removes at least 95% of the bias associated with that covariate. Since all bias, under unconfoudnedness, is

associated with the propensity score, this suggests that under normality five blocks removes most of the bias associated with all the covariates. This has often been the starting point of empirical analyses using this estimator (e.g., Rosenbaum and Rubin, 1983b; Dehejia and Wahba, 1999), and has been implemented in STATA by Becker and Ichino (2002). Often, however, researchers subsequently check the balance of the covariates within each block. If the true propensity score per block is constant, the distribution of the covariates among the treated and controls should be identical, or, in the evaluation terminology, the covariates should be balanced. Hence one can assess the adequacy of the statistical model by comparing the distribution of the covariates among treated and controls within blocks. If the distributions are found to be different, one can either split the blocks into a number of subblocks, or generalize the specification of the propensity score. Often some informal version of the following algorithm is used: If within a block the propensity score itself is unbalanced, the blocks are too large and need to be split. If, conditional on the propensity score being balanced, the covariates are unbalanced, the specification of the propensity score is not adequate. No formal algorithm exists for implementing these blocking methods.

3.3.3 Regression on the Propensity Score

The third method of using the propensity score is to estimate the conditional expectation of Y given W and $e(X)$ and average the difference. Although this method has been used in practice, there is no particular reason why this is an attractive method compared to the regression methods based on the covariates directly. In addition, the large sample properties have not been established.

3.3.4 Matching on the Propensity Score

The Rosenbaum-Rubin result implies that it is sufficient to adjust solely for differences in the propensity score between treated and control units. Since one of the ways in which one can adjust for differences in covariates is matching, another natural way to use the propensity score is through matching. Because the propensity score is a scalar function of the covariates, the bias results in Abadie and Imbens (2002) imply that the bias term is of lower order

than the variance term and matching leads to a $\sqrt{N}$-consistent, asymptotically normally distributed estimator. The variance for the case with matching on the true propensity score also follows directly from their results. More complicated is the case with matching on the estimated propensity score. We are not aware of any results that give the asymptotic variance for this case.

3.4. Mixed Methods

A number of approaches have been proposed that combine two of the three methods described earlier, typically regression with one of its alternatives. These methods appear to be the most attractive in practice. The motivation for these combinations is that, although one method alone is often sufficient to obtain consistent or even efficient estimates, incorporating regression may eliminate remaining bias and improve precision. This is particularly useful because neither matching nor the propensity score methods directly address the correlation between the covariates and the outcome. The benefit associated with combining methods is made explicit in the notion developed by Robins and Ritov (1997) of "double robustness." They propose a combination of weighting and regression where, as long as the parametric model for either the propensity score or the regression functions is specified correctly, the resulting estimator for the average treatment effect is consistent. Similarly, because matching is consistent with few assumptions beyond strong ignorability, thus methods that combine matching and regressions are robust against misspecification of the regression function.

3.4.1 Weighting and Regression

One can rewrite the HIR weighting estimator discussed above as estimating the following regression function by weighted least squares,

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i,$$

with weights equal to

$$\lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1 - W_i}{1 - e(X_i)}}.$$

Without the weights the least squares estimator would not be consistent for the average treatment effect; the weights ensure that the covariates are uncorrelated with the treatment indicator and hence the weighted estimator is consistent.

This weighted-least-squares representation suggests that one may add covariates to the regression function to improve precision, for example as

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i,$$

with the same weights λ_i. Such an estimator, using a more general semiparametric regression model, is suggested in Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), Robins and Ritov (1997), and implemented in Hirano and Imbens (2001). In the parametric context Robins and Ritov argue that the estimator is consistent as long as either the regression model or the propensity score (and thus the weights) are specified correctly. That is, in the Robins-Ritov terminology, the estimator is doubly robust.

3.4.2 Blocking and Regression

Rosenbaum and Rubin (1983b) suggest modifying the basic blocking estimator by using least squares regression within the blocks. Without the additional regression adjustment the estimated treatment effect within blocks can be written as a least squares estimator of τ_m for the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \varepsilon_i,$$

using only the units in block m. As above, one can also add covariates to the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \beta_m' X_i + \varepsilon_i,$$

again estimated on the units in block m.

3.4.3 Matching and Regression

Since Abadie and Imbens (2002) show that the bias of the simple matching estimator can dominate the variance if the dimension of the covariates is too large, additional bias corrections through regression can be particularly relevant in this case. A number of such corrections have been proposed, first by Rubin (1973b) and Quade (1982) in a parametric setting. Let $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ be the observed or imputed potential outcomes for unit i; where these estimated potential outcomes equal observed outcomes for some unit i and its match $\ell(i)$. The bias in their comparison, $\mathbb{E}[\hat{Y}_i(1) - \hat{Y}_i(0)] - (Y_i(1) - Y_i(0))$, arises from the fact that the covariates for units i and $\ell(i)$, X_i and $X_{\ell(i)}$ are not equal, although close because of the matching process.

To further explore this, focusing on the single match case, define for each unit:

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell(i)} & \text{if } W_i = 1, \end{cases} \qquad \hat{X}_i(1) = \begin{cases} X_{\ell(i)} & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

If the matching is exact $\hat{X}_i(0) = \hat{X}_i(1)$ for each unit. If not, these discrepancies will lead to potential bias. The difference $\hat{X}_i(1) - \hat{X}_i(0)$ will therefore be used to reduce the bias of the simple matching estimator.

Suppose unit i is a treated unit ($W_i = 1$), so that $\hat{Y}_i(1) = Y_i(1)$ and $\hat{Y}_i(0)$ is an imputed value for $Y_i(0)$. This imputed value is unbiased for $\mu_0(X_{\ell(i)})$ (since $\hat{Y}_i(0) = Y_{\ell(i)}$), but not necessarily for $\mu_0(X_i)$. One may therefore wish to adjust $\hat{Y}_i(0)$ by an estimate of $\mu_0(X_i) - \mu_0(X_{\ell(i)})$. Typically these corrections are taken to be linear in the difference in the covariates for units i and its match, that is, of the form $\beta_0'(\hat{X}_i(1) - \hat{X}_i(0) = \beta_0'(X_i - X_{\ell(i)})$. One proposed correction is to estimate $\mu_0(x)$ directly by taking the control units that are used as matches for the treated units, with weights corresponding to the number of times a control observations is used as a match, and estimate a linear regression of the form

$$Y_i = \alpha_0 + \beta_0' X_i + \varepsilon_i,$$

on the weighted control observations by least squares. (If unit i is a control unit the correction would be done using an estimator for the regression function $\mu_1(x)$ based on a linear

specification $Y_i = \alpha_1 + \beta_1' X_i$ estimated on the treated units.) AI show that if this correction is done nonparametrically, the resulting matching estimator is consistent and asymptotically normal, with its bias dominated by the variance.

4. Estimating Variances

The variances of the estimators considered so far typically involve unknown functions. For example, as discussed earlier, the variance of efficient estimators of PATE is equal to

$$V_P = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\mu_1(X_i) - \mu_0(X_i) - \tau)^2\right],$$

involving the two regression functions, the two conditional variances and the propensity score.

4.1 Estimating The Variance of Efficient Estimators for τ_P

For efficient estimators for τ_P the asymptotic variance is equal to the efficiency bound V_P. There are a number of ways we can estimate this. The first is essentially by brute force. All five components of the variance, $\sigma_0^2(x)$, $\sigma_1^2(x)$, $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, are consistently estimable using kernel methods or series, and hence the asymptotic variance can be estimated consistently. However, if one estimates the average treatment effect using only the two regression functions, it is an additional burden to estimate the conditional variances and the propensity score in order to estimate V_P. Similarly, if one efficiently estimates the average treatment effect by weighting with the estimated propensity score, it is a considerable additional burden to estimate the first two moments of the conditional outcome distributions just to estimate the asymptotic variance.

A second method applies to the case where either the regression functions or the propensity score is estimated using series or sieves. In that case one can interpret the estimators, given the number of terms in the series, as parametric estimators, and calculate the variance this way. Under some conditions that will lead to valid standard errors and confidence intervals.

A third approach is to use bootstrapping (Efron and Tibshirani, 1993; Horowitz, 2002). Although there is little formal evidence specific for these estimators, given that the estimators are asymptotically linear, it is likely that bootstrapping will lead to valid standard errors and confidence intervals at least for the regression and propensity score methods. Bootstrapping is not valid for matching estimators, as shown by Abadie and Imbens (2007) Subsampling (Politis and Romano, 1999) will still work in this setting.

4.2 Estimating The Conditional Variance

Here we focus on estimation of the variance of estimators for τ_S, which is the conditional variance of the various estimators, conditional on the covariates $\mathbf{X}$ and the treatment indicators $\mathbf{W}$. All estimators used in practice are linear combinations of the outcomes,

$$\hat{\tau} = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W}) \cdot Y_i,$$

with the $\lambda(\mathbf{X}, \mathbf{W})$ known functions of the covariates and treatment indicators. Hence the conditional variance is

$$V\left(\hat{\tau}|\mathbf{X}, \mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W})^2 \cdot \sigma^2_{W_i}(X_i).$$

The only unknown component of this variance is $\sigma^2_w(x)$. Rather than estimating this through nonparametric regression, AI suggest using matching to estimate $\sigma^2_w(x)$. To estimate $\sigma^2_{W_i}(X_i)$ one uses the closest match within the set of units with the same treatment indicator. Let $v(i)$ be the closest unit to i with the same treatment indicator ($W_{v(i)} = W_i$). The sample variance of the outcome variable for these 2 units can then be used to estimate $\sigma^2_{W_i}(X_i)$:

$$\hat{\sigma}^2_{W_i}(X_i) = \left(Y_i - Y_{v(i)}\right)^2/2.$$

Note that this estimator is not consistent estimators of the conditional variances. However this is not important, as we are interested not in the variances at specific points in the

covariates distribution, but in the variance of the average treatment effect. Following the process introduce above, this is estimated as:

$$\hat{V}\left(\hat{\tau}|\mathbf{X},\mathbf{W}\right)=\sum_{i=1}^{N}\lambda_i(\mathbf{X},\mathbf{W})^2\cdot\hat{\sigma}^2_{W_i}(X_i).$$

5. Assessing Unconfoundedness

The unconfoundedness assumption used throughout this discussion is not directly testable. It states that the conditional distribution of the outcome under the control treatment, $Y_i(0)$, given receipt of the active treatment and given covariates, is identical to the distribution of the control outcome given receipt of the control treatment and given covariates. The same is assumed for the distribution of the active treatment outcome, $Y_i(1)$. Yet since the data are completely uninformative about the distribution of $Y_i(0)$ for those who received the active treatment and of $Y_i(1)$ for those receiving the control, the data cannot directly reject the unconfoundedness assumption. Nevertheless, there are often indirect ways of assessing the this, a number of which are developed in Heckman and Hotz (1989) and Rosenbaum (1987). These methods typically rely on estimating a causal effect that is known to equal zero. If based on the test we reject the null hypothesis that this causal effect varies from zero, the unconfoundedness assumption is considered less plausible. These tests can be divided into two broad groups.

The first set of tests focuses on estimating the causal effect of a treatment that is known not to have an effect, relying on the presence of multiple control groups (Rosenbaum, 1987). Suppose one has two potential control groups, for example eligible nonparticipants and ineligibles, as in Heckman, Ichimura and Todd (1997). One interpretation of the test is to compare average treatment effects estimated using each of the control groups. This can also be interpreted as estimating an "average treatment effect" using only the two control groups, with the treatment indicator now a dummy for being a member of the first group. In that case the treatment effect is known to be zero, and statistical evidence of a non-zero effect implies that at least one of the control groups is invalid. Again, not rejecting the test does not imply the unconfoundedness assumption is valid (as both control groups could

suffer the same bias), but non-rejection in the case where the two control groups are likely to have different biases makes it more plausible that the unconfoundness assumption holds. The key for the power of this test is to have available control groups that are likely to have different biases, if at all. Comparing ineligibles and eligible nonparticipants is a particularly attractive comparison. Alternatively one may use different geographic controls, for example from areas bordering on different sides of the treatment group.

One can formalize this test by postulating a three-valued indicator $T_i \in \{-0, 1, 1\}$ for the groups (e.g., ineligibles, eligible nonnonparticipants and participants), with the treatment indicator equal to $W_i = 1\{T_i = 1\}$, so that

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i \in \{-1, 0\} \\ Y_i(1) & \text{if } T_i = 1. \end{cases}$$

If one extends the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates,

$$Y_i(0), Y_i(1) \perp\!\!\!\perp T_i \,\Big|\, X_i,$$

then a testable implication is

$$Y_i(0) \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\},$$

and thus

$$Y_i \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\}.$$

An implication of this independence condition is being tested by the tests discussed above. Whether this test has much bearing on the unconfoundedness assumption depends on whether the extension of the assumption is plausible given unconfoundedness itself.

The second set of tests of unconfoundedness focuses on estimating the causal effect of the treatment on a variable known to be unaffected by it, typically because its value is

determined prior to the treatment itself. Such a variable can be time-invariant, but the most interesting case is in considering the treatment effect on a lagged outcome, commonly observed in labor market programs. If the estimated effect differs from zero, this implies that the treated observations are different from the controls in terms of this particular covariate given the others. If the treatment effect is estimated to be close to zero, it is more plausible that the unconfoundedness assumption holds. Of course this does not directly test this assumption; in this setting, being able to reject the null of no effect does not directly reflect on the hypothesis of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power. For these tests it is clearly helpful to have a number of lagged outcomes.

To formalize this, let us suppose the covariates consist of a number of lagged outcomes $Y_{i,-1},\ldots,Y_{i,-T}$ as well as time-invariant individual characteristics Z_i, so that $X_i = (Y_{i,-1},\ldots,Y_{i,-T},Z_i)$. By construction only units in the treatment group after period -1 receive the treatment; all other observed outcomes are control outcomes. Also suppose that the two potential outcomes $Y_i(0)$ and $Y_i(1)$ correspond to outcomes in period zero. Now consider the following two assumptions. The first is unconfoundedness given only $T-1$ lags of the outcome:

$$Y_{i,0}(1), Y_{i,0}(0) \;\perp\!\!\!\perp\; W_i \;\Big|\; Y_{i,-1},\ldots,Y_{i,-(T-1)},Z_i,$$

and the second assumes stationarity and exchangeability:

$$f_{Y_{i,s}(0)|Y_{i,s-1}(0),\ldots,Y_{i,s-(T-1)}(0),Z_i,W_i}(y_s|y_{s-1},\ldots,y_{s-(T-1)},z,w), \quad \text{does not depend on } i \text{ and } s.$$

Then it follows that

$$Y_{i,-1} \;\perp\!\!\!\perp\; W_i \;\Big|\; Y_{i,-2},\ldots,Y_{i,-T},Z_i,$$

which is testable. This hypothesis is what the procedure described above tests. Whether this test has much bearing on unconfoundedness depends on the link between the two assumptions and the original unconfoundedness assumption. With a sufficient number of lags

unconfoundedness given all lags but one appears plausible conditional on unconfoundedness given all lags, so the relevance of the test depends largely on the plausibility of the second assumption, stationarity and exchangeability.

6. Assessing Overlap

The second of the key assumptions in estimating average treatment effects requires that the propensity score is strictly between zero and one. Although in principle this is testable, as it restricts the joint distribution of observables, formal tests are not the main concern. In practice, this assumption raises a number of issues. The first question is how to detect a lack of overlap in the covariate distributions. A second is how to deal with it, given that such a lack exists.

6.1 Propensity Score Distributions

The first method to detect lack of overlap is to plot distributions of covariates by treatment groups. In the case with one or two covariates one can do this directly. In high dimensional cases, however, this becomes more difficult. One can inspect pairs of marginal distributions by treatment status, but these are not necessarily informative about lack of overlap. It is possible that for each covariate the distribution for the treatment and control groups are identical, even though there are areas where the propensity score is zero or one.

A more direct method is to inspect the distribution of the propensity score in both treatment groups, which can reveal lack of overlap in the multivariate covariate distributions. Its implementation requires nonparametric estimation of the propensity score, however, and misspecification may lead to failure in detecting a lack of overlap, just as inspecting various marginal distributions may be insufficient. In practice one may wish to undersmooth the estimation of the propensity score, either by choosing a bandwidth smaller than optimal for nonparametric estimation or by including higher order terms in a series expansion.

6.2 Selecting a Sample with Overlap

Once one determines that there is a lack of overlap one can either conclude that the

average treatment effect of interest cannot be estimated with sufficient precision, and/or decide to focus on an average treatment effect that is estimable with greater accuracy. To do the latter it can be useful to discard some of the observations on the basis of their covariates. For example one may decide to discard control (treated) observations with propensity scores below (above) a cutoff level. To do this sytematically, we follow Crump, Hotz, Imbens and Mitnik (2006), who focus on sample averate treatment effects. Their starting point is the definition of average treatment effects for subsets of the covariate space. Let $\mathbb{X}$ be the covariate space, and $\mathbb{A} \subset \mathbb{X}$ be some subset. Then define

$$\tau(\mathbb{A}) = \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\} \cdot \tau(X_i) \Big/ \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\}.$$

Crump et al calculate the efficiency bound for $\tau(\mathbb{A})$, assuming homoskedasticity, as

$$\frac{\sigma^2}{q(\mathbb{A})} \cdot \mathbb{E}\left[\frac{1}{e(X)} + \frac{1)}{1-e(X)} \middle| X \in \mathbb{A} \right],$$

where $q(\mathbb{A}) = \Pr(X \in \mathbb{A})$. They derive the characterization for the set $\mathbb{A}$ that minimizes the asymptotic variance and show that it has the form

$$\mathbb{A}^* = \{x \in \mathbb{X} | \alpha \leq e(X) \leq 1-\alpha\},$$

dropping observations with extreme values for the propensity score, with the cutoff value α determined by the equation

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X) \cdot (1-e(X))} \middle| \frac{1}{e(X) \cdot (1-e(X))} \leq \frac{1}{\alpha \cdot (1-\alpha)} \right].$$

Crump et al then suggest estimating $\tau(\mathbb{A}^*)$. Note that this subsample is selected solely on the basis of the joint distribution of the treatment indicators and the covariates, and therefore does not introduce biases associated with selection based on the outcomes. Calculations for Beta distributions for the propensity score suggest that $\alpha = 0.1$ approximates the optimal set well in practice.

7. The Lalonde Data

Here we look at application of the ideas discussed in these notes. We take the NSW job training data orinally collected by Lalonde (1986), and subsequently analyzed by Dehejia and Wahba (1999). The starting point is an experimental evaluation of this training program. Lalonde then constructed non-experimental comparison groups to investigate the ability of various econometric techniques to replicate the experimental results. In the current analysis we use three subsamples, the (experimental) trainees, the experimental controls, and a CPS comparison group.

In the next two subsections we do the design part of the analysis. Without using the outcome data we assess whether strong ignorability has some credibility.

7.1 Summary Statistics

First we give some summary statistics

Table 1: Summary Statistics for Experimental Sample

	Controls (N=260)		Trainees (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	diff / sd	mean	(s.d.)	diff / sd
Age	25.05	7.06	25.82	7.16	0.11	33.23	11.05	-0.67
Black	0.83	0.38	0.84	0.36	0.04	0.07	0.26	2.80
Education	10.09	1.61	10.35	2.01	0.14	12.03	2.87	-0.59
Hispanic	0.11	0.31	0.06	0.24	-0.17	0.07	0.26	-0.05
Married	0.15	0.36	0.19	0.39	0.09	0.71	0.45	-1.15
Earnings '74	2.11	5.69	2.10	4.89	-0.00	14.02	9.57	-1.24
Earnings '75	1.27	3.10	1.53	3.22	0.08	0.12	0.32	1.77
Unempl '74	0.75	0.43	0.71	0.46	-0.09	13.65	9.27	-1.30
Unempl. '75	0.68	0.47	0.60	0.49	-0.18	0.11	0.31	1.54

In this table we report averages and standard deviations for the three subsamples. In addition we report for the trainee/experimental-control and for the trainee/CPS-comparison-group

pairs the difference in average covariate values by treatment status, normalized by the standard deviation of these covariates. So, in Table 1 we see that in the experimental data set the difference in average age between treated and controls is 0.11 standard deviations. In the nonexperimental comparison the difference in age is 0.67 standard deviations.

Note that we do not report the t-statistic for the difference. Essentially the t-statistic is equal to the normalized difference multiplied by the square root of the sample size. As such, the t-statistic partly reflects the sample size. Given a difference of 0.25 standard deviations between the two groups in terms of average covariate values, a larger t-statistic just indicates a larger sample size, and therefore in fact an easier problem in terms of finding credible estimators for average treatment effects. As this example illustrates, a larger t-statistic for the difference between average covariates by treatment group does not indicate that the problem of finding credible estimates of the treatment effect is more difficult. A larger normalized difference does unambiguously indicate a more severe overlap problem.

In general a difference in average means bigger than 0.25 standard deviations is substantial. In that case one may want to be suspicious of simple methods like linear regression with a dummy for the treatment variable. Recall that estimating the average effect essentially amounts to using the controls to estimate the conditional mean $\mu_0(x) = \mathbb{E}[Y_i|W_i = 1, X_i = x]$ and using this estimated regression function to predict the (missing) control outcomes for the treated units. With such a large difference between the two groups in covariate distributions, linear regression is going to rely heavily on extrapolation, and thus will be sensitive to the exact functional form.

Right away we can see that the experimental data set is well balanced. The difference in averages between treatment and control group is never more than 0.18 standard deviations. In contrast, with the CPS comparison group the differences between the averages are up to 1.77 standard deviations from zero, suggesting there will be serious issues in obtaining credible estimates of the average effect of the treatment.

In Figures 1 and 2 we present histogram estimates of the distribution of the propensity

Figure 1: histogram propensity score for controls, exper full sample

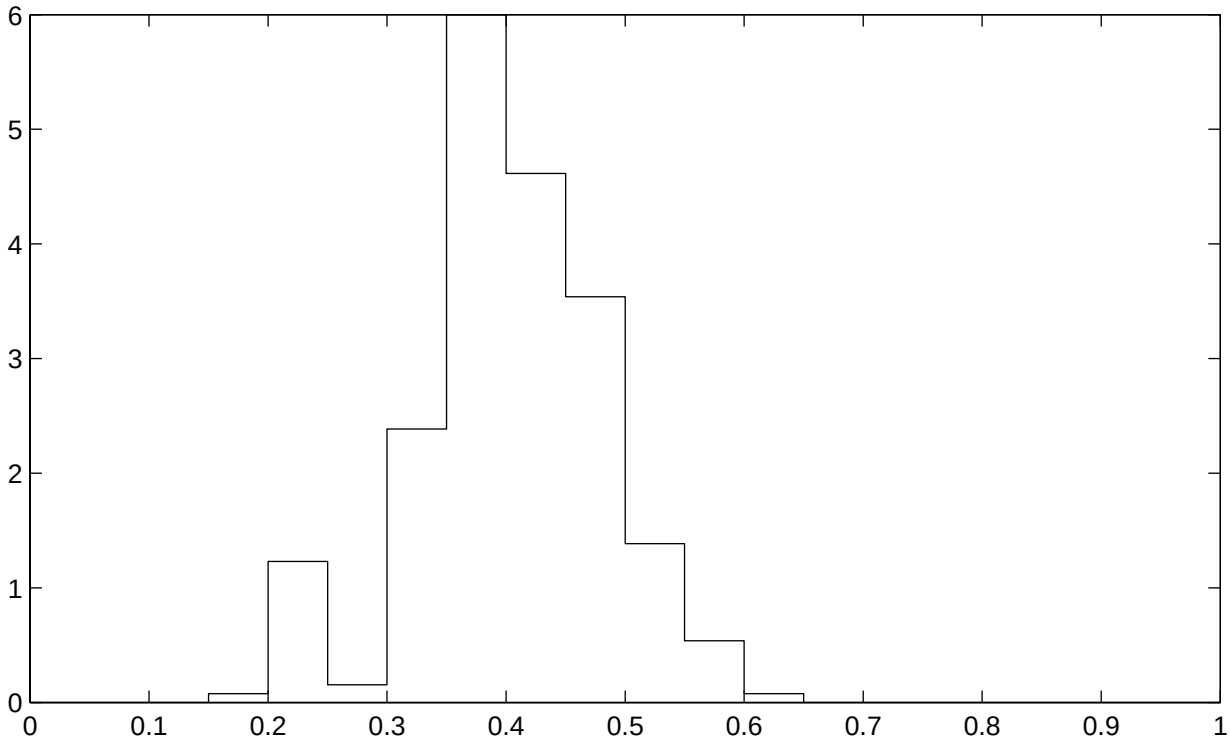


Figure 2: histogram propensity score for treated, exper full sample

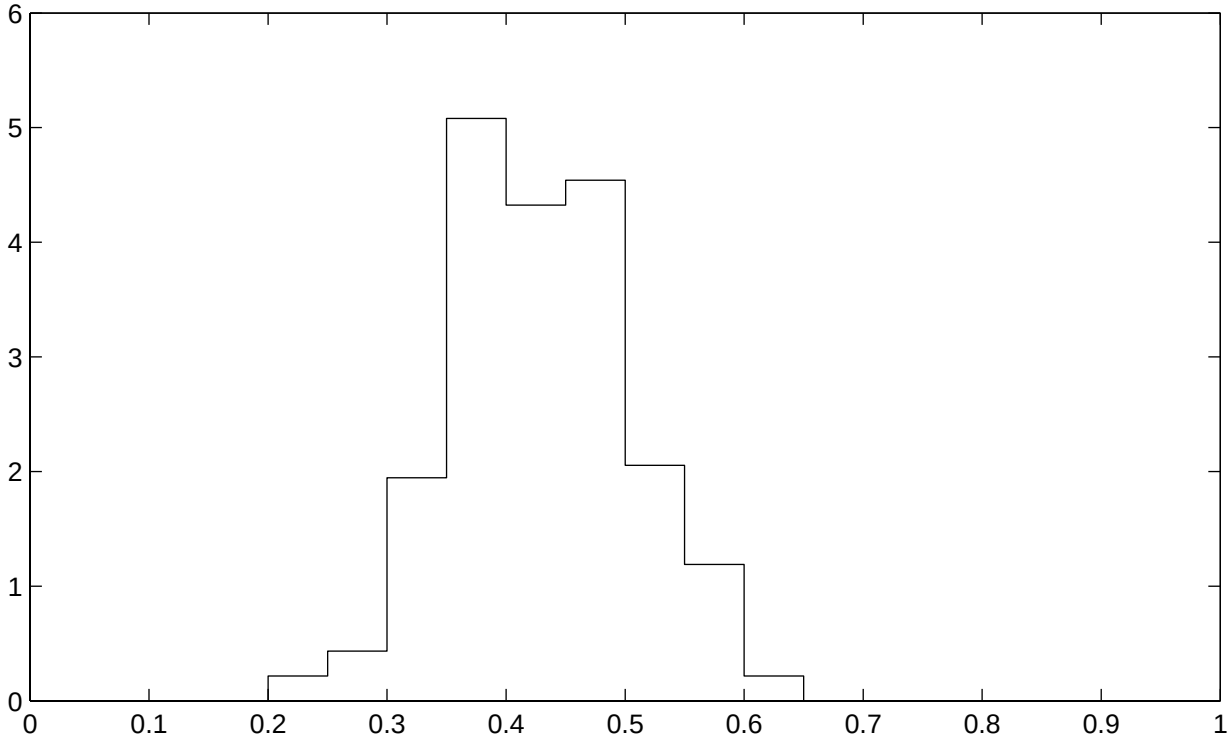


Figure 3: hist p−score for controls, cps full sample

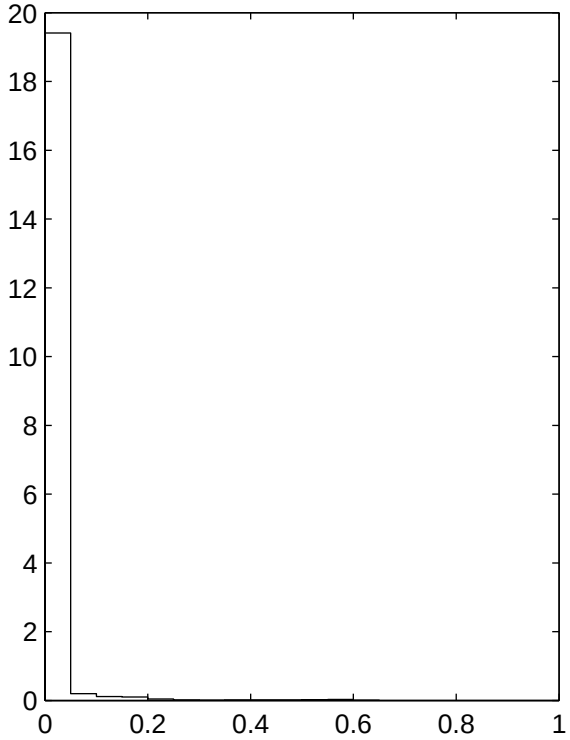


Figure 5: hist p−score for controls, cps selected sample

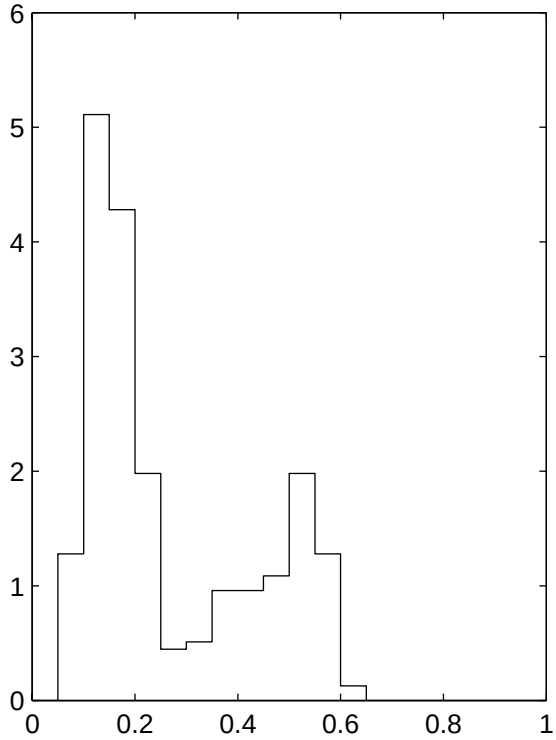


Figure 4: hist p−score for treated, cps full sample

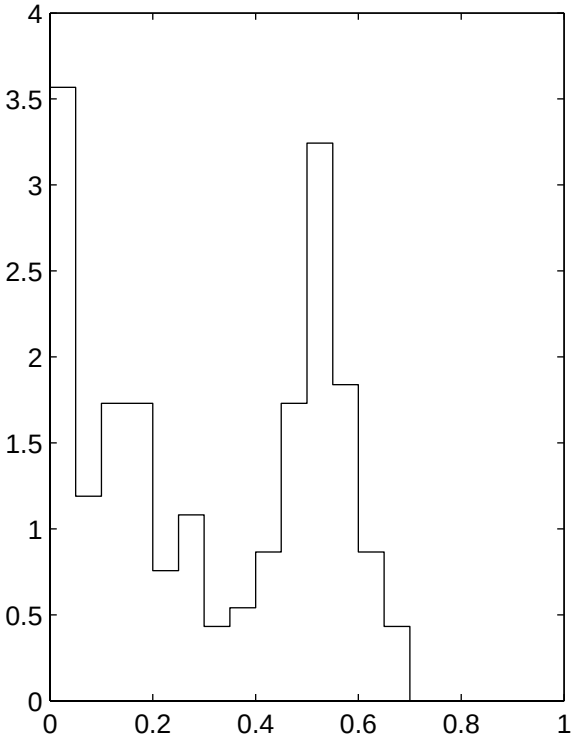
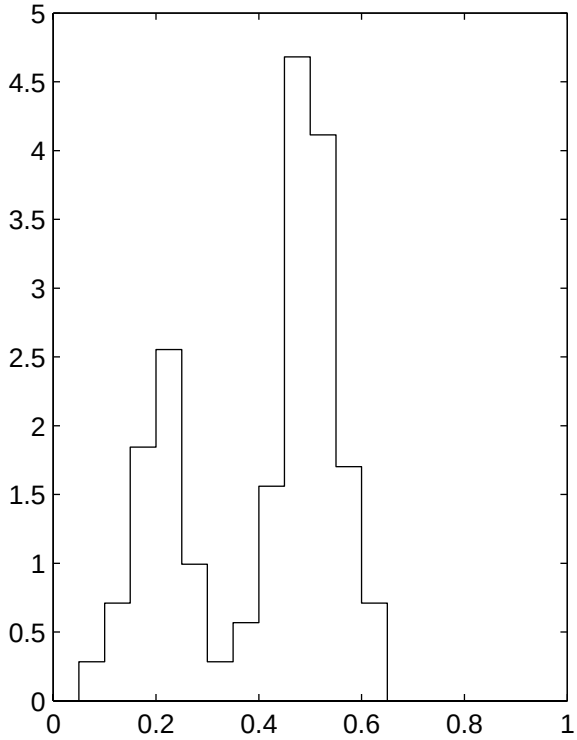


Figure 6: hist p−score for treated, cps selected sample



score for the treatment and control group in the experimental Lalonde data. These distributions again suggest that there is considerable overlap in the covariate distributions. In Figures 3 and 4 we present the histogram estimates for the propensity score distributions for the CPS comparison group. Now there is a clear lack of overlap. For the CPS comparison group almost all mass of the propensity score distribution is concentrated in a small interval to the right of zero, and the distribution for the treatment group is much more spread out.

7.2 Assessing Unconfoundedness

First we use the experimental data. We analyze the data as if earnings in 1975 (Earn '75) is the outcome. This is in fact a covariate, and so it cannot be affected by the treatment. Table 2 reports the results for eleven estimators .

Table 2: Estimates for Lalonde Data with Earnings '75 as Outcome

	Experimental Controls			CPS Comparison Group		
	mean	(s.e.)	t-stat	mean	(s.e.)	t-stat
Simple Dif	0.27	0.30	0.9	-12.12	0.68	-17.8
OLS (parallel)	0.15	0.22	0.7	-1.15	0.36	-3.2
OLS (separate)	0.12	0.22	0.6	-1.11	0.36	-3.1
Propensity Score Weighting	0.15	0.30	0.5	-1.17	0.26	-4.5
Propensity Score Blocking	0.10	0.17	0.6	-2.80	0.56	-5.0
Propensity Score Regression	0.16	0.30	0.5	-1.68	0.79	-2.1
Propensity Score Matching	0.23	0.37	0.6	-1.31	0.46	-2.9
Matching	0.14	0.28	0.5	-1.33	0.41	-3.2
Weighting and Regression	0.15	0.21	0.7	-1.23	0.24	-5.2
Blocking and Regression	0.09	0.15	0.6	-1.30	0.50	-2.6
Matching and Regression	0.06	0.28	0.2	-1.34	0.42	-3.2

For all eleven estimators the estimated effect is close to zero and statistically insignificant at conventional levels. The results suggest that unconfoundedness is plausible. With the CPS comparison group the results are very different. All estimators suggest substantial and statistically significant differences in earnings in 1975 after adjusting for all other covariates,

including earnings in 1974. This suggests that relying on the unconfoundedness assumption, in combination with these estimators, is not very credible for this sample.

7.3 Selecting A Subsample

Next we consider the effects of trimming the sample. We use the simple 0.1 rule where we drop observations with the propensity score outside of the interval $[0.1, 0.9]$. Table 3 we report the subsample sizes by treatment status and propensity score block.

Table 3: Sample Sizes for CPS Sample

	$\hat{e}(X_i) < 0.1$	$0.1 \leq \hat{e}(X_i) \leq 0.9$	$0.9 < \hat{e}(X_i)$	All
Controls	15679	313	0	15992
Trainees	44	141	0	185
All	15723	454	0	16177

Dropping observations with a propensity score less than 0.1 leads to discarding most of the controls, 15679 to be precise, leaving only 313 control observations. In addition 44 out of the 185 treated units are dropped. Nevertheless, the improved balance suggests that we obtain more precise estimates for the remaining sample.

Now let us consider the selected CPS sample. First we assess the balance by looking at the summary statistics.

Table 4: Summary Statistics for Selected CPS Sample

	Controls (N=313)		Trainees (N=141)		
	mean	(s.d.)	mean	(s.d.)	diff / sd
Age	26.60	10.97	25.69	7.29	-0.09
Black	0.94	0.23	0.99	0.12	0.21
Education	10.66	2.81	10.26	2.11	-0.15
Hispanic	0.06	0.23	0.01	0.12	-0.21
Married	0.22	0.42	0.13	0.33	-0.24
Earnings '74	1.96	4.08	1.34	3.72	-0.15
Earnings '75	0.57	0.50	0.80	0.40	0.49
Unempl '74	0.92	1.57	0.75	1.48	-0.11
Unempl. '75	0.55	0.50	0.69	0.46	0.28

These suggest that the balance is much improved, with the largest differences now on the order or 0.5 of a standard deviation, where before they difference was as high as 1.7.

Next we estimate the pseudo treatment effect on earnings in 1975.

Table 5: Estimates on Selected CPS Lalonde Data

	Earn '75 Outcome			Earn '78 Outcome		
	mean	(s.e.)	t-stat	mean	(s.e.)	t-stat
Simple Dif	-0.17	0.16	-1.1	1.73	0.68	2.6
OLS (parallel)	-0.09	0.14	-0.7	2.10	0.71	3.0
OLS (separate)	-0.19	0.14	-1.4	2.18	0.72	3.0
Propensity Score Weighting	-0.16	0.15	-1.0	1.86	0.75	2.5
Propensity Score Blocking	-0.25	0.25	-1.0	1.73	1.23	1.4
Propensity Score Regression	-0.07	0.17	-0.4	2.09	0.73	2.9
Propensity Score Matching	-0.01	0.21	-0.1	0.65	1.19	0.5
Matching	-0.10	0.20	-0.5	2.10	1.16	1.8
Weighting and Regression	-0.14	0.14	-1.1	1.96	0.77	2.5
Blocking and Regression	-0.25	0.25	-1.0	1.73	1.22	1.4
Matching and Regression	-0.11	0.19	-0.6	2.23	1.16	1.9

Here we find that all estimators find only small and insignificant effects of the treatment on

earnings in 1975. This suggests that for this sample unconfoundedness may well be a reasonable assumption, and that the estimators considered here can lead to credible estimates. Finally we report the estimates for earnings in 1978. Only now do we use the outcome data. Note that with the exclusion of the propensity score matching estimator the estimates are all between 1.73 and 2.23, and thus relatively insensitive to the choice of estimator.

References

Athey, S., and S. Stern, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", NBER working paper 6600.

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics,"Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Chen, X., H. Hong, and Tarozzi, (2005), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects,"unpublished working paper, Department of Economics, New York University.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2006), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand,"unpublished manuscript, Department of Economics, UC Berkeley.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2007), "Nonparametric Tests for Treatment Effect Heterogeneity,"forthcoming, *Review of Economics and Statistics*.

Dehejia, R. (2005) "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141-173.

Engle, R., D. Hendry, and J.-F. Richard, (1983) "Exogeneity," *Econometrica*, 51(2): 277-304.

Firpo, S. (2003), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259-276.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Heckman, J., and J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association.*, Vol. 84, No. 804, 862-874.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact

of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261–294.

Heckman, J., R. Lalonde, and J. Smith (2000), "The Economics and Econometrics of Active Labor Markets Programs,"in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Hirano, K., and J. Porter, (2005), "Asymptotics for Statistical Decision Rules,"Working Paper, Dept of Economics, University of Wisconsin.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189. July

Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions,"*Biometrika*, Vol. 87, No. 3, 706-710.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,"*Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and J. Wooldridge., (2007), "Recent Developments in the Econometrics of Program Evaluation,"unpublished manuscript, department of economics, Harvard University.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data,"*American Economic Review*, 76, 604-620.

Lechner, M., (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption,"in Lechner and Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg, Physica.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects,"*American Economic*

Review Papers and Proceedings, 80, 319-323.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., (2004), "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, 72(4), 1221-1246.

Manski, C. (2005), *Social Choice with Partial Knowledge of Treatment Response*, Princeton University Press.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Robins, J., and Y. Ritov, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16, 285-319.

Rosenbaum, P., (1987), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292–316.

Rosenbaum, P., (1995), *Observational Studies*, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", emphBiometrika, 70, 41-55.

Rosenbaum, P., and D. Rubin, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212-218.

Rubin, D., (1973a), "Matching to Remove Bias in Observational Studies", *Biometrics*, 29, 159-183.

Rubin, D., (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.

RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688-701.

RUBIN, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1-26.

RUBIN, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6:34–58.

RUBIN, D., (1990), "Formal Modes of Statistical Inference for Causal Effects", emphJournal of Statistical Planning and Inference, 25, 279-292.

What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 2, **Monday**, **July 30th**, **11.00-12.30 am**

Linear Panel Data Models

These notes cover some recent topics in linear panel data models. They begin with a "modern" treatment of the basic linear model, and then consider some embellishments, such as random slopes and time-varying factor loads. In addition, fully robust tests for correlated random effects, lack of strict exogeneity, and contemporaneous endogeneity are presented. Section 4 considers estimation of models without strictly exogenous regressors, and Section 5 presents a unified framework for analyzing pseudo panels (constructed from repeated cross sections).

1. Quick Overview of the Basic Model

Most of these notes are concerned with an unobserved effects model defined for a large population. Therefore, we assume random sampling in the cross section dimension. Unless stated otherwise, the asymptotic results are for a fixed number of time periods, T, with the number of cross section observations, N, getting large.

For some of what we do, it is critical to distinguish the underlying population model of interest and the sampling scheme that generates data that we can use to estimate the population parameters. The standard model can be written, for a generic i in the population, as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t = 1, \ldots, T, \tag{1.1}$$

where η_t is a separate time period intercept (almost always a good idea), $\mathbf{x}_{it}$ is a $1 \times K$ vector of explanatory variables, c_i is the time-constant unobserved effect, and the $\{u_{it} : t = 1, \ldots, T\}$ are idiosyncratic errors. Thanks to Mundlak (1978) and Chamberlain (1982), we view the c_i as random draws along with the observed variables. Then, one of the key issues is whether c_i is correlated with elements of $\mathbf{x}_{it}$.

It probably makes more sense to drop the i subscript in (1.1), which would emphasize that the equation holds for an entire population. But (1.1) is useful to emphasizing which factors change only across t, which change only change across i, and which change across i and t.It is sometimes convenient to subsume the time dummies in $\mathbf{x}_{it}$.

Ruling out correlation (for now) between u_{it} and $\mathbf{x}_{it}$, a sensible assumption is *contemporaneous exogeneity conditional on* c_i :

$$E(u_{it}|\mathbf{x}_{it}, c_i) = 0, t = 1, \ldots, T. \tag{1.2}$$

This equation really defines $\boldsymbol{\beta}$ in the sense that under (1.1) and (1.2),

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i, \tag{1.3}$$

so the β_j are partial effects holding fixed the unobserved heterogeneity (and covariates other than x_{tj}).

As is now well known, $\boldsymbol{\beta}$ is not identified only under (1.2). Of course, if we added $Cov(\mathbf{x}_{it}, c_i) = \mathbf{0}$ for any t, then $\boldsymbol{\beta}$ is identified and can be consistently estimated by a cross section regression using period t. But usually the whole point is to allow the unobserved effect to be correlated with time-varying $\mathbf{x}_{it}$.

We can allow general correlation if we add the assumption of *strict exogeneity conditional on* c_i:

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = 0, t = 1, \ldots, T, \tag{1.4}$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{1.5}$$

If the elements of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ have suitable time variation, $\boldsymbol{\beta}$ can be consistently estimated by fixed effects (FE) or first differencing (FD), or generalized least squares (GLS) or generalized method of moments (GMM) versions of them. If the simpler methods are used, and even if GLS is used, standard inference can and should be made fully robust to heteroksedasticity and serial dependence that could depend on the regressors (or not). These are the now well-known "cluster" standard errors. With large N and small T, there is little excuse not to compute them.

(Note: Some call (1.4) or (1.5) "strong" exogeneity. But in the Engle, Hendry, and Richard (1983) work, strong exogeneity incorporates assumptions on parameters in different conditional distributions being variation free, and that is not needed here.)

The strict exogeneity assumption is always violated if $\mathbf{x}_{it}$ contains lagged dependent variables, but it can be violated in other cases where $\mathbf{x}_{i,t+1}$ is correlated with u_{it} – a "feedback effect." An assumption more natural than strict exogeneity is *sequential exogeneity condition on* c_i:

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it}, c_i) = 0, t = 1, \ldots, T \tag{1.6}$$

or

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{1.7}$$

This allows for lagged dependent variables (in which case it implies that the dynamics in the

mean have been completely specified) and, generally, is more natural when we take the view that $\{\mathbf{x}_{it}\}$ might react to shocks that affect y_{it}. Generally, $\boldsymbol{\beta}$ is identified under sequential exogeneity. First differencing and using lags of $\mathbf{x}_{it}$ as instruments, or forward filtering, can be used in simple IV procedures or GMM procedures. (More later.)

If we are willing to assume c_i and $\mathbf{x}_i$ are uncorrelated, then many more possibilities arise (including, of course, identifying coefficients on time-constant explanatory variables). The most convenient way of stating the random effects (RE) assumption is

$$E(c_i|\mathbf{x}_i) = E(c_i), \tag{1.8}$$

although using the linear projection in place of $E(c_i|\mathbf{x}_i)$ suffices for consistency (but usual inference would not generally be valid). Under (1.8), we can used pooled OLS or any GLS procedure, including the usual RE estimator. Fully robust inference is available and should generally be used. (Note: The usual RE variance matrix, which depends only on σ_c^2 and σ_u^2, need not be correctly specified! It still makes sense to use it in estimation but make inference robust.)

It is useful to define two *correlated random effects* assumptions:

$$L(c_i|\mathbf{x}_i) = \psi + \mathbf{x}_i\boldsymbol{\xi}, \tag{1.9}$$

which actually is not an assumption but a definition. For nonlinear models, we will have to actually make assumptions about $D(c_i|\mathbf{x}_i)$, the conditional distribution. Methods based on (1.9) are often said to implement the *Chamberlain device*, after Chamberlain (1982).

Mundlak (1978) used a restricted version, and used a conditional expectation:

$$E(c_i|\mathbf{x}_i) = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \tag{1.10}$$

where $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$. This formulation conserves on degrees of freedom, and extensions are useful for nonlinear models.

If we write $c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i$ or $c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i$ and plug into the original equation, for example

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it} \tag{1.11}$$

(absorbing ψ into the time intercepts), then we are tempted to use pooled OLS, or RE estimation because $E(a_i + u_{it}|\mathbf{x}_i) = 0$. Either of these leads to the FE estimator of $\boldsymbol{\beta}$, and to a simple test of $H_0 : \boldsymbol{\xi} = \mathbf{0}$. Later, when we discuss control function methods, it will be handy to run regressions directly that include the time averages. (Somewhat surprisingly, obtain the

same algebraic equivalence using Chamberlain's devise. The pooled OLS estimator of $\boldsymbol{\beta}$ is still the FE estimator, even though the $\boldsymbol{\xi}_t$ might change substantially across t.)

Some of us have been pushing for several years the notion that specification tests should be made robust to assumptions that are not directly being tested. (Technically, they should be robust to assumptions that they have no asymptotic power for detecting violations of.) Much progress has been made, but one still sees Hausman statistics computed that maintain a full set of assumptions under the null. Take comparing random effects to fixed effects. The key assumption is (1.8). whether $Var(\mathbf{v}_i|\mathbf{x}_i)$ has the random effects structure, where $v_{it} = c_i + u_{it}$, should not be a critical issue. It makes no sense to report a fully robust variance matrix for FE and RE but then to compute a Hausman test that maintains the full set of RE assumptions. (In addition to (1.4) and (1.8), these are $Var(\mathbf{u}_i|\mathbf{x}_i, c_i) = \sigma_u^2\mathbf{I}_T$ and $Var(c_i|\mathbf{x}_i) = Var(c_i)$.) The regression-based Hausman test from (1.11) is very handy for obtaining a fully robust test. More specifically, suppose the model contains a full set of year intercepts as well as time-constant and time-varying explanatory variables:

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it}.$$

Now, it is clear that, because we cannot estimate $\boldsymbol{\gamma}$ by FE, it is not part of the Hausman test comparing RE and FE. What is less clear, but also true, is that the coefficients on the time dummies, $\boldsymbol{\eta}$, cannot be included, either. (RE and FE estimation only with aggregate time effects are identical.) In fact, we can only compare the $M \times 1$ estimates of $\boldsymbol{\delta}$, say $\hat{\boldsymbol{\delta}}_{FE}$ and $\hat{\boldsymbol{\delta}}_{RE}$. If we include $\hat{\boldsymbol{\eta}}_{FE}$ and $\hat{\boldsymbol{\eta}}_{RE}$ we introduce a nonsingularity in the asymptotic variance matrix. The regression based test, from the pooled regression

$$y_{it} \text{ on } \mathbf{g}_t, \mathbf{z}_i, \mathbf{w}_{it}, \bar{\mathbf{w}}_i, \; t = 1, \ldots, T; \; i = 1, \ldots, N$$

makes this clear (and that the are M restrictions to test). (Mundlak (1978) suggested this test and Arellano (1993) described the robust version.). Unfortunately, the usual form of the Hausman test does not, and, for example, Stata gets it wrong and tries to include the year dummies in the test (in addition to being nonrobust). The most important problem is that unwarranted degrees of freedom are added to the chi-square distribution, often many extra df, which can produce seriously misleading p-values.

2. New Insights Into Old Estimators

In the past several years, the properties of traditional estimators used for linear models, particularly fixed effects and its instrumental variable counterparts, have been studied under

weaker assumptions. We review some of those results here. In these notes, we focus on models without lagged dependent variables or other non-strictly exogenous explanatory variables, although the instrumental variables methods applied to linear models can, in some cases, be applied to models with lagged dependent variables.

2.1. Fixed Effects Estimation in the Correlated Random Slopes Model

The fixed effects (FE) estimator is still the workhorse in empirical studies that employ panel data methods to estimate the effects of time-varying explanatory variables. The attractiveness of the FE estimator is that it allows arbitrary correlation between the additive, unobserved heterogeneity and the explanatory variables. (Pooled methods that do not remove time averages, as well as the random effects (RE) estimator, essentially assume that the unobserved heterogeneity is uncorrelated with the covariates.) Nevertheless, the framework in which the FE estimator is typically analyzed is somewhat restrictive: the heterogeneity is assumed to be additive and is assumed to have a constant coefficients (factor loads) over time. Recently, Wooldridge (2005a) has shown that the FE estimator, and extensions that sweep away unit-specific trends, has robustness properties for estimating the population average effect (PAE) or average partial effect (APE).

We begin with an extension of the usual model to allow for unit-specific slopes,

$$y_{it} = c_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} \tag{2.1}$$

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{b}_i) = 0, t = 1, \ldots, T, \tag{2.2}$$

where $\mathbf{b}_i$ is $K \times 1$. Rather than acknowledge that $\mathbf{b}_i$ is unit-specific, we ignore the heterogeneity in the slopes and act as if $\mathbf{b}_i$ is constant for all i. We think c_i might be correlated with at least some elements of $\mathbf{x}_{it}$, and therefore we apply the usual fixed effects estimator. The question we address here is: when does the usual FE estimator consistently estimate the population average effect, $\boldsymbol{\beta} = \mathrm{E}(\mathbf{b}_i)$.

In addition to assumption (2.2), we naturally need the usual FE rank condition,

$$rank \sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}) = K. \tag{2.3}$$

Write $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$ where the unit-specific deviation from the average, $\mathbf{d}_i$, necessarily has a zero mean. Then

$$y_{it} = c_i + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{x}_{it}\mathbf{d}_i + u_{it} \equiv c_i + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} \tag{2.4}$$

where $v_{it} \equiv \mathbf{x}_{it}\mathbf{d}_i + u_{it}$. A sufficient condition for consistency of the FE estimator along with

(3) is

$$\mathrm{E}(\ddot{\mathbf{x}}_{it}'\ddot{v}_{it}) = \mathbf{0}, t = 1, \ldots, T. \tag{2.5}$$

Along with (2.2), it suffices that $\mathrm{E}(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\mathbf{d}_i) = \mathbf{0}$ for all t. A sufficient condition, and one that is easier to interpret, is

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1, \ldots, T. \tag{2.6}$$

Importantly, condition (2.6) allows the slopes, $\mathbf{b}_i$, to be correlated with the regressors $\mathbf{x}_{it}$ through permanent components. What it rules out is correlation between idiosyncratic movements in $\mathbf{x}_{it}$. We can formalize this statement by writing $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1, \ldots, T$. Then (2.6) holds if $\mathrm{E}(\mathbf{b}_i|\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iT}) = \mathrm{E}(\mathbf{b}_i)$. So $\mathbf{b}_i$ is allowed to be arbitrarily correlated with the permanent component, $\mathbf{f}_i$. (Of course, $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}$ is a special representation of the covariates, but it helps to illustrate condition (2.6).) Condition (2.6) is similar in spirit to the Mundlak (1978) assumption applied to the slopes (rather to the intercept):
$\mathrm{E}(\mathbf{b}_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}) = \mathrm{E}(\mathbf{b}_i|\bar{\mathbf{x}}_i)$

One implication of these results is that it is a good idea to use a fully robust variance matrix estimator with FE even if one thinks idiosyncratic errors are serially uncorrelated: the term $\ddot{\mathbf{x}}_{it}\mathbf{d}_i$ is left in the error term and causes heteroskedasticity and serial correlation, in general.

These results extend to a more general class of estimators that includes the usual fixed effects and random trend estimator. Write

$$y_{it} = \mathbf{w}_t\mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it}, \quad t = 1, \ldots, T \tag{2.7}$$

where $\mathbf{w}_t$ is a set of deterministic functions of time. We maintain the standard assumption (2.2) but with $\mathbf{a}_i$ in place of c_i. Now, the "fixed effects" estimator sweeps away $\mathbf{a}_i$ by netting out $\mathbf{w}_t$ from $\mathbf{x}_{it}$. In particular, now let $\ddot{\mathbf{x}}_{it}$ denote the residuals from the regression $\mathbf{x}_{it}$ on $\mathbf{w}_t, t = 1, \ldots, T$.

In the random trend model, $\mathbf{w}_t = (1, t)$, and so the elements of $\mathbf{x}_{it}$ have unit-specific linear trends removed in addition to a level effect. Removing even more of the heterogeneity from $\{\mathbf{x}_{it}\}$ makes it even more likely that (2.6) holds. For example, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, then $\mathbf{b}_i$ can be arbitrarily correlated with $(\mathbf{f}_i, \mathbf{h}_i)$. Of course, individually detrending the $\mathbf{x}_{it}$ requires at least three time periods, and it decreases the variation in $\ddot{\mathbf{x}}_{it}$ compared to the usual FE estimator. Not surprisingly, increasing the dimension of $\mathbf{w}_t$ (subject to the restriction $\dim(\mathbf{w}_t) < T$), generally leads to less precision of the estimator. See Wooldridge (2005a) for further discussion.

Of course, the first differencing transformation can be used in place of, or in conjunction

with, unit-specific detrending. For example, if we first difference followed by the within transformation, it is easily seen that a condition sufficient for consistency of the resulting estimator for $\boldsymbol{\beta}$ is

$$\mathrm{E}(\mathbf{b}_i|\Delta\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i), \quad t = 2,\ldots,T, \tag{2.8}$$

where $\Delta\ddot{\mathbf{x}}_{it} = \Delta\mathbf{x}_{it} - \overline{\Delta\mathbf{x}}$ are the demeaned first differences.

Now consider an important special case of the previous setup, where the regressors that have unit-specific coefficients are time dummies. We can write the model as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t c_i + u_{it}, t = 1,\ldots,T, \tag{2.9}$$

where, with small T and large N, it makes sense to treat $\{\eta_t : t = 1,\ldots,T\}$ as parameters, like $\boldsymbol{\beta}$. Model (2.9) is attractive because it allows, say, the return to unbobserved "talent" to change over time. Those who estimate, say, firm-level production functions like to allow the importance of unobserved factors, such as managerial skill, to change over time. Estimation of $\boldsymbol{\beta}$, along with the η_t, is a nonlinear problem. What if we just estimate $\boldsymbol{\beta}$ by fixed effects? Let $\mu_c = \mathrm{E}(c_i)$ and write (2.9) as

$$y_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t d_i + u_{it}, t = 1,\ldots,T, \tag{2.10}$$

where $\alpha_t = \eta_t\mu_c$ and $d_i = c_i - \mu_c$ has zero mean In addition, the composite error, $v_{it} \equiv \eta_t d_i + u_{it}$, is uncorrelated with $(\mathbf{x}_{i1}, \mathbf{x}_2, \ldots, \mathbf{x}_{iT})$ (as well as having a zero mean). It is easy to see that consistency of the usual FE estimator, which allows for different time period intercepts, is ensured if

$$\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, c_i) = \mathbf{0}, t = 1,\ldots,T. \tag{2.11}$$

In other words, the unobserved effects is uncorrelated with the deviations $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$.

If we use the extended FE estimators for random trend models, as above, then we can replace $\ddot{\mathbf{x}}_{it}$ with detrended covariates. Then, c_i can be correlated with underlying levels and trends in $\mathbf{x}_{it}$ (provided we have a sufficient number of time periods).

Using usual FE (with full time period dummies) does not allow us to estimate the η_t, or even determine whether the η_t change over time. Even if we are interested only in $\boldsymbol{\beta}$ when c_i and $\mathbf{x}_{it}$ are allowed to be correlated, being able to detect time-varying factor loads is important because (2.11) is not completely general. It is useful to have a simple test of $\mathrm{H}_0 : \eta_2 = \eta_3 = \ldots = \eta_T$ with some power against the alternative of time-varying coefficients. Then, we can determine whether a more sophisticated estimation method might be needed.

We can obtain a simple variable addition test that can be computed using linear estimation

methods if we specify a particular relationship between c_i and $\mathbf{x}_i$. We use the Mundlak (1978) assumption

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i. \tag{2.12}$$

Then

$$y_{it} = \eta_t\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t\bar{\mathbf{x}}_i\boldsymbol{\xi} + \eta_t a_i + u_{it} = \alpha_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + \lambda_t\bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + \lambda_t a_i + u_{it}, \tag{2.13}$$

where $\lambda_t = \eta_t - 1$ for all t. Under the null hypothesis, $\lambda_t = 0, t = 2, \ldots, T$. If we impose the null hypothesis, the resulting model is linear, and we can estimate it by pooled OLS of y_{it} on $1, d2_t, \ldots, dT_t, \mathbf{x}_{it}, \bar{\mathbf{x}}_i$ across t and i, where the dr_t are time dummies. A variable addition test that all λ_t are zero can be obtained by applying FE to the equation

$$y_{it} = \alpha_1 + \alpha_2 d2_t + \ldots + \alpha_T dT_t + \mathbf{x}_{it}\boldsymbol{\beta} + \lambda_2 d2_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}) + \ldots + \lambda_T dT_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}) + error_{it}, \tag{2.14}$$

and test the joint significance of the $T - 1$ terms $d2_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}), \ldots, dT_t(\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}})$. (The term $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$ would drop out of an FE estimation, and so we just omit it.) Note that $\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}$ is a scalar and so the test as $T - 1$ degrees of freedom. As always, it is prudent to use a fully robust test (even though, under the null, $\lambda_t a_i$ disappears from the error term).

A few comments about this test are in order. First, although we used the Mundlak device to obtain the test, it does not have to represent the actual linear projection because we are simply adding terms to an FE estimation. Under the null, we do not need to restrict the relationshp between c_i and $\mathbf{x}_i$. Of course, the power of the test may be affected by this choice. Second, the test only makes sense if $\boldsymbol{\xi} \neq 0$; in particular, it cannot be used in a pure random effects environment. Third, a rejection of the null does not necessarily mean that the usual FE estimator is inconsistent for $\boldsymbol{\beta}$: assumption (11) could still hold. In fact, the change in the estimate of $\boldsymbol{\beta}$ when the interaction terms are added can be indicative of whether accounting for time-varying η_t is likely to be important. But, because $\hat{\boldsymbol{\xi}}$ has been estimated under the null, the estimated $\boldsymbol{\beta}$ from (1.14) is not generally consistent.

If we want to estimate the η_t along with $\boldsymbol{\beta}$, we can impose the Mundlak assumption and estimate all parameteres, including $\boldsymbol{\xi}$, by pooled nonlinear regression or some GMM version. Or, we can use Chamberlain's (1982) less restrictive assumption. But, typically, when we want to allow arbitrary correlation between c_i and $\mathbf{x}_i$, we work directly from (9) and eliminate the c_i. There are several ways to do this. If we maintain that all η_t are different from zero then we can use a quas-differencing method to eliminat c_i. In particular, for $t \geq 2$ we can multiply the $t - 1$ equation by η_t/η_{t-1} and subtract the result from the time t equation:

$$y_{it} - (\eta_t/\eta_{t-1})y_{i,t-1} = [\mathbf{x}_{it} - (\eta_t/\eta_{t-1})\mathbf{x}_{i,t-1}]\boldsymbol{\beta} + [\eta_t c_i - (\eta_t/\eta_{t-1})\eta_{t-1}c_i] + [u_{it} - (\eta_t/\eta_{t-1})u_{i,t-1}]$$
$$= [\mathbf{x}_{it} - (\eta_t/\eta_{t-1})\mathbf{x}_{i,t-1}]\boldsymbol{\beta} + [u_{it} - (\eta_t/\eta_{t-1})u_{i,t-1}], \quad t \geq 2.$$

We define $\theta_t = \eta_t/\eta_{t-1}$ and write

$$y_{it} - \theta_t y_{i,t-1} = (\mathbf{x}_{it} - \theta_t \mathbf{x}_{i,t-1})\boldsymbol{\beta} + e_{it}, \; t = 2, \ldots, T, \tag{2.15}$$

where $e_{it} \equiv u_{it} - \theta_t u_{i,t-1}$. Under the strict exogeneity assumption, e_{it} is uncorrelated with every element of $\mathbf{x}_i$, and so we can apply GMM to (2.15) to estimate $\boldsymbol{\beta}$ and $(\theta_2, \ldots, \theta_T)$. Again, this requires using nonlinear GMM methods, and the e_{it} would typically be serially correlated. If we do not impose restrictions on the second moment matrix of $\mathbf{u}_i$, then we would not use any information on the second moments of $\mathbf{e}_i$; we would (eventually) use an unrestricted weighting matrix after an initial estimation.

Using all of $\mathbf{x}_i$ in each time period can result in too many overidentifying restrictions. At time t we might use, say, $\mathbf{z}_{it} = (\mathbf{x}_{it}, \mathbf{x}_{i,t-1})$, and then the instrument matrix $\mathbf{Z}_i$ (with $T-1$ rows) would be $\text{diag}(\mathbf{z}_{i2}, \ldots, \mathbf{z}_{iT})$. An initial consistent estimator can be gotten by choosing weighting matrix $(N^{-1}\sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{Z}_i)^{-1}$. Then the optimal weighting matrix can be estimated. Ahn, Lee, and Schmidt (2002) provide further discussion.

If $\mathbf{x}_{it}$ contains sequentially but not strictly exogenous explanatory variables – such as a lagged dependent variable – the instruments at time t can only be chosen from $(\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1})$. Holtz-Eakin, Newey, and Rosen (1988) explicitly consider models with lagged dependent variables; more on these models later.

Other transformations can be used. For example, at time $t \geq 2$ we can use the equation

$$\eta_{t-1}y_{it} - \eta_t y_{i,t-1} = (\eta_{t-1}\mathbf{x}_{it} - \eta_t \mathbf{x}_{i,t-1})\boldsymbol{\beta} + e_{it}, \; t = 2, \ldots, T,$$

where now $e_{it} = \eta_{t-1}u_{it} - \eta_t u_{i,t-1}$. This equation has the advantage of allowing $\eta_t = 0$ for some t. The same choices of instruments are available depending on whether $\{\mathbf{x}_{it}\}$ are strictly or sequentially exogenous.

2.2. Fixed Effects IV Estimation with Random Slopes

The results for the fixed effects estimator (in the generalized sense of removing unit-specific means and possibly trends), extend to fixed effects IV methods, provided we add a constant conditional covariance assumption. Murtazashvili and Wooldridge (2007) derive a simple set of sufficient conditions. In the model with general trends, we assume the natural extension of Assumption FEIV.1, that is, $\text{E}(u_{it}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i) = 0$ for all t, along with Assumption FEIV.2. We modify assumption (2.6) in the obvious way: replace $\ddot{\mathbf{x}}_{it}$ with $\ddot{\mathbf{z}}_{it}$, the

invididual-specific detrended instruments:

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1,\ldots,T \tag{2.16}$$

But something more is needed. Murtazashvili and Wooldridge (2007) show that, along with the previous assumptions, a sufficient condition is

$$\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i), t = 1,\ldots,T. \tag{2.17}$$

Note that the covariance $\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i)$, a $K \times K$ matrix, need not be zero, or even constant across time. In other words, we can allow the detrended covariates to be arbitrarily correlated with the heterogeneous slopes, and that correlation can change in any way across time. But the *conditional* covariance cannot depend on the time-demeaned instruments. (This is an example of how it is important to distinguish between a conditional expectation and an unconditional one: the implicit error in the equation generally has an unconditional mean that changes with *t*, but its conditional mean does not depend on $\ddot{\mathbf{z}}_{it}$, and so using $\ddot{\mathbf{z}}_{it}$ as IVs is valid provided we allow for a full set of dummies.) Condition (2.17) extends to the panel data case the assumption used by Wooldridge (2003a) in the cross section case.

We can easily show why (2.17) suffices with the previous assumptions. First, if $\mathrm{E}(\mathbf{d}_i|\ddot{\mathbf{z}}_{it}) = \mathbf{0}$ – which follows from $\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\mathbf{b}_i)$ – then $\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{d}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\ddot{\mathbf{x}}_{it}\mathbf{d}_i'|\ddot{\mathbf{z}}_{it})$, and so $\mathrm{E}(\ddot{\mathbf{x}}_{it}\mathbf{d}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\ddot{\mathbf{x}}_{it}\mathbf{d}_i) \equiv \gamma_t$ under the previous assumptions. Write $\ddot{\mathbf{x}}_{it}\mathbf{d}_i = \gamma_t + r_{it}$ where $\mathrm{E}(r_{iti}|\ddot{\mathbf{z}}_{it}) = 0, t = 1,\ldots,T$. Then we can write the transformed equation as

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{\mathbf{x}}_{it}\mathbf{d}_i + \ddot{u}_{it} = \ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \gamma_t + r_{it} + \ddot{u}_{it}. \tag{2.18}$$

Now, if $\mathbf{x}_{it}$ contains a full set of time period dummies, then we can absorb γ_t into $\ddot{\mathbf{x}}_{it}$, and we assume that here. Then the sufficient condition for consistency of IV estimators applied to the transformed equations is $\mathrm{E}[\ddot{\mathbf{z}}_{it}'(r_{it} + \ddot{u}_{it})] = \mathbf{0}$,.and this condition is met under the maintained assumptions. In other words, under (2.16) and (2.17), the fixed effects 2SLS estimator is consistent for the average population effect, $\boldsymbol{\beta}$. (Remember, we use "fixed effects" here in the general sense of eliminating the unit-specific trends, $\mathbf{a}_i$.) We must remember to include a full set of time period dummies if we want to apply this robustness result, something that should be done in any case. Naturally, we can also use GMM to obtain a more efficient estimator. If $\mathbf{b}_i$ truly depends on *i*, then the composite error $r_{it} + \ddot{u}_{it}$ is likely serially correlated and heteroskedastic. See Murtazashvili and Wooldridge (2007) for further discussion and similation results on the peformance of the FE2SLS estimator. They also provide examples where the key assumptions cannot be expected to hold, such as when endogenous elements of

$\mathbf{x}_{it}$ are discrete.

3. Behavior of Estimators without Strict Exogeneity

As is well known, both the FE and FD estimators are inconsistent (with fixed T, $N \to \infty$) without the conditional strict exogeneity assumption. But it is also pretty well known that, at least under certain assumptions, the FE estimator can be expected to have less "bias" (actually, inconsistency) for larger T. One assumption is contemporaneous exogeneity, (1.2). If we maintain this assumption, assume that the data series $\{(\mathbf{x}_{it}, u_{it}) : t = 1, \ldots, T\}$ is "weakly dependent" – in time series parlance, integrated of order zero, or I(0) – then we can show that

$$\text{plim } \hat{\boldsymbol{\beta}}_{FE} = \boldsymbol{\beta} + O(T^{-1}) \tag{3.1}$$

$$\text{plim } \hat{\boldsymbol{\beta}}_{FD} = \boldsymbol{\beta} + O(1). \tag{3.2}$$

In some special cases – the AR(1) model without extra covariates – the "bias" terms can be calculated. But not generally. The FE (within) estimator averages across T, and this tends to reduce the bias.

Interestingly, the same results can be shown if $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ has unit roots as long as $\{u_{it}\}$ is I(0) and contemporaneous exogeneity holds. But there is a catch: if $\{u_{it}\}$ is I(1) – so that the time series version of the "model" would be a spurious regression (y_{it} and $\mathbf{x}_{it}$ are not cointegrated), then (3.1) is no longer true. And, of course, the first differencing means any unit roots are eliminated. So, once we start appealing to "large T" to prefer FE over FD, we must start being aware of the time series properties of the series.

The same comments hold for IV versions of the estimators. Provided the instruments are contemporaneously exogenous, the FEIV estimator has bias of order T^{-1}, while the bias in the FDIV estimator does not shrink with T. The same caveats about applications to unit root processes also apply.

Because failure of strict exogeneity causes inconsistency in both FE and FD estimation, it is useful to have simple tests. One possibility is to obtain a Hausman test directly comparing the FE and FD estimators. This is a bit cumbersome because, when aggregate time effects are included, the difference in the estimators has a singular asymptotic variance. Plus, it is somewhat difficult to make the test fully robust.

Instead, simple regression-based strategies are available. Let $\mathbf{w}_{it}$ be the $1 \times Q$ vector, a subset of $\mathbf{x}_{it}$ suspected of failing strict exogeneity. A simple test of strict exogeneity, specifically looking for feedback problems, is based on

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{i,t+1}\boldsymbol{\delta} + c_i + e_{it}, \; t = 1, \ldots, T-1. \tag{3.3}$$

Estimate the equation by fixed effects and test $H_0 : \boldsymbol{\delta} = \mathbf{0}$ (using a fully robust test). Of course, the test may have little power for detecting contemporaneous endogeneity.

In the context of FEIV we can test whether a subset of instruments fails strict exogeneity by writing

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{h}_{i,t+1}\boldsymbol{\delta} + c_i + e_{it}, \; t = 1, \ldots, T-1, \tag{3.4}$$

where $\mathbf{h}_{it}$ is a subset of the instruments, $\mathbf{z}_{it}$. Now, estimate the equation by FEIV using instruments $(\mathbf{z}_{it}, \mathbf{h}_{i,t+1})$ and test coefficients on the latter.

It is also easy to test for contemporaneous endogeneity of certain regressors, even if we allow some regressors to be endogenous under the null. Write the model now as

$$y_{it1} = \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + \mathbf{y}_{it3}\boldsymbol{\gamma}_1 + c_{i1} + u_{it1}, \tag{3.5}$$

where, in an FE environment, we want to test $\mathrm{H}_0 : \mathrm{E}(\mathbf{y}_{it3}' u_{it1}) = \mathbf{0}$. Actually, because we are using the within transformation, we are really testing strict exogeneity of $\mathbf{y}_{it3}$, but we allow all variables to be correlated with c_{i1}. The variables $\mathbf{y}_{it2}$ are allowed to be endogenous under the null – provided, of course, that we have sufficient instruments excluded from the structural equation that are uncorrelated with u_{it1} in every time period. We can write a set of reduced forms for elements of $\mathbf{y}_{it3}$ as

$$\mathbf{y}_{it3} = \mathbf{z}_{it}\boldsymbol{\Pi}_3 + \mathbf{c}_{i3} + \mathbf{v}_{it3}, \tag{3.6}$$

and obtain the FE residuals, $\ddot{\hat{\mathbf{v}}}_{it3} = \ddot{\mathbf{y}}_{it3} - \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\Pi}}_3$, where the columns of $\hat{\boldsymbol{\Pi}}_3$ are the FE estimates of the reduced forms, and the double dots denotes time-demeaning, as usual. Then, estimate the equation

$$\ddot{y}_{it1} = \ddot{\mathbf{z}}_{it1}\boldsymbol{\delta}_1 + \ddot{\mathbf{y}}_{it2}\boldsymbol{\alpha}_1 + \ddot{\mathbf{y}}_{it3}\boldsymbol{\gamma}_1 + \ddot{\hat{\mathbf{v}}}_{it3}\boldsymbol{\rho}_1 + error_{it1} \tag{3.7}$$

by pooled IV, using instruments $(\ddot{\mathbf{z}}_{it}, \ddot{\mathbf{y}}_{it3}, \ddot{\hat{\mathbf{v}}}_{it3})$. The test of the null that $\mathbf{y}_{it3}$ is exogenous is just the (robust) test that $\boldsymbol{\rho}_1 = \mathbf{0}$, and the usual robust test is valid with adjusting for the first-step estimation.

An equivalent approach is to define $\hat{\mathbf{v}}_{it3} = \mathbf{y}_{it3} - \mathbf{z}_{it}\hat{\boldsymbol{\Pi}}_3,$ where $\hat{\boldsymbol{\Pi}}_3$ is still the matrix of FE coefficients, add these to equation (3.5), and apply FE-IV, using a fully robust test. Using a built-in command can lead to problems because the test is rarely made robust and the degrees of freedom are often incorrectly counted.

4. Instrumental Variables Estimation under Sequential Exogeneity

We now consider IV estimation of the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it},\ t = 1,\ldots,T, \tag{4.1}$$

under sequential exogeneity assumptions. Some authors simply use

$$E(\mathbf{x}_{is}' u_{it}) = 0,\ s = 1,\ldots,T, t = 1,\ldots,T. \tag{4.2}$$

As always, $\mathbf{x}_{it}$ probably includes a full set of time period dummies. This leads to simple moment conditions after first differencing:

$$\mathrm{E}(\mathbf{x}_{is}'\Delta u_{it}) = \mathbf{0},\ s = 1,\ldots,t-1;\ t = 2,\ldots,T. \tag{4.3}$$

Therefore, at time t, the available instruments in the FD equation are in the vector $\mathbf{x}_{i,t-1}^o$, where

$$\mathbf{x}_{it}^o \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it}). \tag{4.4}$$

Therefore, the matrix of instruments is simply

$$\mathbf{W}_i = \mathrm{diag}(\mathbf{x}_{i1}^o, \mathbf{x}_{i2}^o, \ldots, \mathbf{x}_{i,T-1}^o), \tag{4.5}$$

which has $T-1$ rows. Because of sequential exogeneity, the number of valid instruments increases with t.

Given $\mathbf{W}_i$, it is routine to apply GMM estimation. But some simpler strategies are available that can be used for comparison or as the first-stage estimator in computing the optimal weighting matrix. One useful one is to estimate a reduced form for $\Delta\mathbf{x}_{it}$ separately for each t. So, at time t, run the regression $\Delta\mathbf{x}_{it}$ on $\mathbf{x}_{i,t-1}^o$, $i = 1,\ldots,N$, and obtain the fitted values, $\widehat{\Delta\mathbf{x}}_{it}$. Of course, the fitted values are all $1 \times K$ vectors for each t, even though the number of available instruments grows with t. Then, estimate the FD equation

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it},\ t = 2,\ldots,T \tag{4.6}$$

by pooled IV using instruments (not regressors) $\widehat{\Delta\mathbf{x}}_{it}$. It is simple to obtain robust standard errors and test statistics from such a procedure because the first stage estimation to obtain the instruments can be ignored (asymptotically, of course).

One potential problem with estimating the FD equation by IVs that are simply lags of $\mathbf{x}_{it}$ is that changes in variables over time are often difficult to predict. In other words, $\Delta\mathbf{x}_{it}$ might have little correlation with $\mathbf{x}_{i,t-1}^o$, in which case we face a problem of weak instruments. In one case, we even lose identification: if $\mathbf{x}_{it} = \boldsymbol{\lambda}_t + \mathbf{x}_{i,t-1} + \mathbf{e}_{it}$ where $\mathrm{E}(\mathbf{e}_{it}|\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}) = \mathbf{0}$ – that is, the elements of $\mathbf{x}_{it}$ are random walks with drift – then $\mathrm{E}(\Delta\mathbf{x}_{it}|\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}) = \mathbf{0}$, and the rank condition for IV estimation fails.

If we impose what is generally a stronger assumption, **dynamic completeness in the conditional mean**,

$$\mathrm{E}(u_{it}|\mathbf{x}_{it}, y_{i,t-1}\mathbf{x}_{i,t-1}, \ldots, y_{i1}, \mathbf{x}_{i1}, c_i) = 0, \quad t = 1, \ldots, T, \tag{4.7}$$

then more moment conditions are available. While (4.7) implies that virtually any nonlinear function of the $\mathbf{x}_{it}$ can be used as instruments, the focus has been only on zero covariance assumptions (or (4.7) is stated as a linear projection). The key is that (4.7) implies that $\{u_{it} : t = 1, \ldots, T\}$ is a serially uncorrelated sequence and u_{it} is uncorrelated with c_i for all t. If we use these facts, we obtain moment conditions first proposed by Ahn and Schmidt (1995) in the context of the AR(1) unosberved effects model; see also Arellano and Honoré (2001). They can be written generally as

$$E[(\Delta y_{i,t-1} - \Delta \mathbf{x}_{i,t-1}\boldsymbol{\beta})'(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}, t = 3, \ldots, T. \tag{4.8}$$

Why do these hold? Because all u_{it} are uncorrelated with c_i, and $\{u_{i,t-1}, \ldots, u_{i1}\}$ are uncorrelated with $c_i + u_{it}$. So $(u_{i,t-1} - u_{i,t-2})$ is uncorrelated with $(c_i + u_{it})$, and the resulting moment conditions can be written in terms of the parameters as (4.8). Therefore, under (4.7), we can add the conditions (4.8) to (4.3) to improve efficiency – in some cases quite substantially with persistent data.

Of course, we do not always intend for models to be dynamically complete in the sense of (4.7). Often, we estimate static models or finite distributed lag models – that is, models without lagged dependent variables – that have serially correlated idiosyncratic errors, and the explanatory variables are not strictly exogenous and so GLS procedures are inconsistent. Plus, the conditions in (4.8) are nonlinear in parameters.

Arellano and Bover (1995) suggested instead the restrictions

$$Cov(\Delta \mathbf{x}_{it}', c_i) = 0, \ t = 2, \ldots, T. \tag{4.9}$$

Interestingly, this is zero correlation, FD version of the conditions from Section 2 that imply we can ignore heterogeneous coefficients in estimation under strict exogeneity. Under (4.9), we have the moment conditions from the levels equation:

$$E[\Delta \mathbf{x}_{it}'(y_{it} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}, t = 2, \ldots, T, \tag{4.10}$$

because $y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} = c_i + u_{it}$ and u_{it} is uncorrelated with $\mathbf{x}_{it}$ and $\mathbf{x}_{i,t-1}$. We add an intercept, α, explicitly to the equation to allow a nonzero mean for c_i. Blundell and Bond (1999) apply these moment conditions, along with the usual conditions in (4.3), to estimate firm-level production functions. Because of persistence in the data, they find the moments in (4.3) are not

especially informative for estimating the parameters. Of course, (4.9) is an extra set of assumptions.

The previous discussion can be applied to the AR(1) model, which has received much attention. In its simplest form we have

$$y_{it} = \rho y_{i,t-1} + c_i + u_{it}, t = 1, \ldots, T, \tag{4.11}$$

so that, by convention, our first observation on y is at $t = 0$. Typically the minimal assumptions imposed are

$$E(y_{is}u_{it}) = 0,\ s = 0, \ldots, t-1,\ t = 1, \ldots, T, \tag{4.12}$$

in which case the available instruments at time t are $\mathbf{w}_{it} = (y_{i0}, \ldots, y_{i,t-2})$ in the FD equation

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta u_{it}, t = 2, \ldots, T. \tag{4.13}$$

In oher words, we can use

$$E[y_{is}(\Delta y_{it} - \rho \Delta y_{i,t-1}) = 0,\ s = 0, \ldots, t-2,\ t = 2, \ldots, T. \tag{4.14}$$

Anderson and Hsiao (1982) proposed pooled IV estimation of the FD equation with the single instrument $y_{i,t-2}$ (in which case all $T-1$ periods can be used) or $\Delta y_{i,t-2}$ (in which case only $T-2$ periods can be used). We can use pooled IV where $T-1$ separate reduced forms are estimated for $\Delta y_{i,t-1}$ as a linear function of $(y_{i0}, \ldots, y_{i,t-2})$. The fitted values $\widehat{\Delta y}_{i,t-1}$, can be used as the instruments in (4.13) in a pooled IV estimation. Of course, standard errors and inference should be made robust to the MA(1) serial correlation in Δu_{it}. Arellano and Bond (1991) suggested full GMM estimation using all of the available instruments $(y_{i0}, \ldots, y_{i,t-2})$, and this estimator uses the conditions in (4.12) efficiently.

Under the dynamic completeness assumption

$$\mathrm{E}(u_{it}|y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}, c_i) = 0, \tag{4.15}$$

the Ahn-Schmidt extra moment conditions in (4.8) become

$$E[(\Delta y_{i,t-1} - \rho \Delta y_{i,t-2})(y_{it} - \rho y_{i,t-1})] = 0, t = 3, \ldots, T. \tag{4.16}$$

Blundell and Bond (1998) noted that if the condition

$$Cov(\Delta y_{i1}, c_i) = Cov(y_{i1} - y_{i0}, c_i) = 0 \tag{4.17}$$

is added to (4.15) then the combinded set of moment conditions becomes

$$E[\Delta y_{i,t-1}(y_{it} - \alpha - \rho y_{i,t-1})] = 0, t = 2, \ldots, T, \tag{4.18}$$

which can be added to the usual moment conditions (4.14). Therefore, we have two sets of

moments linear in the parameters. The first, (4.14), use the differenced equation while the second, (4.18), use the levels. Arellano and Bover (1995) analyzed GMM estimators from these equations generally.

As discussed by Blundell and Bond (1998), condition (4.17) can be intepreted as a restriction on the initial condition, y_{i0}. To see why, write $y_{i1} - y_{i0} = \rho y_{i0} + c_i + u_{i1} - y_{i0} = (1 - \rho)y_{i0} + c_i + u_{i1}$. Because u_{i1} is uncorrelated with c_i, (4.17) becomes

$$Cov((1 - \rho)y_{i0} + c_i, c_i) = 0. \tag{4.19}$$

Write y_{i0} as a deviation from its steady state, $c_i/(1 - \rho)$ (obtained for $|\rho| < 1$ be recursive subsitution and then taking the limit), as

$$y_{i0} = c_i/(1 - \rho) + r_{i0}. \tag{4.20}$$

Then $(1 - \rho)y_{i0} + c_i = (1 - \rho)r_{i0}$, and so (4.17) reduces to

$$Cov(r_{i0}, c_i) = 0. \tag{4.21}$$

In other words, the deviation of y_{i0} from its steady state is uncorrelated with the steady state. Blundell and Bond (1998) contains discussion of when this condition is reasonable. Of course, it is not for $\rho = 1$, and it may not be for ρ "close" to one. On the other hand, as shown by Blundell and Bond (1998), this restriction, along with the Ahn-Schmidt conditions, is very informative for ρ close to one. Hahn (1999) shows theoretically that such restrictions can greatly increase the information about ρ.

The Ahn-Schmidt conditions (4.16) are attractive in that they are implied by the most natural statement of the model, but they are nonlinear and therefore more difficult to use. By adding the restriction on the initial condition, the extra moment condition also means that the full set of moment conditions is linear. Plus, this approach extends to general models with only sequentially exogenous variabes as in (4.10). Extra moment assumptions based on homoskedasticity assumptions – either conditional or unconditional – have not been used nearly as much, probably because they impose conditions that have little if anything to do with the economic hypotheses being tested.

Other approaches to dynamic models are based on maximum likelihood estimation or generalized least squares estimation of a particular set of conditional means. Approaches that condition on the initial condition y_{i0}, an approach suggested by Chamberlain (1980), Blundell and Smith (1991), and Blundell and Bond (1998), seem especially attractive. For example,

suppose we assume that

$$D(y_{it}|y_{i,t-1},y_{i,t-2},\ldots,y_{i1},y_{i0},c_i) = \text{Normal}(\rho y_{i,t-1} + c_i,\sigma_u^2),\ t = 1,2,\ldots,T.$$

Then the distribution of $(y_{i1},\ldots,y_{iT})$ given $(y_{i0} = y_0, c_i = c)$ is just the product of the normal distributions:

$$\prod_{t=1}^{T} \sigma_u^{-T}\phi[(y_t - \rho y_{t-1} - c)/\sigma_u].$$

We can obtain a usable density for (conditional) MLE by assuming

$$c_i|y_{i0} \sim \text{Normal}(\varphi_0 + \xi_0 y_{i0},\sigma_a^2).$$

The log likelihood function is obtained by taking the log of

$$\int_{-\infty}^{\infty}\left(\prod_{t=1}^{T}(1/\sigma_u)^T\phi[(y_{it} - \rho y_{i,t-1} - c)/\sigma_u].\right)(1/\sigma_a)\phi[(c - \varphi_0 - \xi_0 y_{i0})/\sigma_a)dc.$$

Of course, if this is this represents the correct density of $(y_{i1},\ldots,y_{iT})$ given y_{i0} then the MLE is consistent and $\sqrt{N}$-asymptotically normal (and efficient among estimators that condition on y_{i0}).

A more robust approach is to use a generalized least squares approach where $E(\mathbf{y}_i|y_{i0})$ and $Var(\mathbf{y}_i|y_{i0})$ are obtained, and where the latter could even be misspecified. Like with the MLE approach, this results in estimation that is highly nonlinear in the parameters and is used less often than the GMM procedures with linear moment conditions.

The same kinds of moment conditions can be used in extensions of the AR(1) model, such as

$$y_{it} = \rho y_{i,t-1} + \mathbf{z}_{it}\boldsymbol{\gamma} + c_i + u_{it},\ \ t = 1,\ldots,T.$$

If we difference to remove c_i, we can then use exogeneity assumptions to choose instruments. The FD equation is

$$\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta\mathbf{z}_{it}\boldsymbol{\gamma} + \Delta u_{it},\ \ t = 1,\ldots,T,$$

and if the $\mathbf{z}_{it}$ are strictly exogenous with respect to $\{u_{i1},\ldots,u_{iT}\}$ then the available instruments (in addition to time period dummies) are $(\mathbf{z}_i, y_{i,t-2},\ldots,y_{i0})$. We might not want to use all of $\mathbf{z}_i$ for every time period. Certainly we would use $\Delta\mathbf{z}_{it}$, and perhaps a lag, $\Delta\mathbf{z}_{i,t-1}$. If we add sequentially exogenous variables, say $\mathbf{h}_{it}$, to (11.62) then $(\mathbf{h}_{i,t-1},\ldots,\mathbf{h}_{i1})$ would be added to the list of instruments (and $\Delta\mathbf{h}_{it}$ would appear in the equation). We might also add the Arellano

and Bover conditions (4.10), or at least the Ahn and Schmidt conditions (4.8).

As a simple example of methods for dynamic models, consider a dynamic air fare equation for routes in the United States:

$$lfare_{it} = \theta_t + \rho\, lfare_{i,t-1} + \gamma\, concen_{it} + c_i + u_{it},$$

where we include a full set of year dummies. We assume the concentration ratio, $concen_{it}$, is strictly exogenous and that at most one lag of *lfare* is needed to capture the dynamics. The data are for 1997 through 2000, so the equation is specified for three years. After differencing, we have only two years of data:

$$\Delta lfare_{it} = \eta_t + \rho \Delta lfare_{i,t-1} + \gamma \Delta concen_{it} + \Delta u_{it},\ t = 1999, 2000.$$

If we estimate this equation by pooled OLS, the estimators are inconsistent because $\Delta lfare_{i,t-1}$ is correlated with Δu_{it}; we include the OLS estimates for comparison. We apply the simple pooled IV procedure, where separate reduced forms are estimated for $\Delta lfare_{i,t-1}$: one for 1999, with $lfare_{i,t-2}$ and $\Delta concen_{it}$ in the reduced form, and one for 2000, with $lfare_{i,t-2}$, $lfare_{imt-3}$ and $\Delta concen_{it}$ in the reduced form. The fitted values are used in the pooled IV estimation, with robust standard errors. (We only use $\Delta concen_{it}$ in the IV list at time *t*.) Finally, we apply the Arellano and Bond (1991) GMM procedure.

Dependent Variable:	*lfare*		
	(1)	(2)	(3)
Explanatory Variable	Pooled OLS	Pooled IV	Arellano-Bond
$lfare_{-1}$	−.126	.219	.333
	(.027)	(.062)	(.055)
concen	.076	.126	.152
	(.053)	(.056)	(.040)
N	1,149	1,149	1,149

As is seen from column (1), the pooled OLS estimate of ρ is actually negative and statistically different from zero. By contrast, the two IV methods give positive and statistically significant estimates. The GMM estimate of ρ is larger, and it also has a smaller standard error (as we would hope for GMM).

The previous example has small *T*, but some panel data applications have reasonable large *T*. Arellano and Alvarez (1998) show that the GMM estimator that accounts for the MA(1) serial correlation in the FD errors has desirable properties when *T* and *N* are both large, while

the pooled IV estimator is actually inconsistent under asymptotics where $T/N \to a > 0$. See Arellano (2003, Chapter 6) for discussion.

5. Pseudo Panels from Pooled Cross Sections

In cases where panel data sets are not available, we can still estimate parameters in an underlying panel population model if we can obtain random samples in different periods. Many surveys are done annually by obtaining a different random (or stratified) sample for each year. Deaton (1985) showed how to identify and estimate parameters in panel data models from pooled cross sections. As we will see, however, identification of the parameterse can be tenuous.

Deaton (1985) was careful about distinguishing between the population model on the one hand and the sampling scheme on the other. This distinction is critical for understanding the nature of the identification problem, and in deciding the appropriate asymptotic analysis. The recent literature has tended to write "models" at the cohort or group level, which is not in the spirit of Deaton's original work. (Angrist (1991) actually has panel data, but uses averages in each t to estimate parameters of a labor supply function.)

In what follows, we are interested in estimating the parameters of the population model

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f + u_t, \; t = 1, \ldots, T, \tag{5.1}$$

which is best viewed as representing a population defined over T time periods. For this setup to make sense, it must be the case that we can think of a stationary population, so that the same units are represented in each time period. Because we allow a full set of period intercepts, $E(f)$ is never separately identified, and so we might as well set it to zero.

The random quantities in (5.1) are the response variable, y_t, the covariates, $\mathbf{x}_t$ (a $1 \times K$ vector), the unobserved effect, f, and the unobserved idiosyncratic errors, $\{u_t : t = 1, \ldots, T\}$. Like our previous analysis, we are thinking of applications with a small number of time periods, and so we view the intercepts, η_t, as parameters to estimate, along with the $K \times 1$ vector parameter – which is ultimately of interest. We consider the case where all elements of $\mathbf{x}_t$ have some time variation.

As it turns out, to use the standard analysis, we do not even have to assume contemporaneous exogeneity conditional on f, that is,

$$E(u_t|\mathbf{x}_t, f) = 0, t = 1, \ldots, T, \tag{5.2}$$

although this is a good starting point to determine reasonable population assumptions.

Naturally, iterated expectations implies

$$E(u_t|f) = 0, t = 1, \ldots, T, \tag{5.3}$$

and (5.3) is sensible in the context of (5.1). Unless stated otherwise, we take it to be true. Because f aggregates all time-constant unobservables, we should think of (5.3) as implying that $E(u_t|g) = 0$ for any time-constant variable g, whether unobserved or observed. In other words, in the leading case we should think of (5.1) as representing $E(y_t|\mathbf{x}_t, f)$ where any time constant factors are lumped into f.

With a (balanced) panel data set, we would have a random sample in the cross section. Therefore, for a random draw i, $\{(\mathbf{x}_{it}, y_{it}), t = 1, \ldots, T\}$, we would then write the model as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + f_i + u_{it}, \; t = 1, \ldots, T. \tag{5.4}$$

While this notation can cause confusion later when we sample from each cross section, it has the benefit of explictly labelling quantities as changing only across t, changing only across i, or changing across both.

The idea of using independent cross sections to estimate parameters from panel data models is based on a simple insight of Deaton's. Assume that the population for which (5.1) holds is divided into G groups (or cohorts). This designation cannot depend on time. For example, it is common to birth year to define the groups, or even ranges of birth year. For a random draw i satisfying (5.4), let g_i be the group indicator, taking on a value in $\{1, 2, \ldots, G\}$. Then, by our earlier discussion,

$$E(u_{it}|g_i) = 0, t = 1, \ldots, T, \tag{5.5}$$

essentially by definition. In other words, the η_t account for any change in the average unobservables over time and f_i accounts for any time-constant factors.

Taking the expected value of (5.4) conditional on group membership and using only (5.5), we have

$$E(y_{it}|g_i = g) = \eta_t + E(\mathbf{x}_{it}|g_i = g)\boldsymbol{\beta} + E(f_i|g_i = g), \; t = 1, \ldots, T. \tag{5.6}$$

Again, this expession represents an underlying population, but where we have partitioned the population into G groups.

Several authors after Deaton, including Collado (1997) and Verbeek and Vella (2005), have left $E(u_{it}|g_i = g)$ as part of the "error term," with the notation $u_{gt}^* = E(u_{it}|g_i = g)$. In fact, these authors have criticized previous work by Moffitt (1993) for making the "asssumption" that $u_{gt}^* = 0$. But, as Deaton showed, if we start with the underlying population model (5.1),

then $E(u_{it}|g_i = g) = 0$ for all g follows directly. Nevertheless, as we will discuss later, the key assumption is that the structural model (5.1) does not require a full set of group/time effects. If such effects are required, then one way to think about the resulting misspecification is that $E(u_{it}|g_i = g)$ is not zero.

If we define the population means

$$\begin{aligned} \alpha_g &= E(f_i|g_i = g) \\ \mu_{gt}^y &= E(y_{it}|g_i = g) \\ \boldsymbol{\mu}_{gt}^{\mathbf{x}} &= E(\mathbf{x}_{it}|g_i = g) \end{aligned} \tag{5.7}$$

for $g = 1, \ldots, G$ and $t = 1, \ldots, T$ we have

$$\mu_{gt}^y = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g, \, g = 1, \ldots, G, \; t = 1, \ldots, T. \tag{5.8}$$

(Many authors use the notation y_{gt}^* in place of μ_{gt}^y, and similarly for $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$, but, at this point, such a notation gives the wrong impression that the means defined in (5.7) are random variables. They are not. They are group/time means defined on the underlying population.)

Equation (5.8) is remarkable in that it holds without any assumptions restricting the dependence between $\mathbf{x}_{it}$ and u_{ir} across t and r. In fact, $\mathbf{x}_{it}$ can contain lagged dependent variables, most commonly $y_{i,t-1}$, or explanatory variables that are contemporaneously endogenous (as occurs under measurement error in the original population model, an issue that was important to Angrist (1991)). This probably should make us a little suspicious, as the problems of lagged dependent variable, measurement error, and other violations of strict exogeneity are tricky to handle with true panel data.

(In estimation, we will deal with the fact that there are not really $T + G$ parameters in η_t and α_g to estimate; there are only $T + G - 1$. The lost degree of freedom comes from $E(f) = 0$, which puts a restriction on the α_g. With the groups of the same size in the population, the restriction is that the α_g sum to zero.)

If we take (5.8) as the starting point for estimating $\boldsymbol{\beta}$ (along with η_t and α_g), then the issues become fairly clear. If we have sufficient observations in the group/time cells, then the means μ_{gt}^y and $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ can be estimated fairly precisly, and these can be used in a minimum distance estimation framework to estimate $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ consists of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\alpha}$ (where, say, we set $\eta_1 = 0$ as the normalization).

Before discussing estimation details, it is useful to study (5.8) in more detail to determine some simple, and common, strategies. Because (5.8) looks itself like a panel data regression

equation, methods such as "OLS," "fixed effects," and "first differencing" have been applied to sample averages. It is informative to apply these to the population. First suppose that we set each α_g to zero and set all of the time intercepts, η_t, to zero. For notational simplicity, we also drop an overall "intercept," but that would be included at a minimum. Then $\mu_{gt}^{y} = \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta}$ and if we premultiply by $\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}$, average across g and t, and then assume we can invert $\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}\boldsymbol{\mu}_{gt}^{\mathbf{x}}$, we have

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}\boldsymbol{\mu}_{gt}^{\mathbf{x}}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}\mu_{gt}^{y}\right). \tag{5.9}$$

This means that the population parameter, $\boldsymbol{\beta}$, can be written as a pooled OLS regression of the population group/time means μ_{gt}^{y} on the group/time means $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$. Naturally, if we have "good" estimates of these means, then it will make sense to estimate $\boldsymbol{\beta}$ by using the same regression on the sample means. But, so far, this is all in the population. We can think of (5.9) as the basis for a method of moments procedure. It is important that we treat $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ and μ_{gt}^{y} symetrically, that is, as population means to be estimated, whether the $\mathbf{x}_{it}$ are strictly, sequentially, or contemporaneous exogenous – or none of these – in the original model.

When we allow different group means for f_i, as seems critical, and different time period intercepts, which also is necessary for a convincing analysis, we can easily write $\boldsymbol{\beta}$ as an "OLS" estimator by subtracting of time and group averages. While we cannot claim that these expressions will result in efficient estimators, they can shed light on whether we can expect (5.8) to lead to precise estimation of $\boldsymbol{\beta}$. First, without separate time intercepts we have

$$\mu_{gt}^{y} - \bar{\mu}_{g}^{y} = (\boldsymbol{\mu}_{gt}^{\mathbf{x}} - \bar{\boldsymbol{\mu}}_{g}^{\mathbf{x}})\boldsymbol{\beta}, g = 1,\ldots,G,; t = 1,\ldots,T, \tag{5.10}$$

where the notation should be clear, and then one expression for $\boldsymbol{\beta}$ is (5.9) but with $\boldsymbol{\mu}_{gt}^{\mathbf{x}} - \bar{\boldsymbol{\mu}}_{g}^{\mathbf{x}}$ in place of $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$. Of course, this makes it clear that identification of $\boldsymbol{\beta}$ more difficult when the α_g are allowed to differ. Further, if we add in the year intercepts, we have

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G}\sum_{t=1}^{T}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{t=1}^{T}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}\mu_{gt}^{y}\right) \tag{5.11}$$

where $\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$ is the vector of residuals from the pooled regression

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} \text{ on } 1, d2,\ldots,dT, c2, \ldots, cG, \tag{5.12}$$

where *dt* denotes a dummy for period *t* and *cg* is a dummy variable for group *g*.

There are other expressions for $\boldsymbol{\beta}$, too. (Because $\boldsymbol{\beta}$ is generally overidentified, there are many ways to write it in terms of the population moments. For example, if we difference and then take away group averages, we have

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G} \sum_{t=2}^{T} \Delta\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \Delta\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=2}^{T} \Delta\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \Delta\mu_{gt}^{y} \right) \tag{5.13}$$

where $\Delta\boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\mu}_{gt}^{\mathbf{x}} - \boldsymbol{\mu}_{g,t-1}^{\mathbf{x}}$ and $\Delta\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} = \Delta\boldsymbol{\mu}_{gt}^{\mathbf{x}} - G^{-1} \sum_{h=1}^{G} \Delta\boldsymbol{\mu}_{ht}^{\mathbf{x}}$.

Equations (5.11) and (5.13) make it clear that the underlying model in the population cannot contain a full set of group/time interactions. So, for example, if the groups (cohorts) are defined by birth year, there cannot be a full set of birth year/time period interactions. We could allow this feature with invidual-level data because we would typically have variation in the covariates within each group/period cell. Thus, the absense of full cohort/time effects in the population model is the key an identifying restriction.

Even if we exclude full group/time effects, $\boldsymbol{\beta}$ may not be precisely estimable. Clearly $\boldsymbol{\beta}$ is not identified if we can write $\boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\lambda}_t + \boldsymbol{\omega}_g$ for vectors $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}_g$, $t = 1, \ldots, T$, $g = 1, \ldots, G$. In other words, while we must exclude a full set of group/time effects in the structural model, we need some interaction between them in the distribution of the covariates. One might be worried about this way of identifying $\boldsymbol{\beta}$. But even if we accept this identification strategy, the variation in $\{\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} : t = 1, .., T, g = 1, \ldots, G\}$ or $\{\Delta\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} : t = 2, .., T, g = 1, \ldots, G\}$ might not be sufficient to learn much about $\boldsymbol{\beta}$ – even if we have pretty good estimates of the population means.

We are now ready to formally discuss estimation of $\boldsymbol{\beta}$. We have two formulas (and there are many more) that can be used directly, once we estimate the group/time means for y_t and $\mathbf{x}_t$. We can use either true panel data or repeated cross sections. Angrist (1991) used panel data and grouped the data by time period (after differencing). Our focus here is on the case where we do not have panel data, but the general discussion applies to either case. One difference is that, with independent cross sections, we need not account for dependence in the sample averages across *g* and *t* (except in the case of dynamic models).

Assume we have a random sample on $(\mathbf{x}_t, y_t)$ of size N_t, and we have specified the *G* groups or cohorts. Write $\{(\mathbf{x}_{it}, y_{it}) : i = 1, \ldots, N_t\}$. Some authors, wanting to avoid confusion with a true panel data set, prefer to replace *i* with $i(t)$ to emphasize that the cross section units are different in each time period. (Plus, several authors actually write the underlying model in

terms of the pooled cross sections rather than using the underlying population model – a mistake, in my view.) As long as we understand that we have a random sample in each time period, and that random sample is used to estimate the group/time means, there should be no confusion.

For each random draw i, it is useful to let $\mathbf{r}_i = (r_{it1}, r_{it2}, \ldots, r_{itG})$ be a vector of group indicators, so $r_{itg} = 1$ if observation i is in group g. Then the sample average on the response variable in group/time cell (g,t) can be written as

$$\mu_{gt}^y = N_{gt}^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it} = (N_{gt}/N_t)^{-1} N_t^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it}, \tag{5.14}$$

where $N_{gt} = \sum_{i=1}^{N_t} r_{itg}$ is properly treated as a random outcome. (This differs from standard stratified sampling, where the groups are first chosen and then random samples are obtained within each group (stratum). Here, we fix the groups and then randomly sample from the population, keeping track of the group for each draw.) Of course, μ_{gt}^y is generally consistent for μ_{gt}^y. First, $\hat{\rho}_{gt} = N_{gt}/N_t$ converges in probability to $\rho_g = P(r_{itg} = 1)$ – the fraction of the population in group or cohort g (which is supposed to be constant across t). So

$$\begin{aligned} &\hat{\rho}_{gt}^{-1} N_t^{-1} \sum_{i=1}^{N_t} r_{itg} y_{it} \xrightarrow{p} \rho_g^{-1} E(r_{itg} y_{it}) \\ &= \rho_g^{-1} [P(r_{itg} = 1) \cdot 0 + P(r_{itg} = 1) E(y_{it} | r_{itg} = 1)] \\ &= E(y_{it} | r_{itg} = 1) = \mu_{gt}^y. \end{aligned}$$

Naturally, the argument for other means is the same. Let $\mathbf{w}_{it}$ denote the $(K+1) \times 1$ vector $(y_{it}, \mathbf{x}_{it})'$. Then the asymptotic distribution of the full set of means is easy to obtain:

$$\sqrt{N_t}\,(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}} - \boldsymbol{\mu}_{gt}^{\mathbf{w}}) \to Normal(\mathbf{0}, \rho_g^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{w}}),$$

where $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ is the sample average for group/time cell (g,t) and

$$\boldsymbol{\Omega}_{gt}^{\mathbf{w}} = Var(\mathbf{w}_t | g)$$

is the $(K+1) \times (K+1)$ variance matrix for group/time cell (g,t). When we stack the means across groups and time periods, it is helpful to have the result

$$\sqrt{N}\,(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}} - \boldsymbol{\mu}_{gt}^{\mathbf{w}}) \to Normal(\mathbf{0}, (\rho_g \kappa_t)^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{w}}), \tag{5.15}$$

where $N = \sum_{t=1}^{T} N_t$ and $\kappa_t = \lim_{N \to \infty} (N_t/N)$ is, essentially, the fraction of all observations accounted for by cross section t. Of course, $\rho_g \kappa_t$ is consistently estimated by N_{gt}/N, and so, the

implication of (5.15) is that the sample average for cell (g,t) gets weighted by N_{gt}/N, the fraction of all observations accounted for by cell (g,t).

In implementing minimum distance estimation, we need a consistent estimator of $\mathbf{\Omega}_{gt}^{\mathbf{w}}$, and the group/time sample variance serves that purpose:

$$\hat{\mathbf{\Omega}}_{gt}^{\mathbf{w}} = N_{gt}^{-1} \sum_{i=1} r_{itg}(\mathbf{w}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}})(\mathbf{w}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}})' \xrightarrow{p} \mathbf{\Omega}_{gt}^{\mathbf{w}}. \tag{5.16}$$

Now let $\boldsymbol{\pi}$ be the vector of all cell means. For each (g,t), there are $K+1$ means, and so $\boldsymbol{\pi}$ is a $GT(K+1) \times 1$ vector. It makes sense to stack $\boldsymbol{\pi}$ starting with the $K+1$ means for $g = 1$, $t = 1$, $g = 1$, $t = 2$, ..., $g = 1$, $t = T$, ..., $g = G$, $t = 1$, ..., $g = G$, $t = T$. Now, the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are always independent across g because we assume random sampling for each t. When $\mathbf{x}_t$ does not contain lags or leads, the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are independent across t, too. (When we allow for lags of the response variable or explanatory variables, we will adjust the definition of $\boldsymbol{\pi}$ and the moment conditions. Thus, we will always assume that the $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{w}}$ are independent across g and t.) Then,

$$\sqrt{N}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \to \text{Normal}(\mathbf{0}, \mathbf{\Omega}), \tag{5.17}$$

where $\mathbf{\Omega}$ is the $GT(K+1) \times GT(K+1)$ block diagonal matrix with (g,t) block $\mathbf{\Omega}_{gt}^{\mathbf{w}}/(\rho_g \kappa_t)$. Note that $\mathbf{\Omega}$ incorporates both different cell variance matrices as well as the different frequencies of observations.

The set of equations in (5.8) constitute the restrictions on $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\alpha}$. Let $\boldsymbol{\theta}$ be the $(K+T+G-1)$ vector of these parameters, written as

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'.$$

There are $GT(K+1)$ restrictions in equations (5.8), so, in general, there are many overidentifying restrictions. We can write the set of equations in (5.8) as

$$\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}, \tag{5.18}$$

where $\mathbf{h}(\cdot,\cdot)$ is a $GT(K+1) \times 1$ vector. Because we have $\sqrt{N}$-asymptotically normal estimator $\hat{\boldsymbol{\pi}}$, a minimum distance approach suggests itself. It is different from the usual MD problem because the parameters do not appear in a separable way, but MD estimation is still possible. In fact, for the current application, $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is linear in each argument, which means MD estimators of $\boldsymbol{\theta}$ are in closed form.

Before obtaining the efficient MD estimator, we need, because of the nonseparability, an initial consistent estimator of $\boldsymbol{\theta}$. Probably the most straightforward is the "fixed effects"

estimator described above, but where we estimate all components of $\boldsymbol{\theta}$. The estimator uses the just identified set of equations.

For notational simplicity, let $\boldsymbol{\mu}_{gt}$ denote the $(K+1)\times 1$ vector of group/time means for each (g,t) cell. Then let $\boldsymbol{\omega}_{gt}$ be the $(K+T+G-1)\times 1$ vector $(\boldsymbol{\mu}_{gt}^{\mathbf{x}}, d_t, c_g)'$, where d_t is a $1\times(T-1)$ vector of time dummies and c_g is a $1\times G$ vector of group dummies. Then the moment conditions are

$$\left(\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\omega}_{gt}\boldsymbol{\omega}_{gt}'\right)\boldsymbol{\theta}-\left(\sum_{g=1}^{G}\sum_{t=1}^{T}\boldsymbol{\omega}_{gt}\mu_{gt}^{y}\right)=\mathbf{0}. \tag{5.19}$$

When we plug in $\hat{\boldsymbol{\pi}}$ – that is, the sample averages for all (g,t), then $\check{\boldsymbol{\theta}}$ is obtained as the so-called "fixed effects" estimator with time and group effects. The equations can be written as

$$\mathbf{q}(\hat{\boldsymbol{\pi}},\check{\boldsymbol{\theta}})=\mathbf{0}, \tag{5.20}$$

and this representation can be used to find the asymptotic variance of $\sqrt{N}(\check{\boldsymbol{\theta}}-\boldsymbol{\theta})$; naturally, it depends on $\boldsymbol{\Lambda}$ and is straightforward to estimate.

But there is a practically important point: there is nothing nonstandard about the MD problem, and bootstrapping is justified for obtaining asymptotic standard errors and test statistics. (Inoue (forthcoming) asserts that the "unconditional" limiting distribution of $\sqrt{N}(\check{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is not standard, but that is because he treats the sample means of the covariates and of the response variable differently; in effect, he conditions on the former.) The boostrapping is simple: resample each cross section separately, find the new groups for the bootstrap sample, and obtain the "fixed effects" estimates. It makes no sense here to resampling the groups.

Because of the nonlinear way that the covariate means appear in the estimation, the bootstrap may be preferred. The usual asymptotic normal approximation obtained from first-order asymptotics may not be especially good in this case, especially if $\sum_{g=1}^{G}\sum_{t=1}^{T}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime}\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$ is close to being singular, in which case $\boldsymbol{\beta}$ is poorly identified. (Inoue (2007) provides evidence that the distribution of the "FE" estimator, and what he calls a GMM estimator that accounts for different cell sample sizes, do not appear to be normal even with fairly large cell sizes. But his setup for generating the data is different – in particular, he specifies equations directly for the repeated cross sections, and that is how he generates data. As mentioned above, his asymptotic analysis differ from the MD framework, and implies nonnormal limiting distributions. If the data are drawn for each cross section to satisfy the population panel data model, the cell sizes are reasonably large, and there is sufficient variation in $\ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$, the minimum

distance estimators should have reasonable finite-sample properties. But because the limiting distribution depends on the $\hat{\bar{\mu}}_{gt}^{\mathbf{x}}$, which appear in a highly nonlinear way, asymptotic normal approximation might still be poor.)

With the restrictions written as in (5.18), Chamberlain (lecture notes) shows that the optimal weighting matrix is the inverse of

$$\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})', \tag{5.21}$$

where $\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is the $GT \times GT(K+1)$ Jacobian of $\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$. (In the standard case, $\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is the identity matrix.) We already have the consistent estimator of $\boldsymbol{\pi}$ – the cell averages – we showed how to consistently estimate $\boldsymbol{\Omega}$ in equations (5.16), and we can use $\breve{\boldsymbol{\theta}}$ as the initial consistent estimator of $\boldsymbol{\theta}$.

$\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\beta}) = \mathbf{I}_{GT} \otimes (-1, \boldsymbol{\beta}')$. Therefore, $\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta})$ is a block diagonal matrix with blocks

$$(-1, \boldsymbol{\beta}')(\rho_g\kappa_t)^{-1}\boldsymbol{\Omega}_{gt}^{w}(-1, \boldsymbol{\beta}')'. \tag{5.22}$$

But

$$\tau_{gt}^2 \equiv (-1, \boldsymbol{\beta}')\boldsymbol{\Omega}_{gt}^{w}(-1, \boldsymbol{\beta}')' = Var(y_t - \mathbf{x}_t\boldsymbol{\beta}|g), \tag{5.23}$$

and a consistent estimator is simply

$$N_{gt}^{-1}\sum_{i=1}^{N_t} r_{itg}(y_{it} - \mathbf{x}_{it}\breve{\boldsymbol{\beta}} - \breve{\eta}_t - \breve{\alpha}_g)^2$$

is the residual variance estimated within cell (g,t).

Now, $\nabla_{\theta}\mathbf{h}(\boldsymbol{\pi},\boldsymbol{\theta}) = \mathbf{W}(\boldsymbol{\pi})$, the $GT \times (K + T + G - 1)$ matrix of "regressors" in the FE estimation, that is, the rows of $\mathbf{W}(\boldsymbol{\pi})$ are $\boldsymbol{\omega}_{gt} = (\boldsymbol{\mu}_{gt}^{\mathbf{x}\prime}, \mathbf{d}_t, \mathbf{c}_g)$. Now, the FOC for the optimal MD estimator is

$$\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}(\hat{\mathbf{W}}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{gt}^{y}) = \mathbf{0},$$

and so

$$\hat{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}_{gt}^{y}. \tag{5.24}$$

So, as in the standard cases, the efficient MD estimator looks like a "weighted least squares" estimator. The estimated asymptotic variance of $\hat{\boldsymbol{\theta}}$, following Chamberlain, is just $(\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}/N$. Because $\hat{\mathbf{C}}^{-1}$ is the diagonal matrix with entries $(N_{gt}/N)/\hat{\tau}_{gt}^2$, it is easy to weight each cell (g,t) and then compute both $\hat{\boldsymbol{\theta}}$ and its asymptotic standard errors via a

weighted regression; fully efficient inference is straightforward. But one must compute the $\hat{\tau}_{gt}^2$ using the individual-level data in each group/time cell.

It is easily seen that the so-called "fixed effects" estimator, $\check{\boldsymbol{\theta}}$, is

$$\check{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\boldsymbol{\mu}}_{gt}^{y}, \tag{5.25}$$

that is, it uses the identity matrix as the weighting matrix. From Chamberlain (lecture notes), the asymptotic variance of $\check{\boldsymbol{\theta}}$ is estimated as $(\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\check{\mathbf{C}}\hat{\mathbf{W}}(\hat{\mathbf{W}}'\hat{\mathbf{W}})^{-1}$, where $\check{\mathbf{C}}$ is the matrix described above but with $\check{\boldsymbol{\beta}}$ used to estimate the cell variances. (Note: This matrix cannot be computed by just using the "heteroskedasticity-robust" standard errors in the regress $\hat{\mu}_{gt}^{y}$ on $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}$, $\mathbf{d}_t$, $\mathbf{c}_g$.) Because inference using $\check{\boldsymbol{\theta}}$ requires calculating the group/time specific variances, we might as well use the efficient MD estimator in (5.24).

Of course, after the efficient MD estimation, we can readily compute the overidentifying restrictions, which would be rejected if the underlying model needs to include cohort/time effects in a richer fashion.

A few remaining comments are in order. First, several papers, including Deaton (1985), Verbeek and Nijman (1993), and Collado (1997), use a different asymptotic analysis. In the current notation, $GT \to \infty$ (Deaton) or $G \to \infty$, with the cell sizes fixed. These approaches seems unnatural for the way pseudo panels are constructed, and the thought experiment about how one might sample more and more groups is convoluted. While $T \to \infty$ conceptually makes sense, it is still the case that the available number of time periods is much smaller than the cross section sample sizes for each T. McKenzie (2004) has shown that estimators derived under large G asymptotics can have good properties under the MD asymptotics used here. One way to see this is that the IV estimators proposed by Collado (1997), Verbeek and Vella (2005), and others are just different ways of using the population moment conditions in (5.8).

(Some authors appear to want it both ways. For example, Verbeek and Nijman (1993) use large G asymptotics, but treat the within-cell variances and covariances as known. This stance assumes that one can get precise estimates of the second moments within each cell, which means that N_{gt} should be large.)

Basing estimation on (5.8) and using minimum ditance, assuming large cell sizes, makes application to models with lags relatively straightforward. The only difference now is that the vectors of means, $\{\boldsymbol{\mu}_{gt}^{\mathbf{w}} : g = 1,\ldots,G; t = 1,\ldots,T\}$ now contain redundancies. (In other approaches to the problem, for example Collado (1997), McKenzie (2004), the problem with

adding y_{t-1} to the population model is that it generates correlation in the estimating equation based on the pooled cross sections. Here, there is no conceptual distinction between having exogenous or endogenous elements in $\mathbf{x}_t$; all that matters is how adding one modifies the MD moment conditions. As an example, suppose we write

$$y_t = \eta_t + \rho y_{t-1} + \mathbf{z}_t\boldsymbol{\gamma} + f + u_t \qquad (5.26)$$
$$E(u_t|g) = 0,\ g = 1,\ldots,G$$

where g is the group number. Then (5.8) is still valid. But, now we would define the vector of means as $(\mu_{gt}^{y}, \boldsymbol{\mu}_{gt}^{\mathbf{z}})$, and appropriately pick off μ_{gt}^{y} in defining the moment conditions. The alternative is to define $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ to include $\mu_{g,t-1}^{y}$, but this results in a singularity in the asymptotic distribution of $\hat{\boldsymbol{\pi}}$. It is much more straightforward to keep only nonredundant elements in $\boldsymbol{\pi}$ and readjust how the moment conditions.are defined in terms of $\boldsymbol{\pi}$. When we take that approach, it becomes clear that we now have fewer moments to estimate the parameters. If $\mathbf{z}_t$ is $1 \times J$, we have now have $J + T + G$ parameters to estimate from $GT(J + 1)$ population moments. Still, we have added just one more parameter.

To the best of my knowledge, the treatment here is the first to follow the MD approach, applied to (5.8), to its logical conclusion. Its strength is that the estimation method is widely known and used, and it separates the underlyng population model from sampling assumptions. It also shows why we need not make any exogeneity assumptions on $\mathbf{x}_t$. Perhaps most importantly, it reveals the key identification condition: that separate group/time effects are not needed in the underlying model, but enough group/time variation in the means $E(\mathbf{x}_t|g)$ is needed to identify the structural parameters. This sort of condition falls out of other approaches to the problem, such as the instrumental variables approach of but it is harder to see. For example, Verbeek and Vella (2005) propose instrumental variables methods on the equation in time averages using interactions between group (cohort) and time dummies. With a full set of separate time and group effects in the main equation – derivable here from the population panel model – the key identification assumption is that a full set of group/time effects can be excluded from the structural equation, but the means of the covariates have to vary sufficiently across group/time. That is exactly the conclusion we reach with a minimum distance approach.

Interestingly, the MD approach applies easily to extensions of the basic model. For example, we can allow for unit-specific time trends (as in the random growth model of Heckman and Hotz (1989)):

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f_1 + f_2 t + u_t, \tag{5.27}$$

where, for a random draw i, the unobserved heterogeneity is of the form $f_{i1} + f_{i2}t$. Then, using the same arguments as before,

$$\mu_{gt}^{y} = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g + \varphi_g t, \tag{5.28}$$

and this set of moment conditions is easily handled by extending the previous analysis. We can even estimate models with time-varying factor loads on the heterogeneity:

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + \lambda_t f + u_t,$$

where $\lambda_1 = 1$ (say) as a normalization. Now the population moments satisfy

$$\mu_{gt}^{y} = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \lambda_t\alpha_g.$$

There are now $K + G + 2(T - 1)$ free parameters to estimate from $GT(K + 1)$ moments. This extension means that the estimating equations allow the group/time effects to enter more flexibly (although, of course, we cannot replace $\eta_t + \lambda_t\alpha_g$ with unrestricted group/time effects.) The MD estimation problem is now nonlinear because of the interaction term, $\lambda_t\alpha_g$. With more parameters and perhaps not much variation in the $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$, practical implementation may be a problem, but the theory is standard.

This literature would benefit from a careful simulation study, where data for each cross section are generated from the underlying population model, and where g_i – the group identifier – is randomly drawn, too. To be realistic, the underlying model should have full time effects. Verbeek and Vella (2005) come close, but they omit aggregate time effects in the main model while generating the explanatory variables to have means that differ by group/time cell. Probably this paints too optimistic a picture for how well the estimators can work in practice. Remember, even if we can get precise estimates of the cell means, the variation in $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ across g and t might not be enough to tie down $\boldsymbol{\beta}$ precisely.

Finally, we can come back to the comment about how the moment conditions in (5.8) only use the assumption $E(u_t|g) = 0$ for all t and g. It seems likely that we should be able to exploit contemporaneous exogeneity assumptions. Let $\mathbf{z}_t$ be a set of observed variables such that $E(u_t|\mathbf{z}_t, f) = \mathbf{0}$, $t = 1, \ldots, T$. (In a true panel, these vary across i and t. We might have $\mathbf{z}_t = \mathbf{x}_t$, but perhaps $\mathbf{z}_t$ is just a subset of $\mathbf{x}_t$, or we have extra instruments.) Then we can add to (5.8) the moment conditions

$$\begin{aligned} E(\mathbf{z}_t'y_t|g) &= \eta_t E(\mathbf{z}_t|g) + E(\mathbf{z}_t'\mathbf{x}_t|g)\boldsymbol{\beta} + E(\mathbf{z}_t'f|g) + E(\mathbf{z}_t'u_t|g) \\ &= \eta_t E(\mathbf{z}_t|g) + E(\mathbf{z}_t'\mathbf{x}_t|g)\boldsymbol{\beta} + E(\mathbf{z}_t'f|g), \end{aligned} \tag{5.29}$$

where $E(\mathbf{z}_t'u_t|g) = \mathbf{0}$ when we view group designation as contained in f. The moments $E(\mathbf{z}_t'y_t|g)$, $E(\mathbf{z}_t|g)$, and $E(\mathbf{z}_t'\mathbf{x}_t|g)$ can all be estimated by random samples from each cross section, where we average within group/time period. (This would not work if $\mathbf{x}_t$ or $\mathbf{z}_t$ contains lags.) This would appear to add many more moment restrictions that should be useful for identifying $\boldsymbol{\beta}$, but that depends on what we assume about the unobserved moments $E(\mathbf{z}_t'f|g)$.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 3, Monday, July 30th, 2.00-3.00pm

Regression Discontintuity Designs[1]

1. Introduction

Since the late 1990s there has been a large number of studies in economics applying and extending Regression Discontinuity (RD) methods from its origins in the statistics literature in the early 60's (Thisthlewaite and Cook, 1960). Here, we review some of the practical issues in implementation of RD methods. The focus is on five specific issues. The first is the importance of graphical analyses as powerful methods for illustrating the design. Second, we suggest using local linear regression methods using only the observations close to the discontinuity point. Third, we discuss choosing the bandwidth using cross validation specifically tailored to the focus on estimation of regression functions on the boundary of the support, following Ludwig and Miller (2005). Fourth, we provide two simple estimators for the asymptotic variance, one of them exploiting the link with instrumental variables methods derived by Hahn, Todd, and VanderKlaauw (2001, HTV). Finally, we discuss a number of specification tests and sensivitity analyses based on tests for (a) discontinuities in the average values for covariates, (b) discontinuities in the conditional density of the forcing variable, as suggested by McCrary (2007), (c) discontinuities in the average outcome at other values of the forcing variable.

2. Sharp and Fuzzy Regression Discontinuity Designs

2.1 Basics

Our discussion will frame the RD design in the context of the modern literature on causal effects and treatment effects, using the potential outcomes framework (Rubin, 1974), rather than the regression framework that was originally used in this literature. For unit i there are two potential outcomes, $Y_i(0)$ and $Y_i(1)$, with the causal effect defined as the difference

[1] These notes draw heavily on Imbens and Lemieux (2007).

$Y_i(1) - Y_i(0)$, and the observed outcome equal to

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases}$$

where $W_i \in \{0, 1\}$ is the binary indicator for the treatment.

The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor (the forcing variable X_i) being on either side of a common threshold. This predictor X_i may itself be associated with the potential outcomes, but this association is assumed to be smooth, and so any discontinuity in the conditional distribution of the outcome, indexed by the value of this covariate at the cutoff value, is interpreted as evidence of a causal effect of the treatment. The design often arises from administrative decisions, where the incentives for units to participate in a program are partly limited for reasons of resource constraints, and clear transparent rules rather than discretion by administrators are used for the allocation of these incentives.

2.2 The Sharp Regression Discontinuity Design

It is useful to distinguish between two designs, the Sharp and the Fuzzy Regression Discontinuity (SRD and FRD from hereon) designs (e.g., Trochim, 1984, 2001; HTV). In the SRD design the assignment W_i is a deterministic function of one of the covariates, the forcing (or treatment-determining) variable X:

$$W_i = 1\{X_i \geq c\}.$$

All units with a covariate value of at least c are in the treatment group (and participation is mandatory for these individuals), and all units with a covariate value less than c are in the control group (members of this group are not eligible for the treatment). In the SRD design we look at the discontinuity in the conditional expectation of the outcome given the covariate to uncover an average causal effect of the treatment:

$$\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(1)|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i(0)|X_i = x], \quad (1)$$

is interpreted as the average causal effect of the treatment at the discontinuity point.

$$\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]. \quad (2)$$

In order to justify this interpretation we make a smoothness assumption. Typically this assumption is formulated in terms of conditional expectations[2]:

Assumption 1 (Continuity of Conditional Regression Functions)

$$\mathbb{E}[Y(0)|X=x] \qquad \text{and} \quad \mathbb{E}[Y(1)|X=x],$$

are continuous in x.

Under this assumption,

$$\tau_{\text{SRD}} = \lim_{x\downarrow c} \mathbb{E}[Y_i|X_i=x] - \lim_{x\uparrow c} \mathbb{E}[Y_i|X_i=x].$$

The estimand is the difference of two regression functions at a point.

There is a unavoidable need for extrapolation, because by design there are no units with $X_i = c$ for whom we observe $Y_i(0)$. We therefore will exploit the fact that we observe units with covariate values arbitrarily close to c.[3]

As an example of a SRD design, consider the study of the effect of party affiliation of a congressman on congressional voting outcomes by Lee (2007). See also Lee, Moretti and Butler (2004). The key idea is that electoral districts where the share of the vote for a Democrat in a particular election was just under 50% are on average similar in many relevant respects to districts where the share of the Democratic vote was just over 50%, but the small difference in votes leads to an immediate and big difference in the party affiliation of the elected representative. In this case, the party affiliation always jumps at 50%, making this is a SRD design. Lee looks at the incumbency effect. He is interested in the probability

[2]More generally, one might want to assume that the conditional distribution function is smooth in the covariate. Let $F_{Y(w)|X}(y|x) = \Pr(Y_i(w) \leq y|X_i = x)$ denote the conditional distribution function of $Y_i(w)$ given X_i. Then the general version of the assumption assume that $F_{Y(0)|X}(y|x)$ and $F_{Y(1)|X}(y|x)$ are continuous in x for all y. Both assumptions are stronger than required, as we will only use continuity at $x = c$, but it is rare that it is reasonable to assume continuity for one value of the covariate, but not at other values of the covariate.

[3]Although in principle the first component in the difference in (1) would be straightforward to estimate if we actually observe individuals with $X_i = x$, with continuous covariates we also need to estimate this term by averaging over units with covariate values close to c.

of Democrats winning the subsequent election, comparing districts where the Democrats won the previous election with just over 50% of the popular vote with districts where the Democrats lost the previous election with just under 50% of the vote.

2.3 The Fuzzy Regression Discontinuity Design

In the Fuzzy Regression Discontinuity (FRD) design the probability of receiving the treatment need not change from zero to one at the threshold. Instead the design allows for a smaller jump in the probability of assignment to the treatment at the threshold:

$$\lim_{x\downarrow c}\Pr(W_i=1|X_i=x)\neq\lim_{x\uparrow c}\Pr(W_i=1|X_i=x),$$

without requiring the jump to equal 1. Such a situation can arise if incentives to participate in a program change discontinuously at a threshold, without these incentives being powerful enough to move all units from nonparticipation to participation. In this design we interpret the ratio of the jump in the regression of the outcome on the covariate to the jump in the regression of the treatment indicator on the covariate as an average causal effect of the treatment. Formally, the estimand is

$$\tau_{\text{FRD}}=\frac{\lim_{x\downarrow c}\mathbb{E}[Y_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[Y_i|X_i=x]}{\lim_{x\downarrow c}\mathbb{E}[W_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[W_i|X_i=x]}.$$

As an example of a FRD design, consider the study of the effect of financial aid on college attendance by VanderKlaauw (2002). VanderKlaauw looks at the effect of financial aid on acceptance on college admissions. Here X_i is a numerical score assigned to college applicants based on the objective part of the application information (SAT scores, grades) used to streamline the process of assigning financial aid offers. During the initial stages of the admission process, the applicants are divided into L groups based on discretized values of these scores. Let

$$G_i=\begin{cases}1 & \text{if } 0\leq X_i<c_1\\ 2 & \text{if } c_1\leq X_i<c_2\\ \vdots & \\ L & \text{if } c_{L-1}\leq X_i\end{cases}$$

denote the financial aid group. For simplicity, let us focus on the case with $L=2$, and a single cutoff point c. Having a score just over c will put an applicant in a higher category and

increase the chances of financial aid discontinuously compared to having a score just below c. The outcome of interest in the VanderKlaauw study is college attendance. In this case, the statistical association between attendance and the financial aid offer is ambiguous. On the one hand, an aid offer by a college makes that college more attractive to the potential student. This is the causal effect of interest. On the other hand, a student who gets a generous financial aid offer from one college is likely to have better outside opportunities in the form of financial aid offers from other colleges. In the VanderKlaauw application College aid is emphatically not a deterministic function of the financial aid categories, making this a fuzzy RD design. Other components of the college application package that are not incorporated in the numerical score such as the essay and recommendation letters undoubtedly play an important role. Nevertheless, there is a clear discontinuity in the probability of receiving an offer of a larger financial aid package.

Let us first consider the interpretation of τ_{FRD}. HTV exploit the instrumental variables connection to interpret the fuzzy regression discontinuity design when the effect of the treatment varies by unit. Let $W_i(x)$ be potential treatment status given cutoff point x, for x in some small neigborhood around c. $W_i(x)$ is equal to one if unit i would take or receive the treatment if the cutoff point was equal to x. This requires that the cutoff point is at least in principle manipulable.[4] For example, if X is age, one could imagine changing the age that makes an individual eligible for the treatment from c to $c + \epsilon$. Then it is useful to assume monotonicity (see HTV):

Assumption 2 *$W_i(x)$ is non-increasing in x at $x = c$.*

Next, define compliance status. This concept is similar to that in instrumental variables, e.g., Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996). A complier is a unit such that

$$\lim_{x \downarrow X_i} W_i(x) = 0, \qquad \text{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

[4] Alternatively, one could think of the individual characteristic X_i as being manipulable, but in many cases this is an immutable characteristic such as age.

Compliers are units who would get the treatment if the cutoff were at X_i or below, but they would not get the treatment if the cutoff were higher than X_i. To be specific, consider an example where individuals with a test score less than c are encouraged for a remedial teaching program (Matsudaira, 2007). Interest is in the effect of the remedial teaching program on subsequent test scores. Compliers are individuals who would participate if encouraged (if the cutoff for encouragement is below or equal to their actual score), but not if not encouraged (if the cutoff for encouragement is higher than their actual score). Then

$$\frac{\lim_{x\downarrow c}\mathbb{E}[Y_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[Y_i|X_i=x]}{\lim_{x\downarrow c}\mathbb{E}[W_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[W_i|X_i=x]}$$

$$=\mathbb{E}[Y_i(1)-Y_i(0)|\text{unit } i \text{ is a complier and } X_i=c].$$

The estimand is an average effect of the treatment, but only averaged for units with $X_i = c$ (by regression discontinuity), and only for compliers (people who are affected by the threshold).

3. The FRD Design, Unconfoundedness and External Validity

3.1 The FRD Design and Unconfoundedness

In the FRD setting it is useful to contrast the RD approach with estimation of average causal effects under unconfoundedness. The unconfoundedness assumption, e.g., Rosenbaum and Rubin (1983), Imbens (2004), requires that

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \,\Big|\, X_i.$$

If this assumption holds, then we can estimate the average effect of the treatment at $X_i = c$ as

$$\mathbb{E}[Y_i(1)-Y_i(0)|X_i=x]=\mathbb{E}[Y_i|W_i=1,X_i=c]-\mathbb{E}[Y_i|W_i=0,X_i=c].$$

This approach does not exploit the jump in the probability of assignment at the discontinuity point. Instead it assumes that differences between treated and control units with $X_i = c$ are interpretable as average causal effects.

In contrast, the assumptions underlying an FRD analysis implies that comparing treated and control units with $X_i = c$ is likely to be the wrong approach. Treated units with $X_i = c$ include compliers and alwaystakers, and control units at $X_i = c$ consist only of nevertakers. Comparing these different types of units has no causal interpretation under the FRD assumptions. Although, in principle, one cannot test the unconfoundedness assumption, one aspect of the problem makes this assumption fairly implausible. Unconfoundedness is fundamentally based on units being comparable if their covariates are similar. This is not an attractive assumption in the current setting where the probability of receiving the treatment is discontinuous in the covariate. Thus units with similar values of the forcing variable (but on different sides of the threshold) must be different in some important way related to the receipt of treatment. Unless there is a substantive argument that this difference is immaterial for the comparison of the outcomes of interest, an analysis based on unconfoundedness is not attractive in this setting.

3.2 The FRD Design and External Validity

One important aspect of both the SRD and FRD designs is that they at best provide estimates of the average effect for a subpopulation, namely the subpopulation with covariate value equal to $X_i = c$. The FRD design restricts the relevant subpopulation even further to that of compliers at this value of the covariate. Without strong assumptions justifying extrapolation to other subpopulations (e.g., homogeneity of the treatment effect) the designs never allow the researcher to estimate the overall (population) average effect of the treatment. In that sense the design has fundamentally only a limited degree of external validity, although the specific average effect that is identified may well be of special interest, for example in cases where the policy question concerns changing the location of the threshold. The advantage of RD designs compared to other non-experimental analyses that may have more external validity such as those based on unconfoundedness, is that RD designs generally have a relatively high degree of internal validity in settings where they are applicable.

4. Graphical Analyses

4.1 Introduction

Graphical analyses should be an integral part of any RD analysis. The nature of RD designs suggests that the effect of the treatment of interest can be measured by the value of the discontinuity in the expected value of the outcome at a particular point. Inspecting the estimated version of this conditional expectation is a simple yet powerful way to visualize the identification strategy. Moreover, to assess the credibility of the RD strategy, it is useful to inspect two additional graphs. The estimators we discuss later use more sophisticated methods for smoothing but these basic plots will convey much of the intuition. For strikingly clear examples of such plots, see Lee, Moretti, and Butler (2004), Lalive (2007), and Lee (2007). Two figures from Lee (2007) are attached.

4.2 Outcomes by Forcing Variable

The first plot is a histogram-type estimate of the average value of the outcome by the forcing variable. For some binwidth h, and for some number of bins K_0 and K_1 to the left and right of the cutoff value, respectively, construct bins $(b_k, b_{k+1}]$, for $k = 1, \ldots, K = K_0 + K_1$, where

$$b_k = c - (K_0 - k + 1) \cdot h.$$

Then calculate the number of observations in each bin,

$$N_k = \sum_{i=1}^{N} 1\{b_k < X_i \leq b_{k+1}\},$$

and the average outcome in the bin:

$$\overline{Y}_k = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Y_i \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

The first plot of interest is that of the $\overline{Y}_k$, for $k = 1, K$ against the mid point of the bins, $\tilde{b}_k = (b_k + b_{k+1})/2$. The question is whether around the threshold c there is any evidence of a jump in the conditional mean of the outcome. The formal statistical analyses discussed below are essentially just sophisticated versions of this, and if the basic plot does not show

any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantially significant magnitudes. In addition to inspecting whether there is a jump at this value of the covariate, one should inspect the graph to see whether there are any other jumps in the conditional expectation of Y_i given X_i that are comparable to, or larger than, the discontinuity at the cutoff value. If so, and if one cannot explain such jumps on substantive grounds, it would call into question the interpretation of the jump at the threshold as the causal effect of the treatment. In order to optimize the visual clarity it is important to calculate averages that are not smoothed over the cutoff point. The attached figure is taken from the paper by Lee (2007).

4.2 Covariates by Forcing Variable

The second set of plots compares average values of other covariates in the K bins. Specifically, let Z_i be the M-vector of additional covariates, with m-th element Z_{im}. Then calculate

$$\overline{Z}_{km} = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Z_{im} \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

The second plot of interest is that of the $\overline{Z}_{km}$, for $k = 1, K$ against the mid point of the bins, $\tilde{b}_k$, for all $m = 1, \ldots, M$. Lee (2007) presents such a figure for a lagged value of the outcome, namely the election results from a prior election, against the vote share in the last election. In the case of FRD designs, it is also particularly useful to plot the mean values of the treatment variable W_i to make sure there is indeed a jump in the probability of treatment at the cutoff point. Plotting other covariates is also useful for detecting possible specification problems (see Section 8) in the case of either SRD or FRD designs.

4.3 The Density of the Forcing Variable

In the third graph one should plot the number of observations in each bin, N_k, against the mid points $\tilde{b}_k$. This plot can be used to inspect whether there is a discontinuity in the distribution of the forcing variable X at the threshold. McCrary (2007) suggests that such discontinuity would raise the question whether the value of this covariate was manipulated

Figure IIa: Candidate's Probability of Winning Election t+1, by Margin of Victory in Election t: local averages and parametric fit

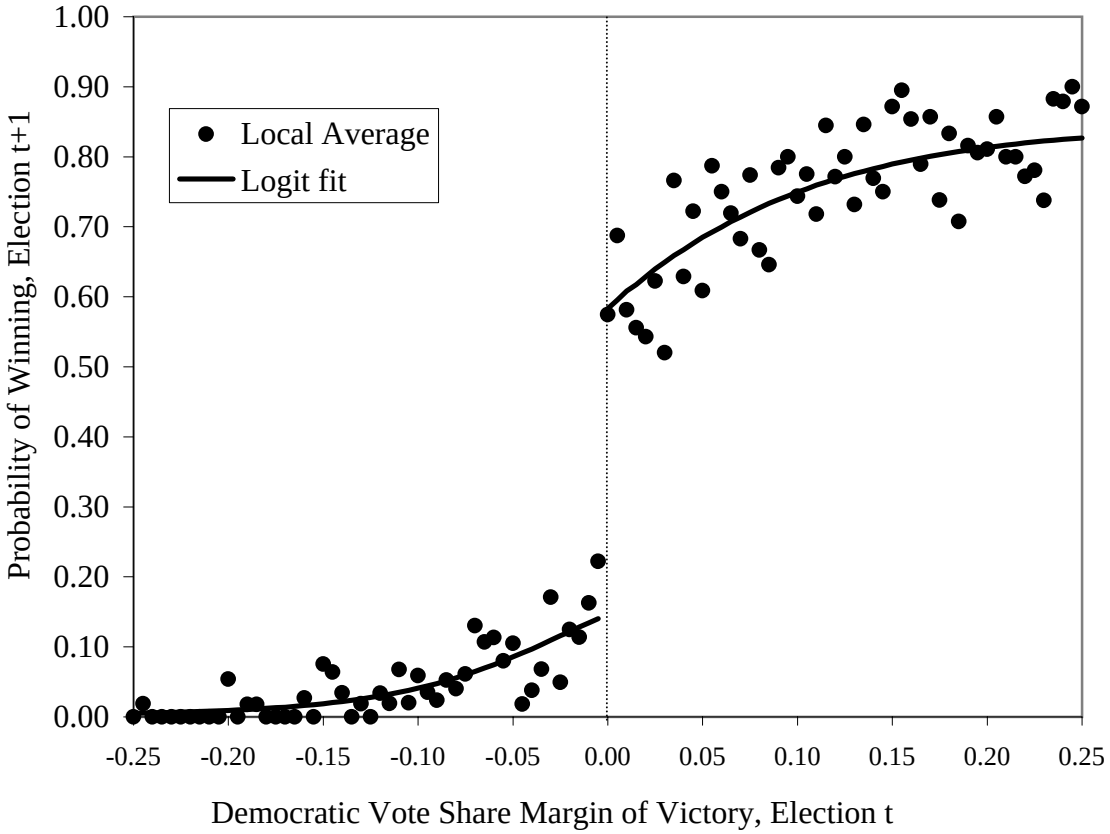
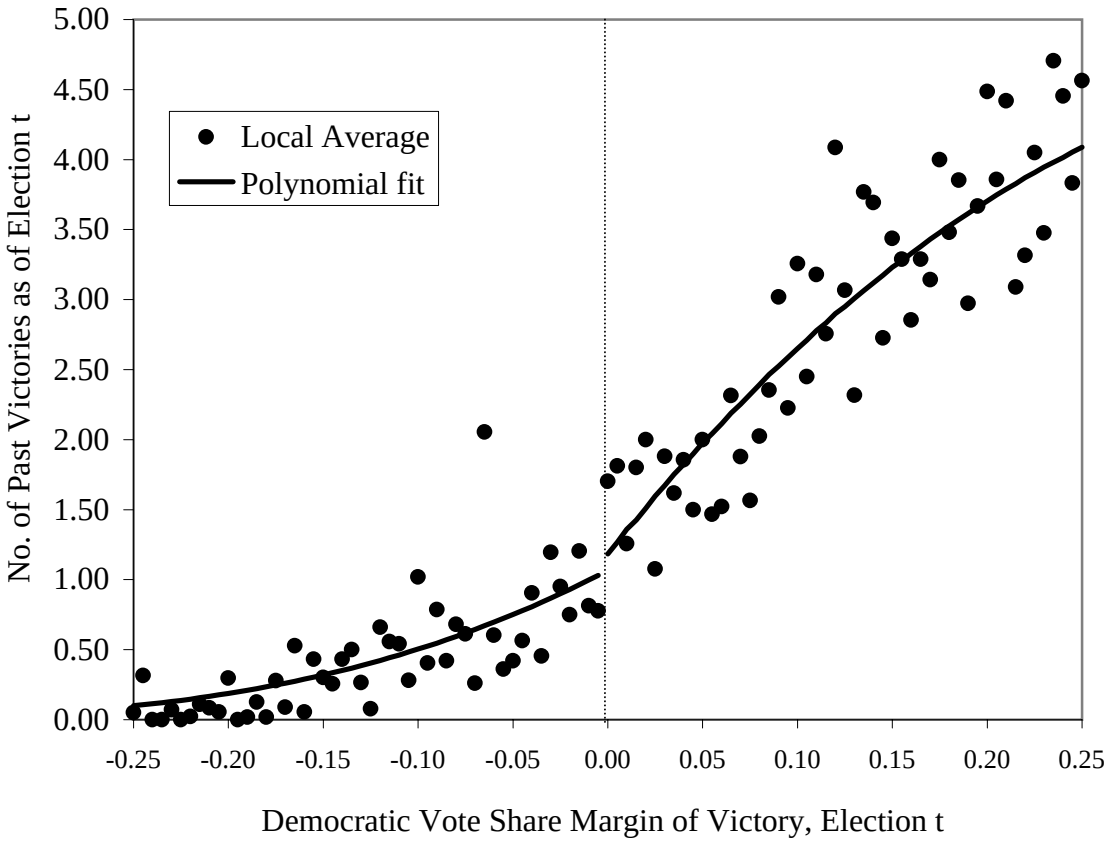


Figure IIb: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and parametric fit



by the individual agents, invalidating the design. For example, suppose that the forcing variable is a test score. If individuals know the threshold and have the option of re-taking the test, individuals with test scores just below the threshold may do so, and invalidate the design. Such a situation would lead to a discontinuity of the conditional density of the test score at the threshold, and thus be detectable in plots such as described here. See Section 8 for more discussion of the specification tests based on this idea.

5. Estimation: Local Linear Regression

5.1 Nonparametric Regression at the Boundary

The practical estimation of the treatment effect τ in both the SRD and FRD designs is largely standard nonparametric regression (e.g., Pagan and Ullah, 1999; Härdle, 1990; Li and Racine, 2007). However, there are two unusual features to estimation in the RD setting. First, we are interested in the regression function at a single point, and second, that single point is a boundary point. As a result, standard nonparametric kernel regression does not work very well. At boundary points, such estimators have a slower rate of convergence than they do at interior points. Standard methods for choosing the bandwidth are also not designed to provide good choices in this setting.

5.2 Local Linear Regression

Here we discuss local linear regression (Fan and Gijbels, 1996). Instead of locally fitting a constant function, we can fit linear regression functions to the observations within a distance h on either side of the discontinuity point:

$$\min_{\alpha_l,\beta_l} \sum_{i|c-h<X_i<c}^{N} \left(Y_i - \alpha_l - \beta_l \cdot (X_i - c)\right)^2,$$

and

$$\min_{\alpha_r,\beta_r} \sum_{i|c\leq X_i<c+h}^{N} \left(Y_i - \alpha_r - \beta_r \cdot (X_i - c)\right)^2.$$

The value of $\mu_l(c)$ and $\mu_r(c)$ are then estimated as

$$\widehat{\mu_l(c)} = \hat{\alpha}_l + \hat{\beta}_l \cdot (c - c) = \hat{\alpha}_l, \qquad \text{and} \quad \widehat{\mu_r(c)} = \hat{\alpha}_r + \hat{\beta}_r \cdot (c - c) = \hat{\alpha}_r,$$

Given these estimates, the average treatment effect is estimated as

$$\hat{\tau}_{\text{SRD}} = \hat{\alpha}_r - \hat{\alpha}_l.$$

Alternatively one can estimate the average effect directly in a single regression, by solving

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{N} 1\{c-h \leq X_i \leq c+h\} \cdot \left(Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i\right)^2,$$

which will numerically yield the same estimate of τ_{SRD}.

We can make the nonparametric regression more sophisticated by using weights that decrease smoothly as the distance to the cutoff point increases, instead of the zero/one weights based on the rectangular kernel. However, even in this simple case the asymptotic bias can be shown to be of order h^2, and the more sophisticated kernels rarely make much difference. Furthermore, if using different weights from a more sophisticated kernel does make a difference, it likely suggests that the results are highly sensitive to the choice of bandwidth. So the only case where more sophisticated kernels may make a difference is when the estimates are not very credible anyway because of too much sensitivity to the choice of bandwidth. From a practical point of view one may just focus on the simple rectangular kernel, but verify the robustness of the results to different choices of bandwidth.

For inference we can use standard least squares methods. Under appropriate conditions on the rate at which the bandwidth goes to zero as the sample size increases, the resulting estimates will be asymptotically normally distributed, and the (robust) standard errors from least squares theory will be justified. Using the results from HTV, the optimal bandwidth is $h \propto N^{-1/5}$. Under this sequence of bandwidths the asymptotic distribution of the estimator $\hat{\tau}$ will have a non-zero bias. If one does some undersmoothing, by requiring that $h \propto N^{-\delta}$ with $1/5 < \delta < 2/5$, then the asymptotic bias disappears and standard least squares variance estimators will lead to valid confidence intervals.

5.3 Covariates

Often there are additional covariates available in addition to the forcing covariate that is the basis of the assignment mechanism. These covariates can be used to eliminate small

sample biases present in the basic specification, and improve the precision. In addition, they can be useful for evaluating the plausibility of the identification strategy, as discussed in Section 8.1. Let the additional vector of covariates be denoted by Z_i. We make three observations on the role of these additional covariates.

The first and most important point is that the presence of these covariates rarely changes the identification strategy. Typically, the conditional distribution of the covariates Z given X is continuous at $x = c$. If such discontinuities in other covariates are found, the justification of the identification strategy may be questionable. If the conditional distribution of Z given X is continuous at $x = c$, then including Z in the regression

$$\min_{\alpha,\beta,\tau,\delta} \sum_{i=1}^{N} 1\{c-h \leq X_i \leq c+h\} \cdot (Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i - \delta' Z_i)^2 ,$$

will have little effect on the expected value of the estimator for τ, since conditional on X being close to c, the additional covariates Z are independent of W.

The second point is that even though with X very close to c, the presence of Z in the regression does not affect any bias, in practice we often include observations with values of X not too close to c. In that case, including additional covariates may eliminate some bias that is the result of the inclusion of these additional observations.

Third, the presence of the covariates can improve precision if Z is correlated with the potential outcomes. This is the standard argument, which also supports the inclusion of covariates even in analyses of randomized experiments. In practice the variance reduction will be relatively small unless the contribution to the $\mathbb{R}^2$ from the additional regressors is substantial.

5.4 Estimation for the Fuzzy Regression Discontinuity Design

In the FRD design, we need to estimate the ratio of two differences. The estimation issues we discussed earlier in the case of the SRD arise now for both differences. In particular, there are substantial biases if we do simple kernel regressions. Instead, it is again likely to be better to use local linear regression. We use a uniform (rectangular) kernel, with the same

bandwidth for estimation of the discontinuity in the outcome and treatment regressions.

First, consider local linear regression for the outcome, on both sides of the discontinuity point. Let

$$\left(\hat{\alpha}_{yl},\hat{\beta}_{yl}\right)=\arg\min_{\alpha_{yl},\beta_{yl}}\sum_{i:c-h\leq X_i<c}\left(Y_i-\alpha_{yl}-\beta_{yl}\cdot(X_i-c)\right)^2, \tag{3}$$

$$\left(\hat{\alpha}_{yr},\hat{\beta}_{yr}\right)=\arg\min_{\alpha_{yr},\beta_{yr}}\sum_{i:c\leq X_i\leq c+h}\left(Y_i-\alpha_{yr}-\beta_{yr}\cdot(X_i-c)\right)^2. \tag{4}$$

The magnitude of the discontinuity in the outcome regression is then estimated as $\hat{\tau}_y=\hat{\alpha}_{yr}-\hat{\alpha}_{yl}$. Second, consider the two local linear regression for the treatment indicator:

$$\left(\hat{\alpha}_{wl},\hat{\beta}_{wl}\right)=\arg\min_{\alpha_{wl},\beta_{wl}}\sum_{i:c-h\leq X_i<c}\left(W_i-\alpha_{wl}-\beta_{wl}\cdot(X_i-c)\right)^2, \tag{5}$$

$$\left(\hat{\alpha}_{wr},\hat{\beta}_{wr}\right)=\arg\min_{\alpha_{wr},\beta_{wr}}\sum_{i:c\leq X_i\leq c+h}\left(Y_i-\alpha_{wr}-\beta_{wr}\cdot(X_i-c)\right)^2. \tag{6}$$

The magnitude of the discontinuity in the treatment regression is then estimated as $\hat{\tau}_w=\hat{\alpha}_{wr}-\hat{\alpha}_{wl}$. Finally, we estimate the effect of interest as the ratio of the two discontinuities:

$$\hat{\tau}_{\rm FRD}=\frac{\hat{\tau}_y}{\hat{\tau}_w}=\frac{\hat{\alpha}_{yr}-\hat{\alpha}_{yl}}{\hat{\alpha}_{wr}-\hat{\alpha}_{wl}}. \tag{7}$$

Because of the specific implementation we use here, with a uniform kernel, and the same bandwidth for estimation of the denominator and the numerator, we can characterize the estimator for τ as a Two-Stage-Least-Squares (TSLS) estimator (See HTV). This equality still holds when we use local linear regression and include additional regressors. Define

$$V_i=\begin{pmatrix}1\\ 1\{X_i<c\}\cdot(X_i-c)\\ 1\{X_i\geq c\}\cdot(X_i-c)\end{pmatrix},\qquad\text{and}\quad\delta=\begin{pmatrix}\alpha_{yl}\\ \beta_{yl}\\ \beta_{yr}\end{pmatrix}. \tag{8}$$

Then we can write

$$Y_i=\delta' V_i+\tau\cdot W_i+\varepsilon_i. \tag{9}$$

Estimating τ based on the regression function (9) by TSLS methods, with the indicator $1\{X_i\geq c\}$ as the excluded instrument and V_i as the set of exogenous variables is numerically identical to $\hat{\tau}_{\rm FRD}$ as given in (7).

6. Bandwidth Selection

An important issue in practice is the selection of the smoothing parameter, the binwidth h. Here we focus on cross-validation procedures rather than plug in methods which would require estimating derivatives nonparametrically. The specific methods discussed here are based on those developed by Ludwig and Miller (2005, 2007). Initially we focus on the SRD case, and in Section 6.2 we extend the recommendations to the FRD setting.

To set up the bandwidth choice problem we generalize the notation slightly. In the SRD setting we are interested in the

$$\tau_{\text{SRD}} = \lim_{x\downarrow c}\mu(x) - \lim_{x\uparrow c}\mu(x),$$

where $\mu(x) = \mathbb{E}[Y_i|X_i = x]$. We estimate the two terms as

$$\widehat{\lim_{x\downarrow c}\mu(x)} = \hat{\alpha}_r(c), \qquad \text{and} \quad \widehat{\lim_{x\uparrow c}\mu(x)} = \hat{\alpha}_l(c),$$

where $\hat{\alpha}_l(x)$ and $\hat{\beta}_l(x)$ solve

$$\Big(\hat{\alpha}_l(x), \hat{\beta}_l(x)\Big) = \arg\min_{\alpha,\beta} \sum_{j|x-h<X_j<x} \left(Y_j - \alpha - \beta\cdot(X_j - x)\right)^2. \tag{10}$$

and $\hat{\alpha}_r(x)$ and $\hat{\beta}_r(x)$ solve

$$\Big(\hat{\alpha}_r(x), \hat{\beta}_r(x)\Big) = \arg\min_{\alpha,\beta} \sum_{j|x<X_j<x+h} \left(Y_j - \alpha - \beta\cdot(X_j - x)\right)^2. \tag{11}$$

Let us focus first on estimating $\lim_{x\downarrow c}\mu(x)$. For estimation of this limit we are interested in the bandwidth h that minimizes

$$Q_r(x,h) = \mathbb{E}\left[\left(\lim_{z\downarrow x}\mu(z) - \hat{\alpha}_r(x)\right)^2\right],$$

at $x = c$. However, we focus on a single bandwidth for both sides of the threshold, and therefore focus on minimizing

$$Q(c,h) = \frac{1}{2}\cdot(Q_l(c,h) + Q_r(c,h)) = \frac{1}{2}\cdot\left(\mathbb{E}\left[\left(\lim_{x\uparrow c}\mu(x) - \hat{\alpha}_l(c)\right)^2\right] + \mathbb{E}\left[\left(\lim_{x\downarrow c}\mu(x) - \hat{\alpha}_r(c)\right)^2\right]\right).$$

We now discuss two methods for choosing the bandwidth.

6.1 Bandwidth Selection for the SRD Design

For a given binwidth h, let the estimated regression function at x be

$$\hat{\mu}(x) = \begin{cases} \hat{\alpha}_l(x) & \text{if } x < c, \\ \hat{\alpha}_r(x) & \text{if } x \geq c, \end{cases}$$

where $\hat{\alpha}_l(x)$, $\hat{\beta}_l(x)$, $\hat{\alpha}_r(x)$ and $\hat{\beta}_r(x)$ solve (10) and (11). Note that in order to mimic the fact that we are interested in estimation at the boundary we only use the observations on one side of x in order to estimate the regression function at x, rather than the observations on both sides of x, that is, observations with $x - h < X_j < x + h$. In addition, the strict inequality in the definition implies that $\hat{\mu}(x)$ evaluated at $x = X_i$ does not depend on Y_i.

Now define the cross-validation criterion as

$$\text{CV}_Y(h) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \hat{\mu}(X_i)\right)^2, \tag{12}$$

with the corresponding cross-validation choice for the binwidth

$$h_{\text{CV}}^{\text{opt}} = \arg\min_h \text{CV}_Y(h).$$

The expected value of this cross-validation function is under some conditions equal to $\mathbb{E}[\text{CV}_Y(h)] = C + \mathbb{E}[Q(X,h)] = C + \int Q(x,h) f_X(dx)$, for some constant that does not depend on h. Although the modification to estimate the regression using one-sided kernels mimics more closely the estimand of interest, this is still not quite what we are interested in. Ultimately we are solely interested in estimating the regression function in the neighborhood of a single point, the threshold c, and thus in minimizing $Q(c,h)$, rather than $\int_x Q(x,h) f_X(x) dx$. If there are few observations in the tails of the distributions, minimizing the criterion in (12) may lead to larger bins than is optimal for estimating the regression function around $x = c$ if c is in the center of the distribution. We may therefore wish to minimize the cross-validation criterion after first discarding observations from the tails. Let $q_{X,\delta,l}$ be δ quantile of the empirical distribution of X for the subsample with $X_i < c$, and let

$q_{X,\delta,r}$ be δ quantile of the empirical distribution of X for the subsample with $X_i \geq c$. Then we may wish to use the criterion

$$\text{CV}_Y^\delta(h) = \frac{1}{N} \sum_{i: q_{X,\delta,l} \leq X_i \leq q_{X,1-\delta,r}} \left(Y_i - \hat{\mu}(X_i)\right)^2. \tag{13}$$

The modified cross-validation choice for the bandwidth is

$$h_{\text{CV}}^{\delta,\text{opt}} = \arg\min_h \text{CV}_Y^\delta(h). \tag{14}$$

The modified cross-validation function has expectation, again ignoring terms that do not involve h, proportional to $\mathbb{E}[Q(X,h)|q_{X,\delta,l} < X < q_{X,\delta,r}]$. Choosing a smaller value of δ makes the expected value of the criterion closer to what we are ultimately interested, that is, $Q(c,h)$, but has the disadvantage of leading to a noisier estimate of $\mathbb{E}[\text{CV}_Y^\delta(h)]$. In practice one may wish to choose $\delta = 1/2$, and discard 50% of the observations on either side of the threshold, and afterwards assess the sensitivity of the bandwidth choice to the choice of δ. Ludwig and Miller (2005) implement this by using only data within 5 percentage points of the threshold on either side.

6.2 Bandwidth Selection for the FRD Design

In the FRD design, there are four regression functions that need to be estimated: the expected outcome given the forcing variable, both on the left and right of the cutoff point, and the expected value of the treatment, again on the left and right of the cutoff point. In principle, we can use different binwidths for each of the four nonparametric regressions.

In the section on the SRD design, we argued in favor of using identical bandwidths for the regressions on both sides of the cutoff point. The argument is not so clear for the pairs of regressions functions by outcome we have here, and so in principle we have two optimal bandwidths, one based on minimizing $\text{CV}_Y^\delta(h)$, and one based on minimizing $\text{CV}_W^\delta(h)$, defined correspondingly. It is likely that the conditional expectation of the treatment is relatively flat compared to the conditional expectation of the outcome variable, suggesting one should use a larger binwidth for estimating the former.[5] Nevertheless, in practice it is appealing

[5] In the extreme case of the SRD design the conditional expectation of W given X is flat on both sides

to use the same binwidth for numerator and denominator. Since typically the size of the discontinuity is much more marked in the expected value of the treatment, one option is to use the optimal bandwidth based on the outcome discontinuity. Alternatively, to minimize bias, one may wish to use the smallest bandwidths selected by the cross validation criterion applied separately to the outcome and treatment regression:

$$h_{\rm CV}^{\rm opt} = \min\left(\arg\min_h {\rm CV}_Y^\delta(h), \arg\min_h {\rm CV}_W^\delta(h)\right),$$

where ${\rm CV}_Y^\delta(h)$ is as defined in (12), and ${\rm CV}_W^\delta(h)$ is defined similarly. Again a value of $\delta = 1/2$ is likely to lead to reasonable estimates in many settings.

7. Inference

We now discuss some asymptotic properties for the estimator for the FRD case given in (7) or its alternative representation in (9).[6] More general results are given in HTV. We continue to make some simplifying assumptions. First, as in the previous sections, we use a uniform kernel. Second, we use the same bandwidth for the estimator for the jump in the conditional expectations of the outcome and treatment. Third, we undersmooth, so that the square of the bias vanishes faster than the variance, and we can ignore the bias in the construction of confidence intervals. Fourth, we continue to use the local linear estimator. Under these assumptions we give an explicit expression for the asymptotic variance, and present two estimators for the asymptotic variance. The first estimator follows explicitly the analytic form for the asymptotic variance, and substitutes estimates for the unknown quantities. The second estimator is the standard robust variance for the Two-Stage-Least-Squares (TSLS) estimator, based on the sample obtained by discarding observations when the forcing covariate is more than h away from the cutoff point. Both are robust to heteroskedasticity.

7.1 The Asymptotic Variance

To characterize the asymptotic variance we need a couple of additional pieces of notation.

of the threshold, and so the optimal bandwidth would be infinity. Therefore, in practice it is likely that the optimal bandwidth would be larger for estimating the jump in the conditional expectation of the treatment than in estimating the jump in the conditional expectation of the outcome.

[6] The results for the SRD design are a special case of these for the FRD design.

Define the four variances

$$\sigma^2_{Yl} = \lim_{x \uparrow c} \text{Var}(Y_i | X_i = x), \qquad \sigma^2_{Yr} = \lim_{x \downarrow c} \text{Var}(Y_i | X_i = x),$$

$$\sigma^2_{Wl} = \lim_{x \uparrow c} \text{Var}(W_i | X_i = x), \qquad \sigma^2_{Wr} = \lim_{x \downarrow c} \text{Var}(W_i | X_i = x),$$

and the two covariances

$$C_{YWl} = \lim_{x \uparrow c} \text{Cov}(Y_i, W_i | X_i = x), \qquad C_{YWr} = \lim_{x \downarrow c} \text{Cov}(Y_i, W_i | X_i = x).$$

Note that because of the binary nature of W, it follows that $\sigma^2_{Wl} = \mu_{Wl} \cdot (1 - \mu_{Wl})$, where $\mu_{Wl} = \lim_{x \uparrow c} \Pr(W_i = 1 | X_i = x)$, and similarly for σ^2_{Wr}. To discuss the asymptotic variance of $\hat{\tau}$ it is useful to break it up in three pieces. The asymptotic variance of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ is

$$V_{\tau_y} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Yr} + \sigma^2_{Yl} \right). \tag{15}$$

The asymptotic variance of $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$V_{\tau_w} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Wr} + \sigma^2_{Wl} \right) \tag{16}$$

The asymptotic covariance of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ and $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$C_{\tau_y, \tau_w} = \frac{4}{f_X(c)} \cdot \left(C_{YWr} + C_{YWl} \right). \tag{17}$$

Finally, the asymptotic distribution has the form

$$\sqrt{Nh} \cdot (\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\tau^2_w} \cdot V_{\tau_y} + \frac{\tau^2_y}{\tau^4_w} \cdot V_{\tau_w} - 2 \cdot \frac{\tau_y}{\tau^3_w} \cdot C_{\tau_y, \tau_w} \right). \tag{18}$$

This asymptotic distribution is a special case of that in HTV (page 208), using the rectangular kernel, and with $h = N^{-\delta}$, for $1/5 < \delta < 2/5$ (so that the asymptotic bias can be ignored).

7.2 A Plug-in Estimator for the Asymptotic Variance

We now discuss two estimators for the asymptotic variance of $\hat{\tau}$. First, we can estimate the asymptotic variance of $\hat{\tau}$ by estimating each of the components, τ_w, τ_y, V_{τ_w}, V_{τ_y}, and C_{τ_y, τ_w} and substituting them into the expression for the variance in (18). In order to do this we first estimate the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{\mu}_y(X_i) = Y_i - 1\{X_i < c\} \cdot \hat{\alpha}_{yl} - 1\{X_i \geq c\} \cdot \hat{\alpha}_{yr},$$

$$\hat{\eta}_i = W_i - \hat{\mu}_w(X_i) = W_i - 1\{X_i < c\} \cdot \hat{\alpha}_{wl} - 1\{X_i \geq c\} \cdot \hat{\alpha}_{wr}.$$

Then we estimate the variances and covariances consistently as

$$\hat{\sigma}^2_{Yl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\varepsilon}_i^2, \qquad \hat{\sigma}^2_{Yr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\varepsilon}_i^2,$$

$$\hat{\sigma}^2_{Wl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\eta}_i^2, \qquad \hat{\sigma}^2_{Wr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\eta}_i^2,$$

$$\hat{C}_{YWl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\varepsilon}_i \cdot \hat{\eta}_i, \qquad \hat{C}_{YWr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\varepsilon}_i \cdot \hat{\eta}_i.$$

Finally, we estimate the density consistently as

$$\hat{f}_X(x) = \frac{N_{hl} + N_{hr}}{2 \cdot N \cdot h}.$$

Then we can plug in the estimated components of V_{τ_y}, V_{τ_W}, and C_{τ_Y,τ_W} from (15)-(17), and finally substitute these into the variance expression in (18).

7.3 The TSLS Variance Estimator

The second estimator for the asymptotic variance of $\hat{\tau}$ exploits the interpretation of the $\hat{\tau}$ as a TSLS estimator, given in (9). The variance estimator is equal to the robust variance for TSLS based on the subsample of observations with $c - h \leq X_i \leq c + h$, using the indicator $1\{X_i \geq c\}$ as the excluded instrument, the treatment W_i as the endogenous regressor and the V_i defined in (8) as the exogenous covariates.

8. Specification Testing

There are generally two main conceptual concerns in the application of RD designs, sharp or fuzzy. A first concern about RD designs is the possibility of other changes at the same cutoff value of the covariate. Such changes may affect the outcome, and these effects may be attributed erroneously to the treatment of interest. The second concern is that of manipulation of the covariate value.

8.1 Tests Involving Covariates

One category of tests involves testing the null hypothesis of a zero average effect on pseudo outcomes known not to be affected by the treatment. Such variables includes covariates that are by definition not affected by the treatment. Such tests are familiar from settings with identification based on unconfoundedness assumptions. In most cases, the reason for the discontinuity in the probability of the treatment does not suggest a discontinuity in the average value of covariates. If we find such a discontinuity, it typically casts doubt on the assumptions underlying the RD design. See the second part of the Lee (2007) figure for an example.

8.2 Tests of Continuity of the Density

The second test is conceptually somewhat different, and unique to the RD setting. McCrary (2007) suggests testing the null hypothesis of continuity of the density of the covariate that underlies the assignment at the discontinuity point, against the alternative of a jump in the density function at that point. Again, in principle, one does not need continuity of the density of X at c, but a discontinuity is suggestive of violations of the no-manipulation assumption. If in fact individuals partly manage to manipulate the value of X in order to be on one side of the boundary rather than the other, one might expect to see a discontinuity in this density at the discontinuity point.

8.3 Testing for Jumps at Non-discontinuity Points

Taking the subsample with $X_i < c$ we can test for a jump in the conditional mean of the outcome at the median of the forcing variable. To implement the test, use the same method for selecting the binwidth as before. Also estimate the standard errors of the jump and use this to test the hypothesis of a zero jump. Repeat this using the subsample to the right of the cutoff point with $X_i \geq c$. Now estimate the jump in the regression function and at $q_{X,1/2,r}$, and test whether it is equal to zero.

8.4 RD Designs with Misspecification

Lee and Card (2007) study the case where the forcing variable variable X is discrete. In practice this is of course always true. This implies that ultimately one relies for identification

on functional form assumptions for the regression function $\mu(x)$. Lee and Card consider a parametric specification for the regression function that does not fully saturate the model, that is, it has fewer free parameters than there are support points. They then interpret the deviation between the true conditional expectation $\mathbb{E}[Y|X=x]$ and the estimated regression function as random specification error that introduces a group structure on the standard errors. Lee and Card then show how to incorporate this group structure into the standard errors for the estimated treatment effect. Within the local linear regression framework discussed in the current paper one can calculate the Lee-Card standard errors (possibly based on slightly coarsened covariate data if X is close to continuous) and compare them to the conventional ones.

8.5 Sensitivity to the Choice of Bandwidth

One should investigate the sensitivity of the inferences to this choice, for example, by including results for bandwidths twice (or four times) and half (or a quarter of) the size of the originally chosen bandwidth. Obviously, such bandwidth choices affect both estimates and standard errors, but if the results are critically dependent on a particular bandwidth choice, they are clearly less credible than if they are robust to such variation in bandwidths.

8.6 Comparisons to Estimates Based on Unconfoundedness in the FRD Design

If we have an FRD design, we can also consider estimates based on unconfoundedness. Inspecting such estimates and especially their variation over the range of the covariate can be useful. If we find that for a range of values of X, our estimate of the average effect of the treatment is relatively constant and similar to that based on the FRD approach, one would be more confident in both sets of estimates.

References

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J.D. and A.B. Krueger, (1991), Does Compulsory School Attendance Affect Schooling and Earnings?, *Quarterly Journal of Economics* 106, 979-1014.

Angrist, J.D., and V. Lavy, (1999), Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics* 114, 533-575.

Black, S., (1999), Do Better Schools Matter? Parental Valuation of Elementary Education, *Quarterly Journal of Economics* 114, 577-599.

Card, D., A. Mas, and J. Rothstein, (2006), Tipping and the Dynamics of Segregation in Neighborhoods and Schools, Unpublished Manuscript, Department of Economics, Princeton University.

Chay, K., and M. Greenstone, (2005), Does Air Quality Matter; Evidence from the Housing Market, *Journal of Political Economy* 113, 376-424.

Cook, T., (2007), "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, forthcoming, *Journal of Econometrics*.

DiNardo, J., and D.S. Lee, (2004), Economic Impacts of New Unionization on Private Sector Employers: 1984-2001, *Quarterly Journal of Economics* 119, 1383-1441.

Fan, J. and I. Gijbels, (1996), *Local Polynomial Modelling and Its Applications* (Chapman and Hall, London).

Hahn, J., P. Todd and W. Van der Klaauw, (2001), Identification and Estimation of Treatment Effects with a Regression Discontinuity Design, *Econometrica* 69, 201-209.

Härdle, W., (1990), *Applied Nonparametric Regression* (Cambridge University Press, New York).

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., and T. Lemieux, (2007) "Regression Discontinuity Designs: A Guide to Practice," forthcoming, *Journal of Econometrics*.

Lee, D.S. and D. Card, (2007), Regression Discontinuity Inference with Specification Error, forthcoming, *Journal of Econometrics*.

Lee, D.S., Moretti, E., and M. Butler, (2004), Do Voters Affect or Elect Policies? Evidence from the U.S. House, *Quarterly Journal of Economics* 119, 807-859.

Lemieux, T. and K. Milligan, (2007), Incentive Effects of Social Assistance: A Regression Discontinuity Approach, forthcoming, *Journal of Econometrics*.

Ludwig, J., and D. Miller, (2005), Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, NBER working paper 11702.

McCrary, J., (2007), Testing for Manipulation of the Running Variable in the Regression Discontinuity Design, forthcoming, *Journal of Econometrics*.

McEwan, P., and J. Shapiro, (2007),The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates using exact Birth Dates," Unpublished manuscript.

Pagan, A. and A. Ullah, (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

Porter, J., (2003), Estimation in the Regression Discontinuity Model," mimeo, Department of Economics, University of Wisconsin, http://www.ssc.wisc.edu/ jporter/reg_discont_2003.pdf.

Shadish, W., T. Campbell and D. Cook, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston).

Thistlewaite, D., and D. Campbell, (1960), Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment, *Journal of Educational Psychology* 51, 309-317.

Trochim, W., (1984), *Research Design for Program Evaluation; The Regression-discontinuity Design* (Sage Publications, Beverly Hills, CA).

Trochim, W., (2001), Regression-Discontinuity Design, in N.J. Smelser and P.B Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences* 19 (Elsevier North-Holland, Amsterdam) 12940-12945.

Van Der Klaauw, W., (2002), Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–discontinuity Approach, *International Economic Review* 43, 1249-1287.

What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 4, **Monday**, **July 30th**, **3.15-4.15 pm**

Nonlinear Panel Data Models

These notes summarize some recent, and perhaps not-so-recent, advances in the estimation of nonlinear panel data models. Research in the last 10 to 15 years has branched off in two directions. In one, the focus has been on parameter estimation, possibly only up to a common scale factor, in semiparametric models with unobserved effects (that can be arbitrarily correlated with covariates.) Another branch has focused on estimating partial effects when restrictions are made on the distribution of heterogeneity conditional on the history of the covariates. These notes attempt to lay out the pros and cons of each approach, paying particular attention to the tradeoff in assumptions and the quantities that can be estimated.

1. Basic Issues and Quantities of Interest

Most microeconomic panel data sets are best characterized as having few time periods and (relatively) many cross section observations. Therefore, most of the discussion in these notes assumes T is fixed in the asymptotic analysis while N is increasing. We assume random sample in the cross section, $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$. Take y_{it} to be a scalar for simplicity. If we are not concerned about traditional (contemporaneous) endogeneity, then we are typically interested in

$$D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \tag{1.1}$$

or some feature of this distribution, such as $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or a conditional median. In the case of a mean, how do we summarize the partial effects? Let $m_t(\mathbf{x}_t, \mathbf{c})$ be the mean function. If x_{tj} is continuous, then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}, \tag{1.2}$$

or look at discrete changes. How do we account for unobserved $\mathbf{c}_i$? If we want to estimate magnitudes of effects, we need to know enough about the distribution of $\mathbf{c}_i$ so that we can insert meaningful values for $\mathbf{c}$. For example, if $\boldsymbol{\mu}_\mathbf{c} = E(\mathbf{c}_i)$, then we can compute the *partial effect at the average (PEA)*,

$$\theta_j(\mathbf{x}_t, \boldsymbol{\mu}_\mathbf{c}). \tag{1.3}$$

Of course, we need to estimate the function m_t and the mean of $\mathbf{c}_i$. If we know more about the distribution of $\mathbf{c}_i$, we can insert different quantiles, for example, or a certain number of standard deviations from the mean.

Alternatively, we can average the partial effects across the distribution of $\mathbf{c}_i$:

$$\text{APE}(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)]. \tag{1.4}$$

The difference between (1.3) and (1.4) can be nontrivial for nonlinear mean functions. The definition in (1.4) dates back at least to Chamberlain (1982), and is closely related to the notion of the average structural function (ASF) (Blundell and Powell (2003)). The ASF is defined as

$$\text{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \tag{1.5}$$

Assuming the derivative passes through the expectation results in (1.5), the average partial effect. Of course, computing discrete changes gives the same result always. APEs are directly across models, and APEs in general nonlinear models are comparable to the estimated coefficients in a standard linear model.

Semiparametric methods, which, by construction, are silent about the distribution of $\mathbf{c}_i$, unconditionally or conditional on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, cannot generally deliver estimates of average partial (marginal) effects. Instead, an index structure is usually imposed so that parameters can be consistently estimated. So, for example, with scalar heterogeneity we might have an index model with additive heterogeneity:

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{1.6}$$

where, say, $G(\cdot)$ is strictly increasing and continuously differentiable (and, in some cases, is known, and in others, is not). Then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{1.7}$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Then estimating β_j means we can estimate the sign of the partial effect, and even the relative effects of any two continuous variables. But, even if $G(\cdot)$ is specified (the more common case), the magnitude of the effect evidently cannot be estimated without making assumptions about the distribution of c_i: the size of the scale factor for a random draw i, $g(\mathbf{x}_t\boldsymbol{\beta} + c_i)$, depends on c_i. Without knowing something about the distribution of c_i we cannot generally locate $g(\mathbf{x}_t\boldsymbol{\beta} + c_i)$.

Returning to the general case, Altonji and Matzkin (2005) focus on what they call the *local average response (LAR)* as opposed to the APE or PAE. The LAR at $\mathbf{x}_t$ for a continuous variable x_{tj} is

$$\int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \text{d}H_t(\mathbf{c}|\mathbf{x}_t), \tag{1.8}$$

where $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $\text{D}(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. This is a "local" partial effect because it averages out the heterogeneity for the slice of the population given by the vector $\mathbf{x}_t$. The APE,

which by comparison could be called a "global average response," averages out over the entire distribution of $\mathbf{c}_i$. See also Florens, Heckman, Meghir, and Vytlacil (2004).

It is important to see that the definitions of partial effects does not depend on the nature of the variables in $\mathbf{x}_t$ (except for whether it makes sense to use the calculus approximation or use changes). In particular, $\mathbf{x}_t$ can include lagged dependent variables and lags of other variables, which may or may not be strictly exogenous. Unfortunately, we cannot identify the APEs, or even relative effects in index models, without some assumptions.

2. Exogeneity Assumptions on the Covariates

Ideally, we would only have to specify a model for $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ or some feature. Unfortunately, it is well known that specifying a full parametric model is not sufficient for identifying either the parameters of the model or the partial effects defined in Section 1. In this section, we consider two useful exogeneity assumptions on the covariates.

It is easiest to deal with estimation under a strict exogeneity assumption. The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \tag{2.1}$$

which means that $\mathbf{x}_{ir}$, $r \neq t$, does not appear in the conditional distribution of $\mathbf{y}_{it}$ once $\mathbf{x}_{it}$ and $\mathbf{c}_i$ have been counted for. Chamberlain (1984) labeled (2.1) *strict exogeneity conditional on the unobserved (or latent) effects* $\mathbf{c}_i$. Sometimes, a conditional mean version is sufficient:

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \tag{2.2}$$

which we already saw for linear models. (In other cases a condition stated in terms of conditional medians is more convenient.) Of course, either version of the assumption rules out lagged dependent variables, but also other situations where there may be feedback from idiosyncratic changes in y_{it} to future movements in $\mathbf{x}_{ir}$, $r > t$. But it is the assumption underlying the most common estimation methods for nonlinear models.

More natural is a *sequential exogeneity* assumption (conditional on the unobserved effects) assumption, which we can state generally as

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \tag{2.3}$$

or, sometimes, in terms of specific features of the distribution. Assumption (2.3) allows for lagged dependent variables and does not restrict feedback. Unfortunately, it is much more difficult to allow, especially in nonlinear models.

Neither (2.2) nor (2.3) allows for contemporaneous endogeneity of one or more elements of

$\mathbf{x}_{it}$, where, say, x_{itj} is correlated with unobserved, time-varying unobservables that affect y_{it}, or where x_{itj} is simultaneously determined along with y_{it}. This will be covered in later notes on control function methods.

3. Conditional Independence Assumption

In some cases – certainly traditionally – a conditional independence assumption is imposed. We can write the condition generally as

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_i,\mathbf{c}_i). \tag{3.1}$$

This assumption is only useful in the context of the strict exogeneity assumption (2.1), in which case we can write

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i). \tag{3.2}$$

In a parametric context, the conditional independence assumption therefore reduces our task to specifying a model for $D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, $\mathbf{c}_i$. In random effects and CRE frameworks, conditional independence plays a critical role in being able to estimate the parameters in distribution the of $\mathbf{c}_i$. We could get by with less restrictive assumptions by parameterizing the dependence, but that increases computational burden. As it turns out, conditional independence plays no role in estimating APEs for a broad class of models. (That is, we do not need to place restrictions on $D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i)$.) Before we can study estimation, we must discuss the critical issue of the dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$.

4. Assumptions about the Unobserved Heterogeneity

The modern approach to panel data analysis with micro data treats the unobserved heterogeneity as random draws along with the observed data, and that is the view taken here. Nevertheless, there are still reasons one might treat them as parameters to estimate, and we allow for that in our discussion.

Random Effects

For general nonlinear models, what we call the "random effects" assumption is independence between $\mathbf{c}_i$ and $\mathbf{x}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$:

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{4.1}$$

If we combine this assumption with a model for $m_t(\mathbf{x}_t, \mathbf{c})$, then the APEs are actually nonparametrically identified. (And, in fact, we do not need to assume strict or sequential exogeneity to use a pooled estimation method, or to use just a single time period.) In fact, if $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) = m_t(\mathbf{x}_{it}, \mathbf{c}_i)$ and $D(\mathbf{c}_i|\mathbf{x}_{it}) = D(\mathbf{c}_i)$, then the APEs are obtained from

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{4.2}$$

(The argument is a simple application of the law of interated expectations; it is discussed in detail in Wooldridge (2005a).) In principle, $E(y_{it}|\mathbf{x}_{it})$ can be estimated nonparametrically, and we only need a single time period to identify the partial effects for that time period.

In some leading cases (for example random effects probit and Tobit with heterogeneity normally distributed), if we want to obtain partial effects for different values of $\mathbf{c}$, we must assume more: the strict exogeneity assumption (2.1), the conditional independence assumption (3.1), and the random effects assumption (4.1) with a parametric distribution for $D(\mathbf{c}_i)$ are typically sufficient. We postpone this discussion because it takes us into the realm of specifying parametric models.

Correlated Random Effects

A "correlated random effects" framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but the dependence in restricted in some way. In a parametric setting, we specify a distribution for $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, as in Mundlak (1978), Chamberlain (1982), and many subsequent authors. For many models, including probit and Tobit, one can allow $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ to depend in a "nonexchangeable" manner on the time series of the covariates; Chamberlain's random effects probit model does this. But the distributional assumptions that lead to simple estimation – namely, homoskedastic normal with a linear conditional mean — are restrictive. But it is aslo possible to assume

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i) \tag{4.3}$$

without specifying $\mathrm{D}(c_i|\bar{\mathbf{x}}_i)$ or restricting any feature of this distribution. We will see in the next section that (4.3) is very powerful.

We can go further. For example, suppose that we think the heterogeneity $\mathbf{c}_i$ is correlated with features of the covariates other than just the time average. Altonji and Matzkin (2005) allow for $\bar{\mathbf{x}}_i$ in equation (4.3) to be replaced by other functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, such as sample variances and covariance. These are examples of "exchangeable" functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ – that is, statistics whose value is the same regardless of the ordering of the

$\mathbf{x}_{it}$. Non-exchangeable functions can be used, too. For example, we might think that $\mathbf{c}_i$ is correlated with individual-specific trends, and so we obtain $\mathbf{w}_i$ as the intercept and slope from the unit-specific regressions $\mathbf{x}_{it}$ on 1, t, $t = 1, \ldots, T$ (for $T \geq 3$); we can also add the error variance from this individual specific regression if we have sufficient time periods. Then, the condition becomes

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\mathbf{w}_i). \tag{4.4}$$

Practically, we need to specify $\mathbf{w}_i$ and then establish that there is enough variation in $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ separate from $\mathbf{w}_i$; this will be clear in the next section.

Fixed Effects

Unfortunately, the label "fixed effects" is used in different ways by different researchers (and, sometimes, by the same researcher). The traditional view was that a fixed effects framework meant $\mathbf{c}_i$, $i = 1, \ldots, N$ were treated as parameters to estimate. This view is still around, and, when researchers say they estimated a nonlinear panel data model by "fixed effects," they sometimes mean the $\mathbf{c}_i$ were treated as parameters to estimate along with other parameters (whose dimension does not change with N). As is well known, except in special cases, estimation of the $\mathbf{c}_i$ generally introduces an "incidental parameters" problem. (More on this later when we discuss estimation methods, and partial effects.) Subject to computational feasilibity, the approach that treats the $\mathbf{c}_i$ as parameters is widely applicable.

The "fixed effects" label can mean that $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted. Even in that case, there are different approaches to estimation of parameters. One is to specify a joint distribution $D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{c}_i)$ such that a sufficient statistic, say $\mathbf{s}_i$, can be found such that

$$D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i) = D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{s}_i), \tag{4.5}$$

and where the latter distribution still depends on the parameters of interest in a way that identifies them. In most cases, the conditional independence assumption (3.1) is maintained, although one CMLE is known to have robustness properties: the so-called "fixed effects" Poisson estimator. We cover that later on.

5. Nonparametric Identification of Average Partial and Local Average Effects

Before considering identification and estimation of parameters in parametric models, it is useful to ask which quantities, if any, are identified without imposing parametric assumptions. Not surprisingly, there are no known results on nonparametric identificiation of partial effects

evaluated at specific values of $\mathbf{c}$, such as $\boldsymbol{\mu}_{\mathbf{c}}$ – except, of course, when the partial effects do not depend on $\mathbf{c}$. Interestingly, identification can fail even under a full set of strong parametric assumptions. For example, in the probit model

$$P(y = 1|\mathbf{x}, c) = \Phi(\mathbf{x}\boldsymbol{\beta} + c), \tag{5.1}$$

where $\mathbf{x}$ is $1 \times K$ an includes unity, the partial effect for a continuous variable x_j is simply $\beta_j \phi(\mathbf{x}\boldsymbol{\beta} + c)$. The partial effect at the mean of c is simply $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$. Suppose we assume that $c|\mathbf{x}$ ~Normal$(0, \sigma_c^2)$. Then it is easy to show that

$$P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}), \tag{5.2}$$

which means that only the scaled parameter vector $\boldsymbol{\beta}_c \equiv \boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}$ is identified. Therefore, $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$, is evidently unidentified. (The fact that probit of y on $\mathbf{x}$ estimates $\boldsymbol{\beta}_c$ has been called the "attenuation bias" that results from omitted variables in the context of probit, even when the omitted variable is independent of the covariates and normally distributed. As mentioned earlier more generally, the average partial effects are obtained directly from $P(y = 1|\mathbf{x})$, and, in fact, are given by $\beta_{cj} \phi(\mathbf{x}\boldsymbol{\beta}_c)$. As discussed in Wooldridge (2002, Chapter 15), $\beta_{cj} \phi(\mathbf{x}\boldsymbol{\beta}_c)$ can be larger or smaller in magnitude than the PEA $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$: $|\beta_{cj}| \leq |\beta_j|$ but $\phi(\mathbf{x}\boldsymbol{\beta}_c) \geq \phi(\mathbf{x}\boldsymbol{\beta})$.)

A related example is due to Hahn (2001), and is related to the nonidentification restuls of Chamberlain (1993). Suppose that x_{it} is a binary indicator (for example, a policy variable). Consider the unobserved effects probit model

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = \Phi(\beta x_{it} + c_i), \tag{5.3}$$

As discussed by Hahn, β is not known to be identified in this model, even under conditional serial independence assumption *and* the random effects assumption $D(c_i|\mathbf{x}_i) = D(c_i)$. But the average partial effect, which in this case is an average treatment effect, is simply $\tau \equiv E[\Phi(\beta + c_i)] - E[\Phi(c_i)]$. By the general result cited earlier, τ is consistently estimated (in fact, unbiasedly estimated) by using a difference of means for the treated and untreated groups, for either time period. In fact, as discussed in Wooldridge (2005a), identification of the APE holds if we replace Φ with an unknown function G and allow $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. But the parameters are still not identified.

To summarize: the APE is identified for any function $G(\cdot)$ whether or not the conditional serial independence holds, even if we add separate year intercepts. But β is not identified under the strongest set of assumptions. This simple example suggests that perhaps our focus on parameters is wrong-headed.

We can establish identification of average partial effects much more generally. Assume only that the strict exogeneity assumption (2.1) holds along with $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. These two assumptions are sufficient to identify the APEs. To see why, note that the average structural function at time *t* can be written as

$$\text{ASF}_t(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] = E_{\bar{\mathbf{x}}_i}\{E[m_t(\mathbf{x}_t, \mathbf{c}_i)|\bar{\mathbf{x}}_i]\} \equiv E_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)], \tag{5.4}$$

where $r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i) \equiv E[r_t(\mathbf{x}_t, \mathbf{c}_i)|\bar{\mathbf{x}}_i]$. It follows that, given an estimator $\hat{r}_t(\cdot, \cdot)$ of the function $r_t(\cdot, \cdot)$, the ASF can be estimated as

$$\widehat{\text{ASF}}_t(\mathbf{x}_t) \equiv N^{-1} \sum_{i=1}^{N} \hat{r}_t(\mathbf{x}_t, \bar{\mathbf{x}}_i), \tag{5.5}$$

and then we can take derivatives or changes with respect to the entries in $\mathbf{x}_t$. Notice that (5.4) holds without the strict exogeneity assumption (2.1) or the assumption $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. However, these assumptions come into play in our ability to estimate $r_t(\cdot, \cdot)$. If we combine (21) and (4.3) we have

$$\begin{aligned} E(y_{it}|\mathbf{x}_i) &= E[E(y_{it}|\mathbf{x}_i, \mathbf{c}_i)|\mathbf{x}_i] = E[m_t(\mathbf{x}_{it}, \mathbf{c}_i)|\mathbf{x}_i] = \int m_t(\mathbf{x}_{it}, \mathbf{c}) dF(\mathbf{c}|\mathbf{x}_i) \\ &= \int m_t(\mathbf{x}_{it}, \mathbf{c}) dF(\mathbf{c}|\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i), \end{aligned} \tag{5.6}$$

where $F(\mathbf{c}|\mathbf{x}_i)$ denotes the cdf of $D(\mathbf{c}_i|\mathbf{x}_i)$ (which can be a discrete, continuous, or mixed distribution), the second equality follows from (2.1), the fourth equality follows from assumption (4.3), and the last equality folllows from the definition of $r_t(\cdot, \cdot)$ Of course, because $E(y_{it}|\mathbf{x}_i)$ depends only on $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$, we must have

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i). \tag{5.7}$$

Further, $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is assumed to have time variation, and so $\mathbf{x}_{it}$ and $\bar{\mathbf{x}}_i$ can be used as separate regressors even in a fully nonparametric setting.

Altonji and Matskin (2005).use this idea more generally, and focus on estimating the local average response. Wooldridge (2005a) used $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$ generally in the case $\mathbf{x}_{it}$ is discrete, in which case a full nonparametric analysis is easy. When

$$D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\mathbf{w}_i) \tag{5.8}$$

for $\mathbf{w}_i$ a function of $\mathbf{x}_i$, Altonji and Matzkin (2005) show that the LAR can be obtained as

$$\int \frac{\partial r_t(\mathbf{x}_t, \mathbf{w})}{\partial x_{tj}} dK_t(\mathbf{w}|\mathbf{x}_t), \tag{5.9}$$

where $r(\mathbf{x}_t, \mathbf{w}) = \mathrm{E}(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{w}_i = \mathbf{w})$ and $K_t(\mathbf{w}|\mathbf{x}_t)$ is the cdf of $\mathrm{D}(\mathbf{w}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. Altonji and Matskin demonstrate how to estimate the LAR based on nonparametric estimation of $\mathrm{E}(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i)$ followed by "local" averaging, that is, averaging $\partial r(y_{it}|\mathbf{x}_t, \mathbf{w}_i)/\partial x_{tj}$ over observations i with $\mathbf{x}_{it}$ "close" to $\mathbf{x}_t$.

This analysis demonstrates that APEs are nonparametrically identified under the conditional mean version of strict exogeneity, (2.1), and (5.8), at least for time-varying covariates if $\mathbf{w}_i$ is restricted in some way. In fact, we can identify the APEs for a single time period with just one year of data on y. We only need to obtain $\bar{\mathbf{x}}_i$ and, in effect, include it as a control. Of course, efficiency would be gained by assuming some stationarity across t and using a pooled method.

6. Dynamic Models

General models with only sequentially exogenous variables are difficult to deal with. Arellano and Carrasco (2003) consider probit models. Wooldridge (2000) suggests a strategy the requires modeling the dynamic distribution of the variables that are not strictly exogenous. Much more is known about models with lagged dependent variables and otherwise strictly exogenous variables. So, we start with a model for

$$D(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \ldots, \mathbf{z}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i), \; t = 1, \ldots, T, \tag{6.1}$$

which we assume also is $D(\mathbf{y}_{it}|\mathbf{z}_i, \mathbf{y}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i)$ where $\mathbf{z}_i$ is the entire history $\{z_{it} : t = 1, \ldots, T\}$. This is the sense in which the z_{it} are strictly exogenous.

Suppose this model depends only on $(\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i)$, so $f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta})$. The joint density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $(\mathbf{y}_{i0}, \mathbf{z}_i, \mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}). \tag{6.2}$$

The problem with using this for estimation is the presence of $\mathbf{c}_i$ along with the initial condition, $\mathbf{y}_{i0}$. Approaches: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (incidental parameters problem, although recent research has attempted to reduce the asymptotic bias in the partial effects). (ii) Try to estimate the parameters without specifying conditional or unconditional distributions for c_i. (Available in some special cases covered below, but other restrictions are needed. And, generally, cannot estimate partial effects.). (iii) Find or, more practically, approximate $D(\mathbf{y}_{i0}|\mathbf{c}_i, z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. After integrating out c_i we obtain the density for $D(\mathbf{y}_{i0}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{z}_i)$ and we can use MLE (conditional on z_i), (iv) Model $D(\mathbf{c}_i|\mathbf{y}_{i0}, \mathbf{z}_i)$. After

integrating out c_i we obtain the density for $D(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{y}_{i0},\mathbf{z}_i)$, and we can use MLE (conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$). As shown by Wooldridge (2005b), in some leading cases – probit, ordered probit, Tobit, Poisson regression – there is a density $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z})$ that mixes with the density $f(\mathbf{y}_1,\ldots,\mathbf{y}_T|\mathbf{y}_0,\mathbf{z},\mathbf{c})$ to produce a log-likelihood that is in a common family and carried out by standard software.

If $m_t(\mathbf{x}_t,\mathbf{c},\boldsymbol{\theta})$ is the mean function $E(y_t|\mathbf{x}_t,\mathbf{c})$ for a scalar y_t, then average partial effects are easy to obtain. The average structural function is

$$ASF(\mathbf{x}_t) = E_{c_i}[m_t(\mathbf{x}_t,\mathbf{c}_i,\theta)] = E\left\{\left[\int m_t(\mathbf{x}_t,\mathbf{c},\boldsymbol{\theta})h(\mathbf{c}|y_{i0},\mathbf{z}_i,\boldsymbol{\gamma})d\mathbf{c}\right]\Big|y_{i0},\mathbf{z}_i\right\}. \tag{6.3}$$

The term inside the brackets, say $r_t(x_t,y_{i0},z_i,\theta,\gamma)$ is available, at least in principle, because $m_t()$ and $h()$ have been specified. Often, they have simple forms, in fact. Generally, it can be simulated. In any case, $ASF(\mathbf{x}_t,\boldsymbol{\theta})$ is consistently estimated by

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{t=1}^{T} r_t(\mathbf{x}_t,y_{i0},\mathbf{z}_i,\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}}).$$

Partial derivatives and differences with respect to elements of x_t (which, remember, can include y_{t-1}) can be computed. With large N and small T, the panel data bootstrap can be used for standard errors and inference.

7. Applications to Specific Models

We now turn to some common parametric models and highlight the difference between estimation partial effects at different values of the heterogeneity and estimating average partial effects. An analysis of Tobit models follows in a very similar way to those in the following two sections. See Wooldridge (2002, Chapter 16) and Honoré and Hu (2004).

7.1 Binary and "Fractional" Response Models

We start with the standard specification for the unobserved effects (UE) probit model, which is

$$P(y_{it} = 1|\mathbf{x}_{it},c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i),\ t = 1,\ldots,T, \tag{7.1}$$

where $\mathbf{x}_{it}$ does not contain an overall intercept but would usually include time dummies. We cannot identify $\boldsymbol{\beta}$ or the APEs without further assumptions. The traditional RE probit models imposes a strong set of assumptions: strict exogeneity, conditional serial independence, and independence between c_i and $\mathbf{x}_i$ with c_i ~Normal(μ_c,σ_c^2). Under these assumptions, $\boldsymbol{\beta}$ and the parameters in the distribution of c_i are identified and are consistently estimated by full MLE

(conditional on $\mathbf{x}_i$).

We can relax independence between c_i and $\mathbf{x}_i$ using the Chamberlain-Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, a_i|\mathbf{x}_i \sim \text{Normal}(0, \sigma_a^2), \tag{7.2}$$

where the time average is often used to save on degrees of freedom. We can relax (7.2) and allow Chamberlain's (1980) more flexible device:

$$c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i = \psi + \mathbf{x}_{i1}\boldsymbol{\xi}_1 + \ldots + \mathbf{x}_{iT}\boldsymbol{\xi}_T + a_i \tag{7.3}$$

Even when the $\boldsymbol{\xi}_r$ seem to be very different, the Mundlak restriction can deliver similar estimates of the other parameters and the APEs. (In the linear case, they both produce the usual FE estimator of $\boldsymbol{\beta}$.)

If we still assume conditional serial independence then all parameters are identified. We can estimate the mean of c_i as $\hat{\mu}_c = \hat{\psi} + \left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}$ and the variance as $\hat{\sigma}_c^2 \equiv \hat{\boldsymbol{\xi}}'\left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i'\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Of course, c_i is not generally normally distributed unless $\bar{\mathbf{x}}_i\boldsymbol{\xi}$ is. The approximation might get better as T gets large. In any case, we can plug in values of c that are a certain number of estimated standard deviations from $\hat{\mu}_c$, say $\hat{\mu}_c \pm \hat{\sigma}_c$.

The APEs are identified from the ASF, which is consistently estimated as

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{7.4}$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1 + \hat{\sigma}_a^2)^{1/2}$. The derivatives or changes of $\widehat{\text{ASF}}(\mathbf{x}_t)$ with respect to elements of $\mathbf{x}_t$ can be compared with fixed effects estimates from a linear model. Often, if we also average out across $\mathbf{x}_{it}$, the linear FE estimates are similar to the averaged effects.

As we discussed generally in Section 5, the APEs are defined without the conditional serial independence assumption. Without $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it}, c_i)$, we can still estimate the scaled parameters because

$$P(y_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a), \tag{7.5}$$

and so pooled probit consistently estimates the scaled parametes. (Time dummies have been supressed for simplicity.) Now we have direct estimates of $\boldsymbol{\beta}_a$, ψ_a, and $\boldsymbol{\xi}_a$, and we insert those directly into (7.4).

Using pooled probit can be inefficient for estimating the scaled parameters, whereas the

full MLE is efficient but not (evidently) robust to violation of the conditional serial independence assumption. It is possible to estimate the parameters more efficiently than pooled probit that is still consistent under the same set of assumptions. One possibility is minimum distance estimation. That is, estimate a separate models for each t, and then impose the restrictions using minimum distance methods. (This can be done with or without the Mundlak device.)

A different approach is to apply the so called "generalized estimating equations" (GEE) approach. Briefly, GEE applied to panel data is essentially weighted multivariate nonlinear least squares (WMNLS) with explicit recognition that the weighting matrix might not be the inverse of the conditional variance matrix. In most nonlinear panel data models, obtaining the actual matrix $Var(\mathbf{y}_i|\mathbf{x}_i)$ is difficult, if not impossible, because integrating out the heterogeneity does not deliver a closed form. The GEE approach uses $Var(y_{it}|\mathbf{x}_i)$ implied by the specific distribution – in the probit case, we have the correct conditional variances,

$$Var(y_{it}|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)[1 - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)] \equiv v_{it}. \tag{7.6}$$

The "working" correlation matrix oftenusually specified as "exchangeable,"

$$Corr(e_{it}, e_{is}|\mathbf{x}_i) \text{ " = "} \rho, \tag{7.7}$$

where $e_{it} = [y_{it} - \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)]v_{it}^{1/2}$ is the standardized error. Or, each pair (t, s) is allowed to have its own correlation but which is assumed to be independent of $\mathbf{x}_i$ ("unstructured"). The conditional correlation $Corr(e_{it}, e_{is}|\mathbf{x}_i)$ is not constant, but that is the working assumption. The hope is to improve efficiency over the pooled probit estimator while maintaining the robustness of the pooled estimator. (The full RE probit estimator is not robust to serial dependence.) A robust sandwich matrix is easily computed provided the conditional mean function (in this case, response probability) is correctly specified.

Because the Bernoulli log-likelihood is in the linear exponential family (LEF), exactly the same methods can be applied if $0 \leq y_{it} \leq 1$ – that is, y_{it} is a "fractional" response – but where the model is for the conditional mean: $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$. Pooled "probit" or minimum distance estimation or GEE can be used. Now, however, we must make inference robust to $Var(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ not having the probit form. (There are cases where $Var(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ is proportional to (7.6), and so it still makes sense to use the probit quasi-log-likelihood. Pooled nonlinear regression is another possibility or weighted multivariate nonlinear regression are also possible and a special case of GEE.)

A more radical suggestion, but in the spirit of Altonji and Matzkin (2005) and Wooldridge (2005a), is to just use a flexible model for $\mathrm{E}(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ directly. For example, if y_{it} is binary, or a fractional response, $0 \leq y_{it} \leq 1$, we might just specify a flexible parametric model, such as

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{7.8}$$

or the "heteroskedastic probit" model

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[(\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma})\exp(-\bar{\mathbf{x}}_i\boldsymbol{\eta})]. \tag{7.9}$$

If we write either of these functions as $r_t(\mathbf{x}_t, \bar{\mathbf{x}})$ then the average structural function is estimated as $\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1}\sum_{i=1}^{N} \hat{r}_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)$, where the "^" indicates that we have substituted in the parameter estimates. We can let all parameters depend on *t*, or we can estimate the parameters separately for each *t* and then use minimum distance estimation to impose the parameter restrictions. The justification for using, say, (7.8) is that we are interested in the average partial effects, and how parameters appear is really not the issue. Even though (7.8) cannot be derived from $E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$ or any other standard model, there is nothing sacred about this formulation. In fact, it is fairly simplistic. We can view (7.8) as the approximation to the true $E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ obtained after integrating c_i out of the unknown function $m(\mathbf{x}_t, c_i)$. (We could formalize this process by using series estimation, as in Newey (1988), where the number of terms is allowed to grow with *N*.) This is the same argument used by, say, Angrist (2001) in justifying linear models for limited dependent variables when the focus on primarily on average effects.

The argument is essentially unchanged if we replace $\bar{\mathbf{x}}_i$ with other statistics $\mathbf{w}_i$. For example, we might run, for each *i*, the regression $\mathbf{x}_{it}$ on $1, t$, $t = 1, \ldots, T$ and use the intercept and slope (on the time trend) as the elements of $\mathbf{w}_i$. Or, we can use sample variances and covariances for each *i*, along with the sample mean. Or, we can use initial values and average growth rates. The key condition is $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\mathbf{w}_i)$, and then we need sufficient variation in $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ not explained by $\mathbf{w}_i$ for identification. (Naturally, as we expand $\mathbf{w}_i$, the number of time periods required generally increases.)

Of course, once we just view (7.8) as an approximation, we can are justified in using the logistic function, say

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Lambda[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{7.10}$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$, which, again, can be applied to binary or fractional responses. The focus on partial effects that average out the heterogeneity can be liberating in

that it means the step of specifying $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ is largely superfluous, and, in fact, can get in the way of pursuing a suitably flexible analysis. On the other hand, if we start with, say, a "structural" model such as $P(y_{i1} = 1|\mathbf{x}_i, \mathbf{c}_i) = \Phi(a_i + \mathbf{x}_{it}\mathbf{b}_i)$, which is a heterogeneous index model, then we cannot derive equations such as (7.8) or (7.9), even under the strong assumption that $\mathbf{c}_i$ is independent of $\mathbf{x}_i$ and multivariate normal. If we imposed the Chamberlain device for the elements of $\mathbf{c}_i$ we can get expressions "close" to a combination of (7.8) and (7.9). Whether one is willing to simply estimate relative simple models such as (7.8) in order to estimate APEs depends on one's taste for bypassing more traditional formulations.

If we start with the logit formulation

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{7.11}$$

then we can estimate the parameters, $\boldsymbol{\beta}$ without restricting $D(c_i|\mathbf{x}_i)$ in any way, but we must add the conditional independence assumption. (No one has been able to show that, unlike in the linear model, or the Poisson model covered below, that the MLE that conditions on the number of successes $n_i = \sum_{t=1}^{T} y_{it}$ is robust to serial dependence. It appears not to be. Plus, the binary nature of y_{it} appears to be critical, so the conditional MLE cannot be applied to fractional responses even under serial independence.) Because we have not restricted $D(c_i|\mathbf{x}_i)$ in any way, it appears that we cannot estimate average partial effects. As commonly happens in nonlinear models, if we relax assumptions about the distribution of heterogeneity, we lose the ability to estimate partial effects. We can estimate the effects of the covariates on the log-odds ratio, and relative partial effects of continuous variables. But for partial effects themselves, we do not have sensible values to plug in for c, and we cannot average across its distribution.

The following table summarizes the features of various approaches to estimating binary response unobserved effects models.

Model, Estimation Method	$P(y_{it}= 1\|x_{it}, c_i)$	Restricts $D(c_i\|x_i)$?	Idiosyncratic Serial	PEs	APEs?
	Bounded in (0,1)?		Dependence?	at $E(c_i)$?	
RE Probit, MLE	Yes	Yes (indep, normal)	No	Yes	Yes
RE Probit, Pooled MLE	Yes	Yes (indep, normal)	Yes	No	Yes
RE Probit, GEE	Yes	Yes (indep, normal)	Yes	No	Yes
CRE Probit, MLE	Yes	Yes (lin. mean, normal)	No	Yes	Yes
CRE Probit, Pooled MLE	Yes	Yes (lin. mean, normal)	Yes	No	Yes
CRE Probit, GEE	Yes	Yes (lin. mean, normal)	Yes	No	Yes
LPM, Within	No	No	Yes	Yes	Yes
FE Logit, MLE	Yes	No	No	No	No

As an example, we apply several of the methods to women's labor force participation data, used by Chay and Hyslop (2001), where the data are for five time periods spaced four months apart. The results are summarized in the following table. The standard errors for the APEs were obtained with 500 bootstrap replications. The time-varying explanatory variables are log of husband's income and number of children, along with a full set of time period dummies. (The time-constant variables race, education, and age are also included in columns (2), (3), and (4).)

	(1)	(2)		(3)		(4)		(5)
Model	Linear	Probit		CRE Probit		CRE Probit		FE Logit
Estimation Method	Fixed Effects	Pooled MLE		Pooled MLE		MLE		MLE
	Coefficient	Coefficient	APE	Coefficient	APE	Coefficient	APE	Coefficient
kids	−.0389	−.199	−.0660	−.117	−.0389	−.317	−.0403	−.644
	(.0092)	(.015)	(.0048)	(.027)	(.0085)	(.062)	(.0104)	(.125)
lhinc	−.0089	−.211	−.0701	−.029	−.0095	−.078	−.0099	−.184
	(.0046)	(.024)	(.0079)	(.014)	(.0048)	(.041)	(.0055)	(.083)
$\overline{kids}$	—	—	—	−.086	—	−.210	—	—
	—	—	—	(.031)	—	(.071)	—	—
$\overline{lhinc}$	—	—	—	−.250	—	−.646	—	—
	—	—	—	(.035)	—	(.079)	—	—
$(1+\hat{\sigma}_a^2)^{-1/2}$	—	—		—		.387		—
Log Likelihood	—	−16,556.67		−16,516.44		−8,990.09		−2,003.42
Number of Women	5,663	5,663		5,663		5,663		1,055

Generally, CMLE approaches are fragile to changes in the specification. For example, a natural extension is

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Lambda(a_i + \mathbf{x}_{it}\mathbf{b}_i), \tag{7.12}$$

where $\mathbf{b}_i$ is a vector of heterogeneous slopes with $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$; let $\alpha \equiv E(a_i)$. This extension of the standard unobserved effects logit model raises several issues. First, what do we want to estimate? Perhaps the partial effects at the mean values of the heterogeneity. But the APEs, or local average effects, are probably of more interest.

Nothing seems to be known about what the logit CMLE would estimate if applied to (7.12), where we assume $\boldsymbol{\beta} = \mathbf{b}_i$. On the other hand, if, say, $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$, a flexible binary response model with covariates $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ (and allowing sufficiently for changes over time) identifies the APEs – without the conditional serial independence assumption. The same is true of the extension to time-varying factor loads, $P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Lambda(\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t c_i)$.

There are methods that allow estimation, up to scale, of the coefficients without even specifying the distribution of u_{it} in

$$y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]. \tag{7.13}$$

under strict exogeneity.conditional on c_i. Arellano and Honoré (2001) survey methods,

including variations on Manski's maximum score estimator.

Estimation of parameters and APEs is much more difficult even in simple dynamic models. Consider

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, c_i) = P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i), \; t = 1, \ldots, T,$$

which combines correct dynamic specification with strict exogeneity of $\{z_{it}\}$. For a dynamic probit model

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i). \tag{7.14}$$

Treating the c_i as parameters to estimate causes inconsistency in β and ρ (although there is recent work by Woutersen and Fernández-Val that shows how to make the asymptotic bias of order $1/T^2$; see the next section). A simple analysis is available if we specify

$$c_i|\mathbf{z}_i, y_{i0} \sim Normal(\psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi}, \sigma_a^2) \tag{7.15}$$

Then

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i), \tag{7.16}$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i\boldsymbol{\xi}$. Because a_i is independent of $(y_{i0}, \mathbf{z}_i)$, it turns out we can use standard random effects probit software, with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t. Easily get the average partial effects, too:

$$\widehat{ASF}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N} \Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a), \tag{7.17}$$

and take differences or derivatives with respect to elements of $(\mathbf{z}_t, y_{t-1})$. As before, the coefficients are multiplied by $(1 + \hat{\sigma}_a^2)^{-1/2}$. Of course, both the structural model and model for $D(c_i|y_{i0}, \mathbf{z}_i)$ can be made more flexible (such as including interactions, or letting $Var(c_i|\mathbf{z}_i, y_{i0})$ be heteroskedastic).

We apply this method to the Chay and Hyslop data and estimate a model for $\mathrm{P}(lfp_{it} = 1|kids_{it}, lhinc_{it}, lfp_{i,t-1}, c_i)$, where one lag of labor force participation is assumed to suffice for the dynamics and $\{(kids_{it}, lhinc_{it}) : t = 1, \ldots, T\}$ is assumed to be strictly exogenous conditional on c_i. Also, we include the time-constant variables *educ*, *black*, *age*, and age^2 and a full set of time-period dummies. (We start with five periods and lose one with the lag. Therefore, we estimate the model using four years of data.) We include among the regressors the initial value, lfp_{i0}, $kids_{i1}$ through $kids_{i4}$, and $lhinc_{i1}$ through $lhinc_{i4}$. Estimating the model by RE probit gives $\hat{\rho} = 1.541$ (se = .067), and so, even after controlling for

unobserved heterogeneity, there is strong evidence of state dependence. But to obtain the size of the effect, we compute the APE for lfp_{t-1}. The calculation involves averaging $\Phi(\mathbf{z}_{it}\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a) - \Phi(\mathbf{z}_{it}\hat{\boldsymbol{\delta}}_a + \hat{\xi}_{a0}y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a)$ across all t and i; we must be sure to scale the original coefficients by $(1 + \hat{\sigma}_a^2)^{-1/2}$, where, in this application, $\hat{\sigma}_a^2 = 1.103$. The APE estimated from this method is about .259. In other words, averaged across all women and all time periods, the probability of being in the labor force at time t is about .26 higher if the women was in the labor force at time $t - 1$ than if she was not. This estimate controls for unobserved heterogeneity, number of young children, husband's income, and the woman's education, race, and age.

It is instructive to compare the APE with the estimate of a dynamic probit model that ignores c_i. In this case, we just use pooled probit of lfp_{it} on $1, kids_{it}, lhinc_{it}, lfp_{i,t-1} educ_i, black_i, age_i$, and age_i^2 and include a full set of period dummies. The coefficient on $lfp_{i,t-1}$ is 2.876 (se = .027), which is much higher than in the dynamic RE probit model. More importantly, the APE for state dependence is about .837, which is much higher than when heterogeneity is controlled for. Therefore, in this example, much of the persistence in labor force participation of married women is accounted for by the unobserved heterogeneity. There is still some state dependence, but its value is much smaller than a simple dynamic probit indicates.

Arellano and Carrasco (2003) use a different approach to estimate the parameters and APEs in dynamic binary response models with only sequentially exogenous variables. Thus, their method applies to models with lagged dependent variables, but also other models where there made be feedback from past shocks to future covariates. (Their assumptions essentially impose serial conditional serial independence.) Rather than impose an assumption such as (7.15), they use a different approximation. Let $v_{it} = c_i + u_{it}$ be the composed error in $y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]$. Then, in the context of a probit model, they assume

$$v_{it}|w_{it} \sim \text{Normal}(E(c_i|w_{it}), \sigma_t^2) \tag{7.18}$$

where $w_{it} = (x_{it}, y_{i,t-1}, x_{i,t-1}, \ldots, y_{i1}x_{i1})$. The mean $E(c_i|w_{it})$ is unrestricted (although, of course, they are linked across time by interacted expectations because $w_{it} \subset w_{i,t+1}$), but the shape of the distribution is assumed to be the same across t. Arellano and Carrasco discuss identification and estimation, and extensions to models with time-varying factor loads.

Honoré and Kyriazidou (2000) extend an idea of Chamberlain's and show how to estimate $\boldsymbol{\delta}$ and ρ in a logit model without distributional assumptions for c_i. They find conditional

probabilities that do not depend on c_i but still depend on $\boldsymbol{\delta}$ and ρ. However, in the case with four time periods, $t = 0, 1, 2$, and 3, the conditioning that removes c_i requires $z_{i2} = z_{i3}$. HK show how to use a local version of this condition to consistenty estimate the parameters. The estimator is also asymptotically normal, but converges more slowly than the usual $\sqrt{N}$-rate.

The condition that $z_{i2} - z_{i3}$ have a distribution with support around zero rules out aggregate year dummies or even linear time trends. Plus, using only observations with $z_{i2} - z_{i3}$ in a neighborhood of zero results in much lost data. Finally, estimates of partial effects or average partial effects are not available.

While semiparametric approaches can be valuable to comparing parameter estimates with more parametric approaches, such comparisons have limitations. For example, the coefficients on y_{t-1} in the dynamic logit model and the dynamic probit model are comparable only in sign; we cannot take the derivative with respect to y_{t-1} because it is discrete. Because we do not know where the evaluate the partial effects – that is, the values of c to plug in, or average out across the distribution of c_i, we cannot compare the magnitudes with CRC approaches. We can compare the relative effects on the continuous elements in $\mathbf{z}_t$ based on partial derivatives. But even here, if we find a difference between semiparametric and parametric methods, is it because aggregate time effects were excluded in the semiparametric estimation or because the model of $D(c_i|y_{i0}, \mathbf{z}_i)$ was misspecified? Currently, we have no good ways of deciding. (Recently, Li and Zheng (2006) use Bayesian methods to estimate a dynamic Tobit model with unobserved heterogeneity, where they distribution of unosberved heterogeneity is an infinite mixture of normals. They find that all of the average partial effects are very similar to those obtained from the much simpler specification in (7.15).)

Honoré and Lewbel (2002) show how to estimate β in the model

$$y_{it} = 1[v_{it} + x_{it}\beta + c_i + u_{it} \geq 0] \tag{7.19}$$

without distributional assumptions on $c_i + u_{it}$. The special continuous explanatory variable v_{it}, which need not be time varying, is assumed to appear in the equation (and its coefficient is normalized to one). More importantly, v_{it} is assumed to satisfy $D(c_i + u_{it}|v_{it}, x_{it}, z_i) = D(c_i + u_{it}|x_{it}, z_i)$, which is a conditional independence assumption. The vector z_i is assumed to be independent of u_{it} in all time periods. (So, if two time periods are used, z_i could be functions of variables determined prior to the earliest time period.) The most likely scenario is when v_{it} is randomized and therefore independent of (x_{it}, z_i, e_{it}), where $e_{it} = c_i + u_{it}$. It seems unlikely to hold if v_{it} is related to past outcomes on y_{it}. The estimator

derived by Honoré and Lewbel is $\sqrt{N}$-asymptotically normal, and fairly easy to compute; it requires estimation of the density of v_{it} given (x_{it}, z_i) and then a simple IV estimation.

Honoré and Tamer (2006) have recently shown how to obtain bounds on parameters and APEs in dynamic models, including the dynamic probit model; these are covered in the notes on partial identification.

7.2 Count and Other Multiplicative Models

Several options are available for models with conditional means multiplicative in the heterogeneity. The most common is

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{7.20}$$

where $c_i \geq 0$ is the unobserved effect and x_{it} would incude a full set of year dummies in most cases. First consider estimation under strict exogeneity (conditional on c_i):

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i). \tag{7.21}$$

If we add independence between c_i and x_i – a random effects approach – then, using $E(c_i) = 1$ as a normalization,

$$E(y_{it}|\mathbf{x}_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}), \tag{7.22}$$

and various estimation methods can be used to account for the serial dependence in $\{y_{it}\}$ if only x_i is conditioned on. (Serial correlation is certainly present because of c_i, but it could be present due to idiosyncratic shocks, too.) Regardless of the actual distribution of y_{it}, or even its nature – other than $y_{it} \geq 0$ – the pooled Poisson quasi-MLE is consistent for β under (7.22) but likely very inefficient; robust inference is straightforward with small T and large N.

Random effects Poisson requires that $D(y_{it}|\mathbf{x}_i, c_i)$ has a Poisson distribution with mean (7.20), and maintains the conditional independence assumption,

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it}, c_i),$$

along with a specific distribution for c_i – usually a Gamma distribution with unit mean. Unfortunately, like RE probit, the full MLE has no known robustness properties. The Poisson distribution needs to hold along with the other assumptions. A generalized estimating approach is available, too. If the Poisson quasi-likelihood is used, the GEE estimator is fully robust provided the mean is correctly specified. One can use an exchangeable, or at least constant, working correlation matrix. See Wooldridge (2002, Chapter 19).

A CRE model can be allowed by writing $c_i = \exp(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi})a_i$ where a_i is independent of x_i

with unit mean. Then

$$E(y_{it}|\mathbf{x}_i) = \exp(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi}) \tag{7.23}$$

and now the same methods described above can be applied but with $\bar{x}_i$ added as regressors. This approach identifies average partial effects. In fact, we could use Altonji and Matzkin (2005) and specify $E(c_i|x_i) = h(\bar{\mathbf{x}}_i)$ (say), and then estimate the semiparametric model $E(y_{it}|\mathbf{x}_i) = h(\bar{\mathbf{x}}_i)\exp(\mathbf{x}_{it}\boldsymbol{\beta})$. Other features of the series $\{\mathbf{x}_{it} : t = 1,\ldots,T\}$, such as individual-specific trends or sample variances, can be added to $h(\cdot)$.

An important estimator that can be used under just

$$E(y_{it}|\mathbf{x}_i, c_i) = c_i\exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{7.24}$$

is the conditional MLE derived under a Poisson distributional assumption and the conditional independence assumption. It is often called the fixed effects Poisson estimator, and, in fact, $\hat{\boldsymbol{\beta}}$ turns out to be identical to using pooled Poisson QMLE and treating the c_i as parameters to estimate. (A rare case, like the linear model, where this does not result in an incidental parameters problem.). It is easy to obtain fully robust inference, too (although it is not currently part of standard software, such as Stata). The fact that the quasi-likelihood is derived for a particular, discrete distribution appears to make people queasy about using it, but it is analogous to using the normal log-likelihood in the linear model: the resulting estimator, the usual FE estimator, is fully robust to nonnormality, heteroskedasticity, and serial correlation.

Estimation of models under sequential exogeneity has been studied by Chamberlain (1992) and Wooldridge (1997). In particular, they obtain moment conditions for models such as

$$E(y_{it}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}, c_i) = c_i\exp(\mathbf{x}_{it}\boldsymbol{\beta}). \tag{7.25}$$

Under this assumption, it can be shown that

$$E\{[y_{it} - y_{i,t+1}\exp((\mathbf{x}_{it} - \mathbf{x}_{i,t+1})\boldsymbol{\beta})|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}] = 0, \tag{7.26}$$

and, because these moment conditions depend only on observed data and the parameter vector $\boldsymbol{\beta}$, GMM can be used to estimate $\boldsymbol{\beta}$, and fully robust inference is straightforward.

The moment conditions in (7.26) involve the differences $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$, and we saw for the linear model that, if elements of $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$ are persistent, IV and GMM estimators can be badly biased and imprecise. If we make more assumptions, models with lagged dependent variables and other regressors that are strictly exogenous can be handled using the conditional MLE approach in Section 6. Wooldridge (2005b) shows how a dynamic Poisson model with conditional Gamma heterogeneity can be easily estimated.

8. Estimating the Fixed Effects

It is well known that, except in special cases (linear and Poisson), treating the c_i as parameters to estimate leads to inconsistent estimates of the common parameters $\boldsymbol{\theta}$. But two questions arise. First, are there ways to adjust the "fixed effects" estimate of $\boldsymbol{\theta}$ to at least partially remove the bias? Second, could it be that estimates of the average partial effects, based generally on

$$N^{-1}\sum_{i=1}^{N}\frac{\partial m_t(\mathbf{x}_t,\hat{\boldsymbol{\theta}},\hat{\mathbf{c}}_i)}{\partial x_{tj}}, \tag{8.1}$$

where $m_t(\mathbf{x}_t,\boldsymbol{\theta},\mathbf{c}) = E(y_t|\mathbf{x}_t,\mathbf{c})$, are better behaved than the parameter estimates, and can their bias be removed? In the unobserved effects probit model, (8.1) becomes

$$N^{-1}\sum_{i=1}^{N}\hat{\beta}_j\phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}+\hat{c}_i), \tag{8.2}$$

which is easy to compute once $\hat{\boldsymbol{\beta}}$ and the $\hat{c}_i$ (N of them) have been obtained.

Hahn and Newey (2004) propose both jackknife and analytical bias corrections and show that they work well for the probit case. Generally, the jackknife procedure to remove the bias in $\hat{\boldsymbol{\theta}}$ is simple but can be computationally intensive. The idea is this. The estimator based on T time periods has probability limit that can be written as

$$\boldsymbol{\theta}_T = \boldsymbol{\theta} + \mathbf{b}_1/T + \mathbf{b}_2/T^2 + O(T^{-3}) \tag{8.3}$$

for vectors $\mathbf{b}_1$ and $\mathbf{b}_2$. Now, let $\hat{\boldsymbol{\theta}}_{(t)}$ denote the estimator that drops time period t. Then, assuming stability across t, the plim of $\hat{\boldsymbol{\theta}}_{(t)}$ is

$$\boldsymbol{\theta}_{(t)} = \boldsymbol{\theta} + \mathbf{b}_1/(T-1) + \mathbf{b}_2/(T-1)^2 + O(T^{-3}). \tag{8.4}$$

It follows that

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\,(T\hat{\boldsymbol{\theta}} - (T-1)\hat{\boldsymbol{\theta}}_{(t)}) &= (T\boldsymbol{\theta} + \mathbf{b}_1 + \mathbf{b}_2/T) - [(T-1)\boldsymbol{\theta} + \mathbf{b}_1 + \mathbf{b}_2/(T-1)] + O(T^{-3}) \\ &= \boldsymbol{\theta} - \mathbf{b}_2/[T(T-1)] + O(T^{-3}) = \boldsymbol{\theta} + O(T^{-2}).\end{aligned} \tag{8.5}$$

If, for given heterogeneity c_i, the data are independent and identically distributed, then (8.5) holds for all leave-one-time-period-out estimators, so we use the average of all such estimators in computing the panel jackknife estimator:

$$\tilde{\boldsymbol{\theta}} = T\hat{\boldsymbol{\theta}} - (T-1)T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\theta}}_{(t)}. \tag{8.6}$$

From the argument above, theasymptotic bias of $\tilde{\boldsymbol{\theta}}$ is on the order of T^{-2}.

Unfortunately, there are some practical limitations to the jackknife procedure, as well as to the analytical corrections derived by Hahn and Newey. First, aggregate time effects are not allowed, and they would be very difficult to include because the analysis is with $T \to \infty$. (In other words, they would introduce an incidental parameters problem in the time dimension as well as cross section dimension.) Generally, heterogeneity in the distributions across t changes the bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ when a time period is dropped, and so the simple transformation in (8.5) does not remove the bias terms. Second, Hahn and Newey assume independence across t conditional on c_i. It is a traditional assumption, but in static models it is often violated, and it must be violated in dynamic models. Plus, as noted by Hahn and Keursteiner, applying the "leave-one-out" method to dynamic models is problematical because the $\mathbf{b}_1$ and $\mathbf{b}_2$ in (8.4) would depend on t so, again, the transformation in (8.5) will not eliminate the $\mathbf{b}_1$ term.

Recently, Dhaene, Jochmans, and Thuysbaert (2006) propose a modification of the Hahn-Newey procedure that appears promising for dynamic models. In the simplest case, in addition to the "fixed effects" estimator using all time periods, they obtain estimators for two subperiods: one uses the earlier time periods, one uses later time periods, and they have some overlap (which is small as T gets large). Unfortunately, the procedure still requires stationarity and rules out aggregate time effects.

For the probit model, Fernández-Val (2007) studies the properties of estimators and average partial effects and allows time series dependence in the strictly exogenous regressors. Interestingly, in the probit model with exogenous regressors under the conditional independence assumption, the estimates of the APEs based on the "fixed" effects estimator has bias of order T^{-2} in the case that there is no heterogeneity. Unfortunately, these findings do not carry over to models with lagged dependent variables, and the bias corrections in that case are difficult to implement (and still do not allow for time heterogeneity).

The correlated random effects estimators restrict $D(c_i|\mathbf{x}_i)$ in some way, although the recent work by Altonji and Matzkin (2005) shows how those restrictions can be made reasonable. The approach generally identifies the APEs, and even the local average effects, and does not rule out aggregate time effects or arbitrary conditional serial dependence.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 5, Monday, July 30th, 4.30-5.30pm

Instrumental Variables with Treatment Effect Heterogeneity: Local Average Treatment Effects

1. Introduction

Here we investigate the interpretation of instrumental variables estimators allowing for general heterogeneity in the effect of the endogenous regressor. We shall see that instrumental variables estimators generally estimate average treatment effects, with the specific average depending on the choice of instruments. Initially we focus on the case where the endogenous regressor is binary. The example we will use is based on work by Joshua Angrist on estimating the effect of veteran status on earnings (Angrist, 1990). We also discuss the case where the endogenous variable takes on multiple values.

The general theme of this lecture is that with heterogenous treatment effects, endogeneity creates severe problems for identification of population averages. Population average causal effects are only estimable under very strong assumptions on the effect of the instrument on the endogenous regressor ("identification at infinity", or under the constant treatment effect assumptions). Without such assumptions we can only identify average effects for subpopulations that are induced by the instrument to change the value of the endogenous regressors. We refer to such subpopulations as *compliers*, and to the average treatment effect that is point identifed as the *local average treatment effect*. This terminology stems from the canonical example of a randomized experiment with noncompliance. In this example a random subpopulation is assigned to the treatment, but some of the individuals do not comply with their assigned treatment.

These complier subpopulations are not necessarily the subpopulations that are *ex ante* the most interesting subpopulations, but the data is in general not informative about average effects for other subpopulations without extrapolation, similar to the way in which a randomized experiment conducted on men is not informative about average effects for

women without extrapolation. The set up here allows the researcher to sharply separate the extrapolation to the (sub-)population of interest from exploration of the information in the data. The latter relies primarily on relatively interpretable, and substantively meaningful assumptions and avoids functional form or distributional assumptions. Given estimates for the compliers, one can then use the data to assess the plausibility of extrapolating the local average treatment effect to other subpopulations, using the information on outcomes given one of the two treatment levels and covariates.

With multiple instruments and or with covariates one can assess the evidence for heterogeneity, and the plausibility of extrapolation to the full population more extensively.

2. Linear Instrumental Variables with Constant Coefficients

First let us briefly review standard linear instrumental variables methods. In the example we are interested in the causal effect of military service on earnings. Let Y_i be the outcome of interest for unit i, W_i the endogenous regressor, and Z_i the instrument. The standard set up is as follows. A linear model is postulated for the relation between the outcome and the endogenous regressor:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \varepsilon_i.$$

This is a structural/behavioral/causal relationshiup. There is concern that the regressor W_i is endogenous, that is, that W_i is correlated with ε_i. Suppose that we are confident that a second variable, the instrument Z_i is both uncorrelated with the unobserved component ε_i and correlated with the endogenous regressor W_i. The solution is to use Z_i as an instrument for W_i. There are a couple of ways to implement this.

In Two-Stage-Least-Squares we first estimate a linear regression of the endogenous regressor on the instrument by least squares. Let the estimated regression function be

$$\hat{W}_i = \hat{\pi}_0 + \hat{\pi}_1 \cdot Z_i.$$

Then we regress the outcome on the predicted value of the endogenousr regressor, using least

squares:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \hat{W}_i.$$

Alternatively, with a single instrument we can estimate the two reduced form regressions

$$Y_i = \gamma_0 + \gamma_1 \cdot Z_i + \eta_i, \qquad \text{and} \quad W_i = \pi_0 + \pi_1 \cdot Z_i + \nu_i,$$

by least squares and estimate β_1 through Indirect Least Squares (ILS) as the ratio

$$\hat{\beta}_1^{\rm IV} = \hat{\gamma}_1 / \hat{\pi}_1.$$

If there is a single instrument and single endogenous regressor, we end up in both cases with the ratio of the sample covariance of Y and Z to the sample covariance of W and Z.

$$\hat{\beta}_1^{\rm IV} = \frac{\frac{1}{N}\sum_{i=1}^N (Y_i - \bar{Y}) \cdot (Z_i - \bar{Z})}{\frac{1}{N}\sum_{i=1}^N (W_i - \bar{W}) \cdot (Z_i - \bar{Z})}.$$

Using a central limit theorem for all the moments and the delta method we can infer the large sample distribution without additional assumptions.

3. Potential Outcomes

First we set up the problem in a slightly different way, using potential outcomes. Let $Y_i(0)$ and $Y_i(1)$ be two potential outcomes for unit i, one for each value of the endogenous regressor or treatment. The first potential outcome $Y_i(0)$ gives the outcome if person i were not to serve in the military, irrespective of whether this person served or not. The second gives the potential outcome given military service, again irrespective of whether the person served or not. We are interested in the causal effect of military service, $Y_i(1) - Y_i(0)$. We cannot directly observe this since we can only observe either $Y_i(0)$ or $Y_i(1)$, but not both. Let W_i be the realized value of the endogenous regressor, equal to zero or one. We observe W_i and

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 1 \\ Y_i(0) & \text{if } W_i = 0. \end{cases}$$

Now we introduce the instrumental variable set up by defining similar potential outcomes for the treatment. We focus on the case with a binary instrument Z_i. In the Angrist example, Z_i is a binary indicator for having a low draft number, and thus for being draft eligible. Define two potential outcomes $W_i(0)$ and $W_i(1)$, representing the value of the endogenous regressor given the two values for the instrument. The actual or realized value of the endogenous variable is

$$W_i = Y_i(Z_i) = \begin{cases} W_i(1) & \text{if } Z_i = 1 \\ W_i(0) & \text{if } Z_i = 0. \end{cases}$$

So we observe the triple Z_i, $W_i = W_i(Z_i)$ and $Y_i = Y_i(W_i(Z_i))$.

4. Local Average Treatment Effects

4.1. Assumptions

The key instrumental variables assumption is

Assumption 1 (Independence)

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), W_i(0), W_i(1)).$$

It requires that the instrument is as good as randomly assigned, and that it does not directly affect the outcome. The assumption is formulated in a nonparametric way, without definitions of residuals that are tied to functional forms.

It is important to note that this assumption is *not* implied by random assignment of Z_i. To see this, an alternative formulation of the assumption, generalizing the notation slightly, is useful. First we postulate the existence of four potential outcomes, $Y_i(z, w)$, corresponding to the outcome that would be observed if the instrument was $Z_i = z$ and the treatment was $W_i = w$. Then the independence assumption is the combination of two assumptions,

Assumption 2 (Random Assignment)

$$Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1)).$$

and

Assumption 3 (Exclusion Restriction)

$$Y_i(z,w) = Y_i(z',w), \qquad \text{for all } z, z', w.$$

The first of these two assumptions is implied by random assignment of Z_i, but the second is substantive, and randomization has no bearing on it.

It is useful for our approach to think about the compliance behavior of the different units, that is how they respond to different values of the instrument in terms of the treatment received. Table 1 gives the four possible pairs of values $(W_i(0), W_i(1))$, given the binary nature of the treatment and instrument: We cannot directly establish the type of a unit based

Table 1: Compliance Types

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker	defier
	1	complier	always-taker

on what we observe for them since we only see the pair (Z_i, W_i), not the pair $(W_i(0), W_i(1))$. Nevertheless, we can rule out some possibilities. Table 2 summarizes the information about compliance behavior from observed treatment status and instrument.

To make additional progress we we consider a *monotonicity* assumption, also known as the *no-defiers* assumption:

Assumption 4 (Monotonicity/No-Defiers)

$$W_i(1) \geq W_i(0).$$

Table 2: Compliance Type by Treatment and Instrument

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker/defier
	1	always-taker/defier	complier/always-taker

This assumption makes sense in a lot of applications. It is implied directly by many (constant coefficient) latent index models of the type:

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \varepsilon_i > 0\},$$

but it is much weaker than that. For example, one can allow for π_1 to vary across the population, as long as it is the same sign for all units. In the canonical non-compliance example this assumption is very plausible: if Z_i is assignment to a treatment, and W_i is an indicator for receipt of treatment, it makes sense that there are few, if any, individuals who always to the exact opposite of what their assignment is.

4.2. The Local Average Treatment Effect

Given this monotonicity assumption the information we can extract from observed compliance behavior increases.

Table 3: Compliance Type by Treatment and Instrument given Monotonicity

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

Let π_c, π_n, and π_a be the population proportions of compliers, never-takers and always-takers respectively. We can estimate those from the population distribution of treatment and instrument status:

$$\mathbb{E}[W_i|Z_i=0]=\pi_a, \qquad \mathbb{E}[W_i|Z_i=1]=\pi_a+\pi_c,$$

which we can invert to infer the population shares of the different types:

$$\pi_a=\mathbb{E}[W_i|Z_i=0], \qquad \pi_c=\mathbb{E}[W_i|Z_i=1]-\mathbb{E}[W_i|Z_i=0],$$

and

$$\pi_n=1-\mathbb{E}[W_i|Z_i=1].$$

Now consider average outcomes by instrument and treatment status:

$$\mathbb{E}[Y_i|W_i=0,Z_i=0]=\frac{\pi_c}{\pi_c+\pi_n}\cdot\mathbb{E}[Y_i(0)|\text{complier}]+\frac{\pi_n}{\pi_c+\pi_n}\cdot\mathbb{E}[Y_i(0)|\text{never}-\text{taker}],$$

$$\mathbb{E}[Y_i|W_i=0,Z_i=1]=\mathbb{E}[Y_i(0)|\text{never}-\text{taker}],$$

$$\mathbb{E}[Y_i|W_i=1,Z_i=0]=\mathbb{E}[Y_i(1)|\text{always}-\text{taker}],$$

and

$$\mathbb{E}[Y_i|W_i=1,Z_i=1]=\frac{\pi_c}{\pi_c+\pi_a}\cdot\mathbb{E}[Y_i(1)|\text{complier}]+\frac{\pi_a}{\pi_c+\pi_a}\cdot\mathbb{E}[Y_i(1)|\text{always}-\text{taker}].$$

From these relationships we can infer the average outcome by treatment status for compliers,

$$\mathbb{E}[Y_i(0)|\text{complier}], \qquad \text{and} \quad \mathbb{E}[Y_i(1)|\text{complier}],$$

and thus the average effect for compliers:

$$\mathbb{E}[Y(1)-Y_i(0)|\text{complier}]=\mathbb{E}[Y_i(1)|\text{complier}]-\mathbb{E}[Y_i(0)|\text{complier}].$$

We can also get there another way. Consider the least squares regression of Y on a constant and Z. The slope coefficient in that regression estimates

$$\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0].$$

Consider the first term:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i=1] &= \mathbb{E}[Y_i|Z_i=1,\text{complier}]\cdot\Pr(\text{complier}|Z_i=1)\\ &\quad+\mathbb{E}[Y_i|Z_i=1,\text{never}-\text{taker}]\cdot\Pr(\text{never}-\text{taker}|Z_i=1)\\ &\quad+\mathbb{E}[Y_i|Z_i=1,\text{always}-\text{taker}]\cdot\Pr(\text{always}-\text{taker}|Z_i=1)\\ &= \mathbb{E}[Y_i(1)|\text{complier}]\cdot\pi_c\\ &\quad+\mathbb{E}[Y_i(0)|\text{never}-\text{taker}]\cdot\pi_0+\mathbb{E}[Y_i(1)|\text{always}-\text{taker}]\cdot\pi_a.\end{aligned}$$

Similarly

$$\begin{aligned}\mathbb{E}[Y_i|Z_i=0] &= \mathbb{E}[Y_i|Z_i=0,\text{complier}]\cdot\Pr(\text{complier}|Z_i=0)\\ &\quad+\mathbb{E}[Y_i|Z_i=0,\text{never}-\text{taker}]\cdot\Pr(\text{never}-\text{taker}|Z_i=0)\\ &\quad+\mathbb{E}[Y_i|Z_i=0,\text{always}-\text{taker}]\cdot\Pr(\text{always}-\text{taker}|Z_i=0)\\ &= \mathbb{E}[Y_i(0)|\text{complier}]\cdot\pi_c\\ &\quad+\mathbb{E}[Y_i(0)|\text{never}-\text{taker}]\cdot\pi_0+\mathbb{E}[Y_i(1)|\text{always}-\text{taker}]\cdot\pi_a.\end{aligned}$$

Hence the difference is

$$\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]=\mathbb{E}[Y_i(1)-Y_i(0)|\text{complier}]\cdot\pi_c.$$

The same argument can be used to show that the slope coefficient in the regression of W on Z is

$$\mathbb{E}[W_i|Z_i=1]-\mathbb{E}[W_i|Z_i=0]=\pi_c.$$

Hence the instrumental variables estimand, the ratio of these two reduced form estimands, is equal to the local average treatment effect

$$\beta^{\rm IV}=\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[W_i|Z_i=1]-\mathbb{E}[W_i|Z_i=0]}=\mathbb{E}[Y_i(1)-Y_i(0)|\text{complier}].$$

The key insight is that the data are informative solely about the average effect for compliers only. Put differently, the data are not informative about the average effect for nevertakers because they are never seen receiving the treatment, and they are not informative about the average effect for alwaystakers because they are never seen without the treatment. A similar insight in a parametric setting is discussed in Björklund and Moffitt (1987).

A special case of considerable interest is that with one-side non-compliance. Suppose that $W_i(0)=0$, so that those assigned to the control group cannot receive the active treatment (but those assigned to the active treatment can decline to take it). In that case only two compliance types remain, compliers and always-takers. Monotonicity is automatically satisfied. The average effect for compliers is now equal to the average effect for the treated, since any one receiving the treatment is by definition a complier. This case was first studied in Bloom (1984).

4.3 Extrapolating to the Full Population

Although we cannot consistently estimate the average effect of the treatment for always-takers and never-takers, we do have some information about the outcomes for these subpopulations given one of the two treatment arms. Specifically, we can estimate

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right].$$

We can learn from these averages whether there is any evidence of heterogeneity in outcomes by compliance status, by comparing the pair of average outcomes of $Y_i(0)$;

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(0)|\text{complier}\right],$$

and the pair of average outcomes of $Y_i(1)$:

$$\mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{complier}\right].$$

If compliers, never-takers and always-takers are found to be substantially different in levels, by evidence of substantial difference between $\mathbb{E}[Y_i(0)|\text{never}-\text{taker}]$ and $\mathbb{E}[Y_i(0)|\text{complier}]$, and or/between $\mathbb{E}[Y_i(1)|\text{always}-\text{taker}]$, and $\mathbb{E}[Y_i(1)|\text{complier}]$, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types. On the other hand, if one finds that outcomes given the control treatment for never-takers and compliers are similar, and outcomes given the treatment are similar for compliers and always-takers, it is more plausible that average treatment effects for these groups are also comparable.

4.4 Covariates

The local average treatment effect result implies in general that one cannot consistently estimate average effects for subpopulations other than compliers. This still holds in cases where we observe covariates. One can incorporate the covariates into the analysis in a number of different ways. Traditionally the TSLS set up is used with the covariates entering in the outcome equation linearly and additively, as

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2' X_i + \varepsilon_i,$$

with the covariates added to the set of instruments. Given the potential outcome set up with general heterogeneity in the effects of the treatment, one may also wish to allow for more heterogeneity in the correlations with the covariates. Here we describe a general way of doing so. Unlike TSLS type approaches, this involves modelling both the dependence of

the outcome and the treatment on the covariates. Although there is often a reluctance to model the relation between the treatment, there appears no particular reason that economic theory is more informative about the relation between covariates and outcomes than about the relation between covariates and the choices that lead to the treatment.

A full model can be decomposed into two parts, a model for the compliance type given covariates, and a model for the potential outcomes given covariates for each compliance type. A traditional parametric model with a dummy endogenous variables might have the form (translated to the potential outcome set up used here):

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \pi_2' X_i + \eta_i \geq 0\},$$

$$Y_i(w) = \beta_0 + \beta_1 \cdot w + \beta_2' X_i + \varepsilon_i,$$

with (η_i, ε_i) jointly normally distributed (e.g., Heckman, 1978). Such a model can be viewed as imposing various restrictions on the relation between compliance types, covariates and outcomes. For example, in this model, if $\pi_1 > 0$, compliance type depends on η_i:

$$\text{unit } i \text{ is a } \begin{cases} \text{never}-\text{taker} & \text{if } \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{complier} & \text{if } -\pi_0 - \pi_1 - \pi_2' X_i \leq \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{always}-\text{taker} & \text{if } -\pi_0 - \pi_2' X_i \leq \eta_i, \end{cases}$$

which imposes strong restrictions on the relationship between type and outcomes.

An alternative approach is to model the potential outcome $Y_i(w)$ for units with compliance type t given covariates X_i through a common functional form with type and treatment specific parameters:

$$f_{Y(w)|X,T}(y(w)|x,t) = f(y|x;\theta_{wt}),$$

for $(w,t) = (0,n),(0,c),(1,c),(1,a)$. A natural model for the distribution of type is a trinomial logit model:

$$\Pr(T_i = \text{complier}|X_i) = \frac{1}{1 + \exp(\pi_n' X_i) + \exp(\pi_a' X_i)},$$

$$\Pr(T_i = \text{never} - \text{taker}|X_i) = \frac{\exp(\pi_n' X_i)}{1 + \exp(\pi_n' X_i) + \exp(\pi_a' X_i)},$$

and

$$\Pr(T_i = \text{always} - \text{taker}|X_i) = 1 - \Pr(T_i = \text{complier}|X_i) - \Pr(T_i = \text{never} - \text{taker}|X_i).$$

The log likelihood function is then, factored in terms of the contribution by observed W_i, Z_i) values:

$$\begin{aligned}
&\mathcal{L}(\pi_n, \pi_a, \theta_{0n}, \theta_{0c}, \theta_{1c}, \theta_{1a}) = \\
&\quad \times \prod_{i|W_i=0,Z_i=1} \frac{\exp(\pi_n' X_i)}{1 + \exp(\pi_n' X_i) + \exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{0n}) \\
&\quad \times \prod_{i|W_i=0,Z_i=0} \left(\frac{\exp(\pi_n' X_i)}{1 + \exp(\pi_n' X_i)} \cdot f(Y_i|X_i;\theta_{0n}) + \frac{1}{1 + \exp(\pi_n' X_i)} \cdot f(Y_i|X_i;\theta_{0c}) \right) \\
&\quad \times \prod_{i|W_i=1,Z_i=1} \left(\frac{\exp(\pi_a' X_i)}{1 + \exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{1a}) + \frac{1}{1 + \exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{1c}) \right) \\
&\quad \times \prod_{i|W_i=1,Z_i=0} \frac{\exp(\pi_a' X_i)}{1 + \exp(\pi_n' X_i) + \exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{1a}).
\end{aligned}$$

For example, the second factor consists of the contributions of individuals with $Z_i = 0$, $W_i = 0$, who are known to be either compliers or never-takers. Maximizing this is straightforward using the EM algorithm (Dempster, Laird, and Rubin, 1977). For an empirical example of this approach see Hirano, Imbens, Rubin and Zhou (2000), and Imbens and Rubin (1997).

In small samples one may wish to incorporate restrictions on the effects of the covariates, and for example assume that the effect of covariates on the outcome is the same irrespective of compliance type. An advantage of this approach is that it can easily be generalized. The type probabilities are nonparametricaly identified as functions of the covariates, and the similarly the outcome distributions by type as a function of the covariates.

5. Effects of Military Service on Earnings

Angrist (1989) was interested in estimating the effect of serving in the military on earnings. Angrist was concerned about the possibility that those choosing to serve in the military are different from those who do not in ways that affects their subsequent earnings irrespective of serving in the military. To avoid biases in simple comparisons of veterans and non-veterans, he exploited the Vietnam era draft lottery. Specifically he uses the binary indicator whether or not your draft lottery number made you eligible to be drafted as an instrument. This was tied to an individual's day of birth, so more or less random. Even so, that does not make it valid as an instrument. As the outcome of interest Angrist uses log earnings.

The simple ols regression leads to:

$$\begin{array}{rcccc} \widehat{\log(\text{earnings})}_i & = & 5.4364 & - & 0.0205 \cdot \widehat{\text{veteran}}_i \\ & & (0079) & & (0.0167) \end{array}$$

In Table 4 we present population sizes of the four treatmen/instrument samples. For example, with a low lottery number 5,948 individuals do not, and 1,372 individuals do serve in the military.

Table 4: Treatment Status by Assignment

		Z_i	
		0	1
W_i	0	5,948	1,915
	1	1,372	865

Using these data we get the following proportions of the various compliance types, given in Table 5, under the non-defiers assumption. For example, the proportion of nevertakers is

estimated as the conditional probability of $W_i = 0$ given $Z_i = 1$:

$$\Pr(\text{nevertaker}) = \frac{1915}{1915 + 865}.$$

Table 5: Compliance Types: Estimated Proportions

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker (0.6888)	defier (0)
	1	complier (0.1237)	always-taker (0.3112)

Table 6 gives the average outcomes for the four groups, by treatment and instrument status.

Table 6: Estimated Average Outcomes by Treatment and Instrument

		Z_i = 0	Z_i = 1
W_i	0	$\widehat{\mathbb{E}[Y]} = 5.4472$	$\widehat{\mathbb{E}[Y]} = 5.4028$
	1	$\widehat{\mathbb{E}[Y]} = 5.4076$,	$\widehat{\mathbb{E}[Y]} = 5.4289$

Table 7 gives the estimated averages for the four compliance types, under the exclusion restriction. This restriction is the key assumption here. There are a number of reasons why it may be violated, e.g., never-takers taking active actions to avoid military service if draft eligible. The local average treatment effect is -0.2336, a 23% drop in earnings as a result of serving in the military.

Simply doing IV or TSLS would give you the same numerical results:

$$\widehat{\log(\text{earnings})}_i = 5.4836 - 0.2336 \cdot \widehat{\text{veteran}}_i$$

Table 7: Compliance Types: Estimated Average Outcomes

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker: $\widehat{\mathbb{E}[Y_i(0)]} = 5.4028$	defier (NA)
	1	complier: $\widehat{\mathbb{E}[Y_i(0)]} = 5.6948$, $\widehat{\mathbb{E}[Y_i(1)]} = 5.4612$	always-taker: $\widehat{\mathbb{E}[Y_i(1)]} = 5.4076$
		(0.0289)	(0.1266)

It is interesting in this application to inspect the average outcome for different compliance groups. Average log earnings for never-takers are 5.40, lower by 29% than average earnings for compliers who do not serve in the military. This suggests that never-takers are substantially different than compliers, and that the average effect of 23% for compliers need not be informative never-takers. In contrast, average log earnings for always-takers are only 6% lower than those for compliers who serve, suggesting that the differences between always-takers and compliers are considerably smaller.

6. Multivalued Instruments

For any two values of the instrument z_0 and z_1 satisfying the local average treatment effect assumptions we can define the corresponding local average treatment effect:

$$\tau_{z_1,z_0} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i(z_1) = 1, W_i(z_0) = 0].$$

Note that these local average treatment effects need not be the same for different pairs of instrument values. Comparisons of estimates based on different instruments underlies tests of overidentifying restrictions in TSLS settings. An alternative interpretation of rejections in such testing procedures is therefore that the effects of interest vary, rather than that some of the instruments are invalid. Without assuming homogenous effects there are no tests in

general for the validity of the instruments.

The presence of multi-valued, or similarly, multiple, instruments, does, however, provide an opportunity to assess variation in treatment effects, as well as an opportunity to obtain average effects for subpopulations closer to the one of ultimate interest. Suppose that we have an instrument Z_i with support $z_0, z_1, \ldots, z_K$. Suppose also that the monotonicity assumption holds for all pairs z and z', and suppose that the instruments are ordered in such a way that

$$p(z_{k-1}) \leq p(z_k), \qquad \text{where} \quad p(z) = \mathbb{E}[W_i|Z_i = z].$$

Also suppose that the instrument is relevant,

$$\mathbb{E}[g(Z_i) \cdot W_i] \neq 0.$$

Then the instrumental variables estimator based on using $g(Z)$ as an instrument for W estimates a weighted average of local average treatment effects:

$$\tau_{g(\cdot)} = \frac{\text{Cov}(Y_i, g(Z_i))}{\text{Cov}(W_i, g(Z_i))} = \sum_{k=1}^{K} \lambda_k \cdot \tau_{z_k, z_{k-1}},$$

where

$$\lambda_k = \frac{(p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]}{\sum_{k=1}^{K} p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]},$$

$$\pi_k = \Pr(Z_i = z_k).$$

These weights are nonnegative and sum up to one.

By choosing $g(z)$ one can choose a different weight function, although there is obviously a limit to what one can do. One can only estimate a weighted average of the local average treatment effects defined for all pairs of instrument values in the support of the instrument.

If the instrument is continuous, and $p(z)$ is continuous in z, we can define the limit of the local average treatment effects

$$\tau_z = \lim_{z' \downarrow z_0, z'' \uparrow z_0} \tau_{z',z''}.$$

In this case with the monotonicity assumption hold for all pairs z and z', we can use the implied structure on the compliance behavior by modelling $W_i(z)$ as a threshold crossing process,

$$W_i(z) = 1\{h(z) + \eta_i \geq 0\},$$

with the scalar unobserved component η_i independent of the instrument Z_i. This type of latent index model is used extensively in work by Heckman (Heckman and Robb, 1985; Heckman,1990; Heckman and Vytlacil, 2005), as well as in Vytlacil (2000). Vytlacil shows that if the earlier three assumptions hold for all pairs z and z', than there is a function $h(\cdot)$ such that this latent index structure is consistent with the joint distribution of the observables. The latent index structure implies that individuals can be ranked in terms of an unobserved component η_i such that if for two individuals i and j we have $\eta_i > \eta_j$, than $W_i(z) \geq W_j(z)$ for all z.

Given this assumption, we can define the marginal treatment effect $\tau(\eta)$ as

$$\tau(\eta) = \mathbb{E}\left[Y_i(1) - Y_i(0) | \, \eta_i = \eta\right].$$

This marginal treatment effect relates directly to the limit of the local average treatment effects

$$\tau(\eta) = \tau_z, \qquad \text{with} \;\; \eta = -h(z)).$$

Note that we can only define this for values of η for which there is a z such that $\tau = -h(z)$. Normalizing the marginal distribution of η to be uniform on $[0, 1]$ (Vytlacil, 2002), this

restricts η to be in the interval $[\inf_z p(z), \sup_z p(z)]$, where $p(z) = \Pr(W_i = 1|Z_i = z)$. Heckman and Vytlacil (2005) characterize various average treatment effects in terms of this limit. For example, the average treatment effect is simply the average of the marginal treatment effect over the marginal distribution of η:

$$\tau = \int_\eta \tau(\eta) dF_\eta(\eta).$$

In practice the same limits remain on the identification of average effects. The population average effect is only identified if the instrument moves the probability of participation from zero to one. In fact identification of the population average treatment effect does not require identification of $\tau(\eta)$ at every value of η. The latter is sufficient, but not necessary. For example, in a randomized experiment (corresponding to a binary instrument with the treatment indicator equal to the instrument) the average treatment effect is obviously identified, but the marginal treatment effect is not for any value of η.

7. Multivalued Endogenous Variables

Now suppose that the endogenous variable W takes on values $0, 1, \ldots, J$. We still assume that the instrument Z is binary. We study the interpretation of the instrumental variables estimand

$$\tau = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(W_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]}.$$

We make the exclusion assumption that

$$Y_i(w)\ W_i(z)\ \perp\!\!\!\perp\ Z_i,$$

and a version of the monotonicity assumption,

$$W_i(1) \geq W_i(0),$$

Then we can write the instrumental variables estimand as

$$\tau = \sum_{j=1}^{J} \lambda_j \cdot \mathbb{E}[Y_i(j) - Y_i(j-1) | W_i(1) \geq j > W_i(0)],$$

where

$$\lambda_j = \frac{\Pr(W_i(1) \geq j > W_i(0)}{\sum_{i=1}^{J} \Pr(W_i(1) \geq i > W_i(0)}.$$

Note that we can estimate the weights λ_j because

$$\Pr(W_i(1) \geq j > W_i(0) = \Pr(W_i(1) \geq j) - \Pr(W_i(0) \geq j)$$

$$= \Pr(W_i(1) \geq j | Z_i = 1) - \Pr(W_i(0) \geq j | Z_i = 0)$$

$$= \Pr(W_i \geq j | Z_i = 1) - \Pr(W_i \geq j | Z_i = 0),$$

using the monotonicity assumption.

8. Instrumental Variables Estimates of the Returns to Education Using Quarter of Birth as an Instrument

Here we use a subset of the data used by Angrist and Krueger in their 1991 study of the returns to education. Angrist and Krueger were concerned with the endogeneity of education, worrying that individuals with higher ability would have had higher earnings given any level of education, as well as be more likely to have high levels of education. In that case simple least squares estimates would over estimate the returns to education. Their idea was that individuals born in different parts of the year are subject to slightly different compulsory schooling laws. If you are born before a fixed cutoff date you enter school at a younger age than if you are born after that cutoff date, and given that you are allowed to leave school when you turn sixteen, those individuals born before the cutoff date are required to completely more years of schooling. The instrument can therefore be thought of

as the tightness of the compulsory schooling laws, with the tightness being measured by the individual's quarter of birth.

Angrist and Krueger implement this using census data with quarter of birth indicators as the instrument. Table 1 gives average years of education and sample sizes by quarter of birth.

Table 8: Average Level of Education by Quarter of Birth

quarter	1	2	3	4
average level of education	12.69	12.74	12.81	12.84
standard error	0.01	0.01	0.01	0.01
number of observations	81,671	80,138	86,856	80,844

In the illustrations below we just use a single instrument, an indicator for being born in the first quarter. First let us look at the reduced form regressions of log earnings and years of education on the first quarter of birth dummy:

$$\begin{aligned} \widehat{\text{educ}_i} &= 12.797 - 0.109 \cdot \text{qob}_i \\ &\quad (0.006) \quad (0.013) \end{aligned}$$

and

$$\begin{aligned} \widehat{\log(\text{earnings})}_i &= 5.903 - 0.011 \cdot \text{qob}_i \\ &\quad (0.001) \quad (0.003) \end{aligned}$$

The instrumental variables estimate is the ratio of the reduced form coefficients,

$$\hat{\beta}^{\text{IV}} = \frac{-0.1019}{-0.011} = 0.1020.$$

Now let us interpret this in the context of heterogeneous returns to education. This estimate is an average of returns to education, consisting of two types of averaging. The first is over different levels of education. That is, it is a weighted average of the return to moving from nine to ten years, to moving from ten to eleven years, to moving from eleven to twelve years, etcetera. In addition, for any level, e.g., to moving from nine to ten years of education, it is an average effect where the averaging is over those people whose schooling would have been at least ten years of education if tigther compulsory schooling laws had been in effect for them, and who would have had less than ten years of education had they been subject to the looser compulsory schooling laws.

Furthermore, we can estimate how large a fraction of the population is in these categories. First we estimate the

$$\gamma_j = \Pr(W_i(1) \geq j > W_i(0) = \Pr(W_i \geq j | Z_i = 1) - \Pr(W_i \geq j | Z_i = 0)$$

as

$$\hat{\gamma}_j = \frac{1}{N_1} \sum_{i|Z_i=1} 1\{W_i \geq j\} - \frac{1}{N_0} \sum_{i|Z_i=0} 1\{W_i \geq j\}.$$

This gives the unnormalized weight function. We then normalize the weights so they add up to one, $\hat{\lambda}_j = \hat{\gamma}_j / \sum_i \hat{\gamma}_i$.

Figure 1-4 present some of the relevant evidence here. First, Figure 1 gives the distribution of years of education. Figure 2 gives the normalized and Figure 3 gives the unnormalized weight functions. Figure 4 gives the distribution functions of years of education by the two values of the instrument. The most striking feature of these figures (not entirely unanticipated) is that the proportion of individuals in the "complier" subpopulations is extremely small, never more than 2% of the population. This implies that these instrumental variables estimates are averaged only over a very small subpopulation, and that there is little reason to believe that they generalize to the general population. (Nevertheless, this may well be a very interesting subpopulation for some purposes.) The nature of the instrument also suggests

that most of the weight would be just around the number of years that would be required under the compulsory schooling laws. The weight function is actually much flatter, putting weight even on fourteen to fifteen years of education.

Figure 1: histogram estimate of density of years of education

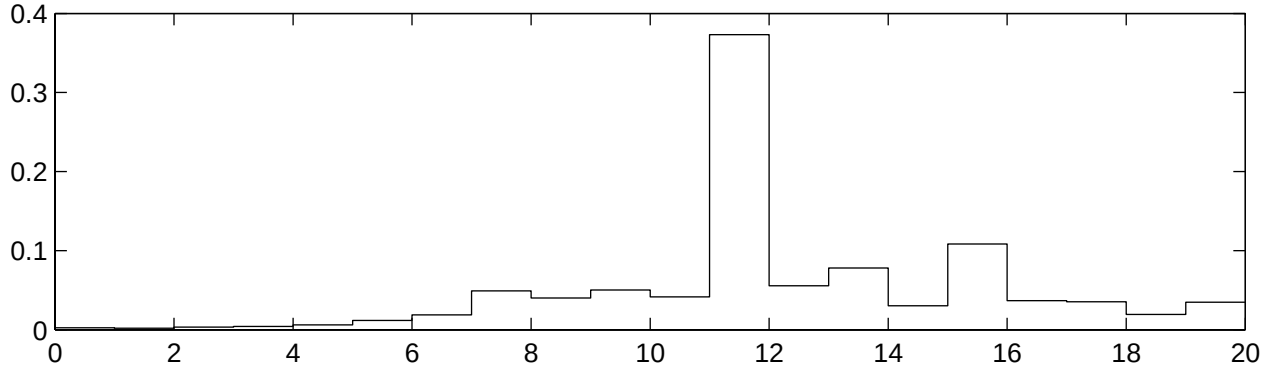


Figure 2: Normalized Weight Function for Instrumental Variables Estimand

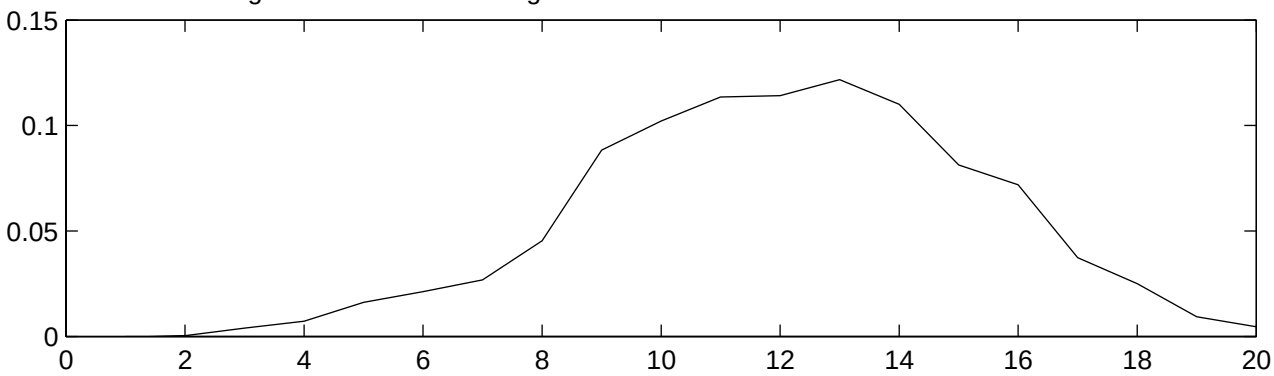


Figure 3: Unnormalized Weight Function for Instrumental Variables Estimand

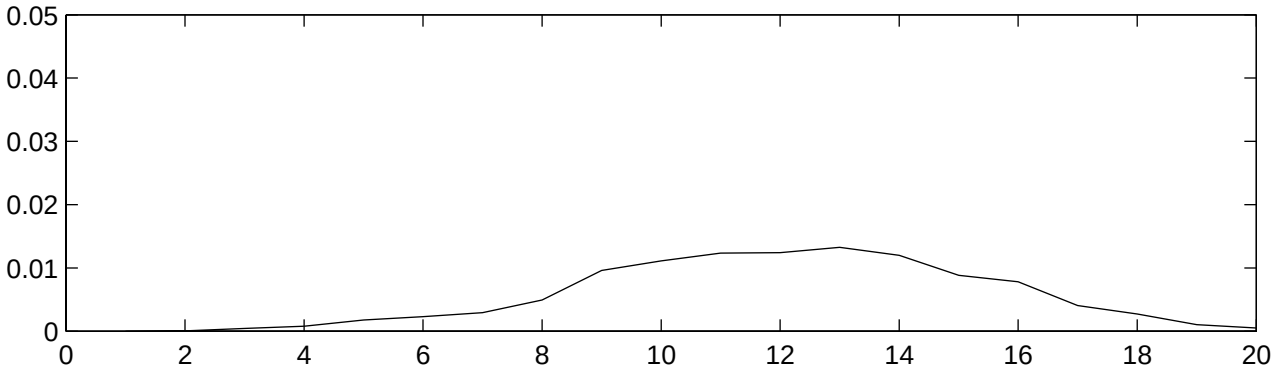
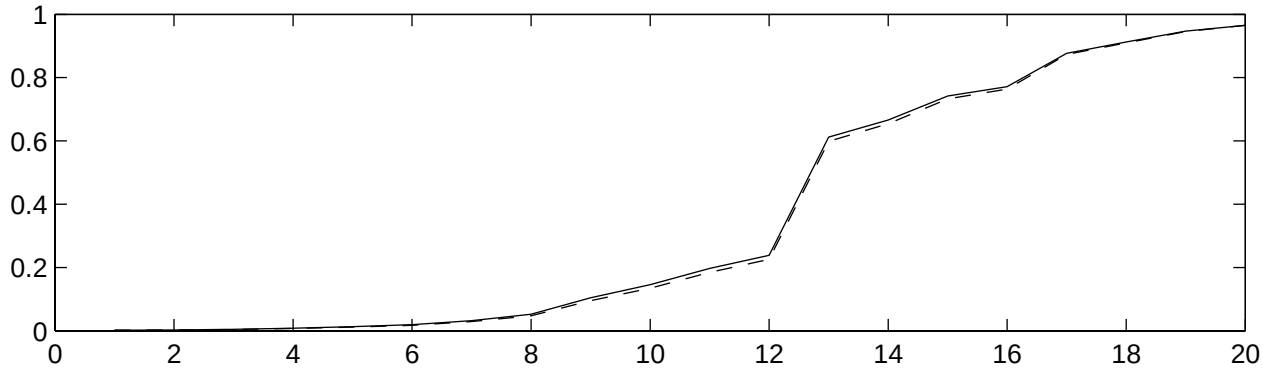


Figure 3: Education Distribution Function by Quarter



References

Angrist, J. D., and G. W. Imbens, (1995), "Two–Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, Vol 90, No. 430, 431-442.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," (with discussion) *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

Angrist, J. and A. Krueger, (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87, June.

Björklund, A. and R. Moffitt, (1987), "The Estimation of Wage Gains and Welfare Gains in Self–Selection Models", *Review of Economics and Statistics*, Vol. LXIX, 42–49.

Bloom, H., (1984), "Accounting for No–shows in Experimental Evaluation Designs," *Evaluation Review*, 8(2) 225–246.

Dempster, A., N. Laird, and D. Rubin (1977), "Maximum Likelihood Estimation from Incomplete Data Using the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Heckman, J. J. (1990), "Varieties of Selection Bias," *American Economic Review* 80, 313-318.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., and E. Vytlacil, (2005), "Structural Equations, Treatment Effects,

and Econometric Policy Evaluation," *Econometrica*, Vol. 73(3), 669-738.

Hirano, K., G. Imbens, D. Rubin, and X. Zhou (2000), "Identification and Estimation of Local Average Treatment Effects," *Biostatistics*, Vol. 1(1), 69-88.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G. W., and D. B. Rubin, (1997), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *Annals of Statistics*, Vol. 25, No. 1, 305–327.

Vytlacil, E., (2002), "Independence, Monotonicity, and Latent Index Models: Am Equivalence Result," *Econometrica*, Vol. 70(1), 331-341.

What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 6, **Tuesday**, **July 31st**, **9.00-10.30 am**

Control Function and Related Methods

These notes review the control function approach to handling endogeneity in models linear in parameters, and draws comparisons with standard methods such as 2SLS. Certain nonlinear models with endogenous explanatory variables are most easily estimated using the CF method, and the recent focus on average marginal effects suggests some simple, flexible strategies. Recent advances in semiparametric and nonparametric control function method are covered, and an example for how one can apply CF methods to nonlinear panel data models is provided.

1. Linear-in-Parameters Models: IV versus Control Functions

Most models that are linear in parameters are estimated using standard IV methods – either two stage least squares (2SLS) or generalized method of moments (GMM). An alternative, the control function (CF) approach, relies on the same kinds of identification conditions. In the standard case where a endogenous explanatory variables appear linearly, the CF approach leads to the usual 2SLS estimator. But there are differences for models nonlinear in endogenous variables even if they are linear in parameters. And, for models nonlinear in parameters, the CF approach offers some distinct advantages.

Let y_1 denote the response variable, y_2 the endogenous explanatory variable (a scalar for simplicity), and $\mathbf{z}$ the $1 \times L$ vector of exogenous variables (which includes unity as its first element). Consider the model

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{1.1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of $\mathbf{z}$ that also includes a constant. The sense in which $\mathbf{z}$ is exogenous is given by the L orthogonality (zero covariance) conditions

$$\mathrm{E}(\mathbf{z}'u_1) = \mathbf{0}. \tag{1.2}$$

Of course, this is the same exogeneity condition we use for consistency of the 2SLS estimator, and we can consistently estimate $\boldsymbol{\delta}_1$ and α_1 by 2SLS under (1.2) and the rank condition, Assumption 2SLS.2.

Just as with 2SLS, the reduced form of y_2 – that is, the linear projection of y_2 onto the exogenous variables – plays a critical role. Write the reduced form with an error term as

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{1.3}$$

$$\mathrm{E}(\mathbf{z}'v_2) = \mathbf{0} \tag{1.4}$$

where $\boldsymbol{\pi}_2$ is $L \times 1$. Endogeneity of y_2 arises if and only if u_1 is correlated with v_2. Write the

linear projection of u_1 on v_2, in error form, as

$$u_1 = \rho_1 v_2 + e_1, \tag{1.5}$$

where $\rho_1 = \mathrm{E}(v_2 u_1)/\mathrm{E}(v_2^2)$ is the population regression coefficient. By definition, $\mathrm{E}(v_2 e_1) = 0$, and $\mathrm{E}(\mathbf{z}' e_1) = \mathbf{0}$ because u_1 and v_2 are both uncorrelated with $\mathbf{z}$.

Plugging (1.5) into equation (1.1) gives

$$y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1, \tag{1.6}$$

where we now view v_2 as an explanatory variable in the equation. As just noted, e_1, is uncorrelated with v_2 and $\mathbf{z}$. Plus, y_2 is a linear function of $\mathbf{z}$ and v_2, and so e_1 is also uncorrelated with y_2.

Because e_1 is uncorrelated with $\mathbf{z}_1$, y_2, and v_2, (1.6) suggests a simple procedure for consistently estimating δ_1 and α_1 (as well as ρ_1): run the OLS regression of y_1 on $\mathbf{z}_1, y_2$, and v_2 using a random sample. (Remember, OLS consistently estimates the parameters in any equation where the error term is uncorrelated with the right hand side variables.) The only problem with this suggestion is that we do not observe v_2; it is the error in the reduced form equation for y_2. Nevertheless, we can write $v_2 = y_2 - \mathbf{z}\boldsymbol{\pi}_2$ and, because we collect data on y_2 and $\mathbf{z}$, we can consistently estimate $\boldsymbol{\pi}_2$ by OLS. Therefore, we can replace v_2 with $\hat{v}_2$, the OLS residuals from the first-stage regression of y_2 on $\mathbf{z}$. Simple substitution gives

$$y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error, \tag{1.7}$$

where, for each i, $error_i = e_{i1} + \rho_1 \mathbf{z}_i (\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2)$, which depends on the sampling error in $\hat{\boldsymbol{\pi}}_2$ unless $\rho_1 = 0$. Standard results on two-step estimation imply the OLS estimators from (1.7) will be consistent for δ_1, α_1, and ρ_1.

The OLS estimates from (1.7) are control function estimates. The inclusion of the residuals $\hat{v}_2$ "controls" for the endogeneity of y_2 in the original equation (although it does so with sampling error because $\hat{\boldsymbol{\pi}}_2 \neq \boldsymbol{\pi}_2$).

It is a simple exercise in the algebra of least squares to show that the OLS estimates of $\boldsymbol{\delta}_1$ and α_1 from (1.7) are *identical* to the 2SLS estimates starting from (1.1) and using $\mathbf{z}$ as the vector of instruments. (Standard errors from (1.7) must adjust for the generated regressor.)

It is trivial to use (1.7) to test $H_0 : \rho_1 = 0$, as the usual t statistic is asymptotically valid under homoskedasticity ($Var(u_1|\mathbf{z}, y_2) = \sigma_1^2$ under H_0); or use the heteroskedasticity-robust version (which does *not* account for the first-stage estimation of $\boldsymbol{\pi}_2$).

Now extend the model:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1 \tag{1.8}$$

$$\mathrm{E}(u_1|\mathbf{z}) = 0. \tag{1.9}$$

For simplicity, assume that we have a scalar, z_2, that is not also in $\mathbf{z}_1$. Then, under (1.9) – which is stronger than (1.2), and is essentially needed to identify nonlinear models – we can use, say, z_2^2 (if z_2 is not binary) as an instrument for y_2^2 because any function of z_2 is uncorrelated with u_1. In other words, we can apply the standard IV estimator with explanatory variables $(\mathbf{z}_1, y_2, y_2^2)$ and instruments $(\mathbf{z}_1, z_2, z_2^2)$; note that we have two endogenous explanatory variables, y_2 and y_2^2.

What would the CF approach entail in this case? To implement the CF approach in (1.8), we obtain the conditional expectation $\mathrm{E}(y_1|\mathbf{z}, y_2)$ – a linear projection argument no longer works because of the nonlinearity – and that requires an assumption about $\mathrm{E}(u_1|\mathbf{z}, y_2)$. A standard assumption is

$$\mathrm{E}(u_1|\mathbf{z}, y_2) = \mathrm{E}(u_1|\mathbf{z}, v_2) = \mathrm{E}(u_1|v_2) = \rho_1 v_2, \tag{1.10}$$

where the first equality follows because y_2 and v_2 are one-to-one functions of each other (given $\mathbf{z}$) and the second would hold if (u_1, v_2) is independent of $\mathbf{z}$ – a nontrivial restriction on the reduced form error in (1.3), not to mention the structural error u_1.. The final assumption is linearity of the conditional expectation $\mathrm{E}(u_1|v_2)$, which is more restrictive than simply defining a linear projection. Under (1.10),

$$\begin{aligned} \mathrm{E}(y_1|\mathbf{z}, y_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1(y_2 - \mathbf{z}\boldsymbol{\pi}_2) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1 v_2. \end{aligned} \tag{1.11}$$

Implementing the CF approach means running the OLS regression y_1 on $\mathbf{z}_1, y_2, y_2^2, \hat{v}_2$,where $\hat{v}_2$ still represents the reduced form residuals. The CF estimates are *not* the same as the 2SLS estimates using any choice of instruments for (y_2, y_2^2).

The CF approach, while likely more efficient than a direct IV approach, is less robust. For example, it is easily seen that (1.9) and (1.10) imply that $\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$. A linear conditional expectation for y_2 is a substantive restriction on the conditional distribution of y_2. Therefore, the CF estimator will be inconsistent in cases where the 2SLS estimator will be consistent. On the other hand, because the CF estimator solves the endogeneity of y_2 and y_2^2 by adding the scalar $\hat{v}_2$ to the regression, it will generally be more precise – perhaps much more precise – than the IV estimator. (I do not know of a systematic analysis comparing the two approaches in models such as (1.8).)

Standard CF approaches impose extra assumptions even in the simple model (1.1) if we allow y_2 to have discreteness in its distribution. For example, suppose y_2 is a binary response. Then the CF approach involves estimating

$$E(y_1|\mathbf{z},y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|\mathbf{z},y_2),$$

and so we must be able to estimate $E(u_1|\mathbf{z},y_2)$. If $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$, (u_1, e_2) is independent of $\mathbf{z}$, $E(u_1|e_2) = \rho_1 e_2$, and $e_2 \sim$Normal$(0,1)$, then

$$\begin{aligned} E(u_1|\mathbf{z},y_2) &= E[E(u_1|\mathbf{z},e_2)|\mathbf{z},y_2] = \rho_1 E(v_2|\mathbf{z},y_2) \\ &= \rho_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \end{aligned}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio (IMR). A simple two-step estimator is to obtain the probit estimator $\hat{\boldsymbol{\delta}}_2$ and then to add the "generalized residual," $\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1-y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as a regressor:

$$y_{i1} \text{ on } \mathbf{z}_{i1},\ y_{i2},\ \widehat{gr}_{i2},\ i = 1,\ldots,N.$$

Consistency of the CF estimators hinges on the model for $D(y_2|\mathbf{z})$ being correctly specified, along with linearity in $E(u_1|v_2)$ (and some sort of independence with $\mathbf{z}$). Of course, if we just apply 2SLS directly to (1.1), it makes no distinction among discrete, continuous, or some mixture for y_2. 2SLS is consistent if $L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ actually depends on $\mathbf{z}_2$ and (1.2) holds. So, while estimating (1.1) using CF methods when y_2 is binary is somewhat popular (Stata's "treatreg" even has the option of full MLE, where (u_1, e_2) is bivariate normal), one should remember that it is less robust than standard IV approaches.

How might one use the binary nature of y_2 in IV estimation? Assume $E(u_1|\mathbf{z}) = 0$ and, nominally, assume a probit model for $D(y_2|\mathbf{z})$. Obtain the fitted probabilities, $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, from the first stage probit, and then use these as IVs for y_{i2}. This method is fully robust to misspecification of the probit model; the standard errors need not be adjusted for the first-stage probit (asymptotically); and it is the efficient IV estimator if $P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2)$ and $Var(u_1|\mathbf{z}) = \sigma_1^2$. But it is probably less efficient than the CF estimator if the additional assumptions needed for CF consistency hold. (Note: Using $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as an IV for y_{i2} is not the same as using $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as a regressor in place of y_{i2}.)

To summarize: except in the case where y_2 appears linearly and a linear reduced form is estimated for y_2, the CF approach imposes extra assumptions not imposed by IV approaches. However, in more complicated models, it is hard to beat the CF approach.

2. Correlated Random Coefficient Models

Control function methods can be used for random coefficient models – that is, models where unobserved heterogeneity interacts with endogenous explanatory variables. However, in some cases, standard IV methods are more robust. To illustrate, we modify equation (1.1) as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + u_1, \tag{2.1}$$

where $\mathbf{z}_1$ is $1 \times L_1$, y_2 is the endogenous explanatory variable, and a_1, the "coefficient" on y_2 – an unobserved random variable. [It is now convenient to set apart the intercept.] We could replace $\boldsymbol{\delta}_1$ with a random vector, say $\mathbf{d}_1$, and this would not affect our analysis of the IV estimator (but would slightly alter the control function estimator). Following Heckman and Vytlacil (1998), we refer to (2.1) as a **correlated random coefficient** (**CRC**) **model**.

It is convenient to write $a_1 = \alpha_1 + v_1$ where $\alpha_1 = \mathrm{E}(a_1)$ is the object of interest. We can rewrite the equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + v_1 y_2 + u_1 \equiv \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + e_1, \tag{2.2}$$

where $e_1 = v_1 y_2 + u_1$. Equation (2.2) shows explicitly a constant coefficient on y_2 (which we hope to estimate) but also an interaction between the observed heterogeneity, v_1, and y_2. Remember, (2.2) is a population model. For a random draw, we would write $y_{i1} = \eta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + v_{i1} y_{i2} + u_{i1}$, which makes it clear that $\boldsymbol{\delta}_1$ and α_1 are parameters to estimate and v_{i1} is specific to observation i.

As discussed in Wooldridge (1997, 2003), the potential problem with applying instrumental variables (2SLS) to (2.2) is that the error term $v_1 y_2 + u_1$ is not necessarily uncorrelated with the instruments $\mathbf{z}$, even if we make the assumptions

$$\mathrm{E}(u_1|\mathbf{z}) = \mathrm{E}(v_1|\mathbf{z}) = 0, \tag{2.3}$$

which we maintain from here on. Generally, the term $v_1 y_2$ can cause problems for IV estimation, but it is important to be clear about the nature of the problem. If we are allowing y_2 to be correlated with u_1 then we also want to allow y_2 and v_1 to be correlated. In other words, $\mathrm{E}(v_1 y_2) = \mathrm{Cov}(v_1, y_2) \equiv \tau_1 \neq 0$. But a nonzero unconditional covariance is *not* a problem with applying IV to (2.2): it simply implies that the composite error term, e_1, has (unconditional) mean τ_1 rather than a zero. As we know, a nonzero mean for e_1 means that the orginal intercept, η_1, would be inconsistenly estimated, but this is rarely a concern.

Therefore, we can allow $\mathrm{Cov}(v_1, y_2)$, the unconditional covariance, to be unrestricted. But the usual IV estimator is generally inconsistent if $\mathrm{E}(v_1 y_2|\mathbf{z})$ depends on $\mathbf{z}$. (There are still cases, which we will cover in Part IV, where the IV estimator is consistent.). Note that, because

$\mathrm{E}(v_1|\mathbf{z}) = 0$, $\mathrm{E}(v_1 y_2|\mathbf{z}) = \mathrm{Cov}(v_1, y_2|\mathbf{z})$. Therefore, as shown in Wooldridge (2003), a sufficient condition for the IV estimator applied to (2.2) to be consistent for $\boldsymbol{\delta}_1$ and α_1 is

$$\mathrm{Cov}(v_1, y_2|\mathbf{z}) = \mathrm{Cov}(v_1, y_2). \tag{2.4}$$

The 2SLS intercept estimator is consistent for $\eta_1 + \tau_1$. Condition (2.4) means that the conditional covariance between v_1 and y_2 is not a function of $\mathbf{z}$, but the unconditional covariance is unrestricted.

Because v_1 is unobserved, we cannot generally verify (2.4). But it is easy to find situations where it holds. For example, if we write

$$y_2 = m_2(\mathbf{z}) + v_2 \tag{2.5}$$

and assume (v_1, v_2) is independent of $\mathbf{z}$ (with zero mean), then (2.4) is easily seen to hold because $\mathrm{Cov}(v_1, y_2|\mathbf{z}) = \mathrm{Cov}(v_1, v_2|\mathbf{z})$, and the latter cannot be a function of $\mathbf{z}$ under independence. Of course, assuming v_2 in (2.5) is independent of $\mathbf{z}$ is a strong assumption even if we do not need to specify the mean function, $m_2(\mathbf{z})$. It is much stronger than just writing down a linear projection of y_2 on $\mathbf{z}$ (which is no real assumption at all). As we will see in various models in Part IV, the representation (2.5) with v_2 independent of $\mathbf{z}$ is not suitable for discrete y_2, and generally (2.4) is not a good assumption when y_2 has discrete characteristics. Further, as discussed in Card (2001), (2.4) can be violated even if y_2 is (roughly) continuous. Wooldridge (2005a) makes some headway in relaxing (2.44) by allowing for parametric heteroskedasticity in u_1 and v_2.

A useful extension of (1.1) is to allow observed exogenous variables to interact with y_2. The most convenient formulation is

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + (\mathbf{z}_1 - \boldsymbol{\psi}_1)y_2\boldsymbol{\gamma}_1 + v_1 y_2 + u_1 \tag{2.6}$$

where $\boldsymbol{\psi}_1 \equiv \mathrm{E}(\mathbf{z}_1)$ is the $1 \times L_1$ vector of population means of the exogenous variables and $\boldsymbol{\gamma}_1$ is an $L_1 \times 1$ parameter vector. As we saw in Chapter 4, subtracting the mean from $\mathbf{z}_1$ before forming the interaction with y_2 ensures that α_1 is the average partial effect.

Estimation of (2.6) is simple if we maintain (2.4) [along with (2.3) and the appropriate rank condition]. Typically, we would replace the unknown $\boldsymbol{\psi}_1$ with the sample averages, $\bar{\mathbf{z}}_1$, and then estimate

$$y_{i1} = \theta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}\boldsymbol{\gamma}_1 + error_i \tag{2.7}$$

by instrumental variables, ignoring the estimation error in the population mean. The only issue

is choice of instruments, which is complicated by the interaction term. One possibility is to use interactions between $\mathbf{z}_{i1}$ and all elements of $\mathbf{z}_i$ (including $\mathbf{z}_{i1}$). This results in many overidentifying restrictions, even if we just have one instrument z_{i2} for y_{i2}. Alternatively, we could obtain fitted values from a first stage linear regression y_{i2} on $\mathbf{z}_i$, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, and then use IVs $[1, \mathbf{z}_i, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}]$, which results in as many overidentifying restrictions as for the model without the interaction. Importantly, the use of $(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}$ as IVs for $(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}$ is asymptotically the same as using instruments $(\mathbf{z}_{i1} - \boldsymbol{\psi}_1) \cdot (\mathbf{z}_i\boldsymbol{\pi}_2)$, where $L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ is the linear projection. In other words, consistency of this IV procedure does not in any way restrict the nature of the distribution of y_2 given $\mathbf{z}$. Plus, although we have generated instruments, the assumptions sufficient for ignoring estimation of the instruments hold, and so inference is standard (perhaps made robust to heteroskedasticity, as usual).

We can just identify the parameters in (2.6) by using a further restricted set of instruments, $[1, \mathbf{z}_{i1}, \hat{y}_{i2}, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}]$. If so, it is important to use these as instruments and not as regressors. If we add the assumption. The latter procedure essentially requires a new assumption:

$$\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2 \tag{2.8}$$

(where $\mathbf{z}$ includes a constant). Under (2.3), (2.4), and (2.8), it is easy to show

$$\mathrm{E}(y_1|\mathbf{z}) = (\eta_1 + \tau_1) + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\pi}_2) + (\mathbf{z}_1 - \boldsymbol{\psi}_1) \cdot (\mathbf{z}\boldsymbol{\pi}_2)\boldsymbol{\gamma}_1, \tag{2.9}$$

which is the basis for the Heckman and Vytlacil (1998) plug-in estimator. The usual IV approach simply relaxes (2.8) and does not require adjustments to the standard errors (because it uses generated instruments, not generated regressors).

We can also use a control function approach if we assume

$$\mathrm{E}(u_1|\mathbf{z}, v_2) = \rho_1 v_2, \mathrm{E}(v_1|\mathbf{z}, v_2) = \xi_1 v_2. \tag{2.10}$$

Then

$$\mathrm{E}(y_1|\mathbf{z}, y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \xi_1 v_2 y_2 + \rho_1 v_2, \tag{2.11}$$

and this equation is estimable once we estimate $\boldsymbol{\pi}_2$. Garen's (1984) control function procedure is to first regress y_2 on $\mathbf{z}$ and obtain the reduced form residuals, $\hat{v}_2$, and then to run the OLS regression y_1 on $1, \mathbf{z}_1, y_2, \hat{v}_2 y_2, \hat{v}_2$. Under the maintained assumptions, Garen's method consistently estimates $\boldsymbol{\delta}_1$ and α_1. Because the second step uses generated regressors, the standard errors should be adjusted for the estimation of $\boldsymbol{\pi}_2$ in the first stage. Nevertheless, a test that y_2 is exogenous is easily obtained from the usual F test of $\mathrm{H}_0 : \xi_1 = 0, \rho_1 = 0$ (or a heteroskedasticity-robust version). Under the null, no adjustment is needed for the generated

standard errors.

Garen's assumptions are more restrictive than those needed for the standard IV estimator to be consistent. For one, it would be a fluke if (2.10) held without the conditional covariance $\text{Cov}(v_1, y_2|\mathbf{z})$ being independent of $\mathbf{z}$. Plus, like HV (1998), Garen relies on a linear model for $\text{E}(y_2|\mathbf{z})$. Further, Garen adds the assumptions that $\text{E}(u_1|v_2)$ and $\text{E}(v_1|v_2)$ are linear functions, something not needed by the IV approach.

Of course, one can make Garen's approach less parametric by replacing the linear functions in (2.10) with unknown functions. But independence of (u_1, v_1, v_2) and $\mathbf{z}$ – or something very close to independence – is needed. And this assumption is not needed for the usual IV estimator,

If the assumptions needed for Garen's CF estimator to be consistent hold, it is likely more efficient than the IV estimator, although a comparison of the correct asymptotic variances is complicated. Again, there is a tradeoff between efficiency and robustness.

In the case of binary y_2, we have what is often called the "switching regression" model. Now, the right hand side of equation (2.11) represents $E(y_1|\mathbf{z}, v_2)$ where $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0]$. If we assume (2.10) and that $v_2|\mathbf{z}$ is Normal$(0, 1)$, then

$$E(y_1|\mathbf{z}, y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2)y_2,$$

where

$$h_2(y_2, \mathbf{z}\boldsymbol{\delta}_2) = y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)$$

is the generalized residual function. The two-step estimation method is the one due to Heckman (1976).

There are two ways to embellish the model. The first is common: interact $(\mathbf{z}_1 - \boldsymbol{\mu}_1)$ with y_2 to allow different slopes for the "treated" and non-treated groups (keeping α_1 as the average treatment effect). This is common, and then the CF regression

$$y_{i1} \text{ on } 1, \ \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}, \ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2), \ h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}$$

is identical to running two separate regressions, including the IMRs for $y_2 = 0$ and $y_2 = 1$. The estimate of α_1 is then the difference in the two intercepts.

An extension that is not so common – in fact, it seems not to appear in the literature – comes from allowing $\mathbf{z}_1$ to also interact with heterogeneity, as in

$$y_1 = \mathbf{z}_1\mathbf{d}_1 + a_1 y_2 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\mathbf{g}_1 + u_1.$$

Now all coefficients are heterogeneous. If we assume that $E(a_1|v_2)$, $E(\mathbf{d}_1|v_2)$, and $E(\mathbf{g}_1|v_2)$ are

linear in v_2, then

$$\begin{aligned} E(y_1|\mathbf{z},y_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\xi}_1 + \rho_1 E(v_2|\mathbf{z},y_2) + \xi_1 E(v_2|\mathbf{z},y_2)y_2 \\ &\quad + \mathbf{z}_1 E(v_2|\mathbf{z},y_2)\boldsymbol{\psi}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)E(v_2|\mathbf{z},y_2)\boldsymbol{\omega}_1 \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)y_2 \\ &\quad + h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)\mathbf{z}_1\boldsymbol{\psi}_1 + h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\omega}_1 \end{aligned}$$

and the second-step estimation after the first stage probit is a regression

$$\begin{gathered} y_{i1} \text{ on } 1,\ \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2},\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}, \\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1},\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1). \end{gathered}$$

across all observations *i*. It is easy use bootstrapping to obtain valid standard errors because the first-stage estimation is just a probit and the second stage is just linear regression.

If not for the term v_1y_2, we could, in a much more robust manner, use an IV procedure (where the standard errors are easier to obtain, too). The IVs would be $[1, \mathbf{z}_{i1}, \hat{\Phi}_{i2}, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1) \cdot \hat{\Phi}_{i2}]$, and the same procedure consistently estimates the average effects whether or not there are random coefficients on $\mathbf{z}_{i1}$.

Interesting, the addition of the terms $h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1}$ and $h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)$ has similarities with methods that allow $E(v_1|v_2)$ and so on to be more flexible. For example, as shown in Heckman and MaCurdy (1986), if $E(u_1|v_2) = \rho_1 v_2 + \kappa_1(v_2^2 - 1)$, then the extra term for $y_2 = 1$ is $-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ and there is a similar expression for $y_{i2} = 0$.

Newey (1988), in the standard switching regression framework, proposed a flexible two-step procedure that estimates $\boldsymbol{\delta}_2$ semiparametrically in the first stage – see Powell (1994) for a survey of such methods – and then uses series in $\mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ in place of the usual IMR terms. He obtains valid standard errors and, in most cases, bootstrapping is valid, too.

3. Some Common Nonlinear Models and Limitations of the CF Approach

Like standard IV methods, control function approaches are more difficult to apply to nonlinear models, even relatively simple ones. Methods are available when the endogenous explanatory variables are continuous, but few if any results apply to cases with discrete y_2.

3.1. Binary and Fractional Responses

The probit model provides a good illustration of the general approach. With a single endogenous explanatory variable, the simplest specification is

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 > u_1 \geq 0], \tag{3.1}$$

where $u_1|z \sim \text{Normal}(0,1)$. But the analysis goes through if we replace (z_1, y_2) with any known function $g_1(z_1, y_2)$, provided we have sufficient identifying assumptions. An example is $y_1 = [\mathbf{z}_1\boldsymbol{\delta}_1 + y_2\mathbf{z}_1\boldsymbol{\alpha}_1 + \gamma_1 y_2^2 + u_1 > 0]$. The nonlinearity in y_2 is not itself a problem (unless we inappropriately try to mimic 2SLS – more on this later).

The Blundell-Smith (1986) and Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for y_2,

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2, \;\; v_2|\mathbf{z} \sim \text{Normal}(0, \tau_2^2). \tag{3.2}$$

A key point is that the RV approach essentially requires

$$(u_1, v_2) \text{ independent of } \mathbf{z}; \tag{3.3}$$

as we will see in the next section, semiparametric and nonparametric CF methods also rely on (3.3), or at least something close to it..

If we assume

$$(u_1, v_2) \sim \text{Bivariate Normal} \tag{3.4}$$

with $\rho_1 = Corr(u_1, v_2)$, then we can proceed with MLE based on $f(y_1, y_2|\mathbf{z})$. A simpler two-step approach, which is convenient for testing $H_0 : \rho_1 = 0$ (y_2 is exogenous) is also available, and works if we replace the normality assumption in (3.2), the independence assumption in (3.3), and joint normality in (3.4) with

$$D(u_1|v_2, \mathbf{z}) = \text{Normal}(\theta_1 v_2, 1 - \rho_1^2), \tag{3.5}$$

where $\theta_1 = \rho_1/\tau_2$ is the regression coefficient. That we can relax the assumptions to some degree using a two-step CF approach has implications for less parametric approaches. Certainly we can relax the homoskedasticity and linear expectation in (3.3) without much additional work, as discussed in Wooldridge (2005a).

Under the weaker assumption (3.5) we can write

$$P(y_1 = 1|\mathbf{z}, y_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_{\rho 1} + \alpha_{\rho 1} y_2 + \theta_{\rho 1} v_2) \tag{3.6}$$

where each coefficient is multiplied by $(1 - \rho_1^2)^{-1/2}$.

The RV two-step approach is

(i) OLS of y_2 on $\mathbf{z}$, to obtain the residuals, $\hat{v}_2$.

(ii) Probit of y_1 on $\mathbf{z}_1, y_2, \hat{v}_2$ to estimate the scaled coefficients.

The original coefficients, which appear in the partial effects, are easily obtained from the set of two-step estimates:

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\rho 1}/(1 + \hat{\theta}_{\rho 1}^2 \hat{\tau}_2^2)^{1/2}, \tag{3.7}$$

where $\hat{\theta}_{\rho 1}$ is the coeffcient on $\hat{v}_2$ and $\hat{\tau}_2^2$ is the usual error variance estimator from the first step OLS, and $\hat{\boldsymbol{\beta}}_{\rho 1}$ includes $\hat{\boldsymbol{\delta}}_{\rho 1}$ and $\hat{\alpha}_{\rho 1}$. Standard errors can be obtained from the delta method of bootstrapping. Of course, they are computed directly from MLE. Partial effects are based on $\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1)$ where $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$. Hopefully it is clear that nothing changes if $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, y_2)$ expect how one computes the partial effects.

A simple *t* test on $\hat{v}_2$ is valid to test $\mathrm{H}_0 : \rho_1 = 0$.

A different way to obtain partial effects is to use the average structural function approach, which leads to $\mathrm{E}_{v_2}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_{\rho 1})]$. Notice this holds under (3.5) without joint normality. A consistent, $\sqrt{N}$-asymptotically normal estimator is

$$\widehat{\mathrm{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\theta}_{\rho 1}\hat{v}_{i2}), \tag{3.8}$$

that is, we average out the reduced form residuals, $\hat{v}_{i2}$. This formulation is useful for more complicated models.

Given that the probit structural model is essentially arbitrary, one might be so bold as to specify models for $P(y_1 = 1|\mathbf{z}_1, y_2, v_2)$ directly. For example, we can add polynomials in v_2 or even interact v_2 with elements of $\mathbf{x}_1$ side a probit or logit function. We return to this in the next section.

The two-step CF approach easily extends to fractional responses. Now, we start with an omitted variables formulation in the conditional mean:

$$E(y_1|\mathbf{z}, y_2, q_1) = E(y_1|\mathbf{z}_1, y_2, q_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1), \tag{3.9}$$

where $\mathbf{x}_1$ is a function of $(\mathbf{z}_1, y_2)$ and q_1 contains unobservables. As usual, we need some exclusion restrictions, embodied by omitting $\mathbf{z}_2$ from $\mathbf{x}_1$. The specification in equation (3.9) allows for responses at the corners, zero and one, and y_1 may take on any values in between. Under the assumption that

$$D(q_1|v_2, \mathbf{z}) \sim \mathrm{Normal}(\theta_1 v_2, \eta_1^2) \tag{3.10}$$

Given (3.9) and (3.10), it can be shown, using the mixing property of the normal distribution, that

$$E(y_1|\mathbf{z}, y_2, v_2) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{\eta 1} + \theta_{\eta 1} v_2), \tag{3.11}$$

where the index "η" denotes coefficients multiplied by $(1 + \eta_1^2)^{-1/2}$. Because the Bernoulli log likelihood is in the linear exponential family, maximizing it consistently estimates the parameters of a correctly specified mean; naturally, the same is true for two-step estimation. That is, the *same* two-step method can be used in the binary and fractional cases. Of course, the variance associated with the Bernoulli distribution is generally incorrect. In addition to correcting for the first-stage estimates, a robust sandwich estimator should be computed to account for the fact that $D(y_1|\mathbf{z},y_2)$ is not Bernoulli. The best way to compute partial effects is to use (3.8), with the slight notational change that the implicit scaling in the coefficients is different. By using (3.8), we can directly use the scaled coefficients estimated in the second stage – a feature common across CF methods for nonlinear models. The bootstrap that reestimates the first and second stages for each iteration is an easy way to obtain standard errors. Of course, having estimates of the parameters up to a common scale allows us to determine signs of the partial effects in (3.9) as well as relative partial effects on the continuous explanatory variables.

Wooldridge (2005) describes some simple ways to make the analysis starting from (3.9) more flexible, including allowing $Var(q_1|v_2)$ to be heteroskedastic. We can also use strictly monotonic transformations of y_2 in the reduced form, say $h_2(y_2)$, regardless of how y_2 appears in the structural model: the key is that y_2 can be written as a function of $(\mathbf{z}, v_2)$. The extension to multivariate $\mathbf{y}_2$ is straightforward with sufficient instruments provide the elements of $\mathbf{y}_2$, or strictly monotonic functions of them, have reduced forms with additive errors that are effectively indendent of $\mathbf{z}$. (This assumption rules out applications to y_2 that are discrete (binary, multinomial, or count)or have a discrete component (corner solution).

The control function approach has some decided advantages over another two-step approach – one that appears to mimic the 2SLS estimation of the linear model. Rather than conditioning on v_2 along with $\mathbf{z}$ (and therefore y_2) to obtain
$\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{P}(y_1 = 1|\mathbf{z}, y_2, v_2)$, we can obtain $\mathrm{P}(y_1 = 1|\mathbf{z})$. To find the latter probability, we plug in the reduced form for y_2 to get $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2) + \alpha_1 v_2 + u_1 > 0]$. Because $\alpha_1 v_2 + u_1$ is independent of $\mathbf{z}$ and (u_1, v_2) has a bivariate normal distribution,
$\mathrm{P}(y_1 = 1|\mathbf{z}) = \Phi\{[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2)]/\omega_1\}$ where
$\omega_1^2 \equiv \mathrm{Var}(\alpha_1 v_2 + u_1) = \alpha_1^2\tau_2^2 + 1 + 2\alpha_1\mathrm{Cov}(v_2, u_1)$. (A two-step procedure now proceeds by using the same first-step OLS regression – in this case, to get the fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ – now followed by a probit of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$. It is easily seen that this method estimates the

coefficients up to the common scale factor $1/\omega_1$, which can be any positive value (unlike in the CF case, where we know the scale factor is greater than unity).

One danger with plugging in fitted values for y_2 is that one might be tempted to plug $\hat{y}_2$ into nonlinear functions, say y_2^2 or $y_2\mathbf{z}_1$. This does not result in consistent estimation of the scaled parameters or the partial effects. If we believe y_2 has a linear RF with additive normal error independent of $\mathbf{z}$, the addition of $\hat{v}_2$ solves the endogeneity problem regardless of how y_2 appears. Plugging in fitted values for y_2 only works in the case where the model is linear in y_2. Plus, the CF approach makes it much easier to test the null that for endogeneity of y_2 as well as compute APEs.

In standard index models such as (3.9), or, if you prefer, (3.1), the use of control functions to estimate the (scaled) parameters and the APEs produces no surprises. However, one must take care when, say, we allow for random slopes in nonlinear models. For example, suppose we propose a random coefficient model

$$E(y_1|\mathbf{z},y_2,\mathbf{c}_1) = E(y_1|\mathbf{z}_1,y_2,\mathbf{c}_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + a_1y_2 + q_1), \tag{3.12}$$

where a_1 is random with mean α_1 and q_1 again has mean of zero. If we want the partial effect of y_2, evaluated at the mean of heterogeneity, we have

$$\alpha_1\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2), \tag{3.13}$$

where $\phi(\cdot)$ is the standard normal pdf, and this equation is obtained by differentiating (3.12) with respect to y_2 and then plugging in $a_1 = \alpha_1$ and $q_1 = 0$. Suppose we write $a_1 = \alpha_1 + d_1$ and assume that (d_1, q_1) is bivariate normal with mean zero. Then, for given $(\mathbf{z}_1, y_2)$, the average structural function can be shown to be

$$E_{(d_1,q_1)}[\Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2 + d_1y_2 + q_1)] = \Phi[(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2)/(\sigma_q^2 + 2\sigma_{dq}y_2 + \sigma_d^2y_2^2)^{1/2}], \tag{3.14}$$

where $\sigma_q^2 = Var(q_1)$, $\sigma_d^2 = Var(d_1)$, and $\sigma_{dq} = Cov(d_1, q_1)$. The average partial effect with respect to, say, y_2, is the derivative of this function with respect to y_2. While this partial effect depends on α_1, it is messier than (3.13) and need not even have the same sign as α_1. Wooldridge (2005) discusses related issues in the context of probit models with exogenous variables and heteroskedasticity. In one example, he shows that, depending on whether heteroskedasticity in the probit is due to heteroskedasticity in $Var(u_1|\mathbf{x}_1)$, where u_1 is the latent error, or random slopes, the APEs are completely different in general. The same is true here: the APE when the coefficient on y_2 is random is generally very different from the APE obtained if we maintain $a_1 = \alpha_1$ but $Var(q_1|v_2)$ is heteroskedastic. In the latter case, the APE

is a positive multiple of α_1.

Incidentally, we can estimate the APE in (3.14) fairly generally. A parametric approach is to assume joint normality of (d_1, q_1, v_2) (and independence with $\mathbf{z}$). Then, with a normalization restriction, it can be shown that

$$E(y_1|\mathbf{z}, v_2) = \Phi[(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2)/(1 + \eta_1 y_2 + \lambda_1 y_2^2)^{1/2}], \tag{3.15}$$

which can be estimated by inserting $\hat{v}_2$ for v_2 and using nonlinear least squares or Bernoulli QMLE. (The latter is often called "heteroskedastic probit" when y_1 is binary.) This procedure can be viewed as an extension to Garen's method for linear models with correlated random coefficients.

Estimation, inference, and interpretation would be especially straightforward (the latter possibly using the bootstrap) if we squint and pretend the term $(1 + \varphi_1 y_2 + \lambda_1 y_2^2)^{1/2}$ is not present. Then, estimation would simply be Bernoulli QMLE of y_{i1} on $\mathbf{z}_{i1}$, y_{i2}, $\hat{v}_{i2}$, and $y_{i2}\hat{v}_{i2}$, which means that we just add the interaction to the usual Rivers-Vuong procedure. The APE for y_2 would be estimated by taking the derivative with respect to y_2 and averaging out $\hat{v}_{i2}$, as usual:

$$N^{-1}\sum_{i=1}^{N}(\hat{\alpha}_1 + \hat{\psi}_1\hat{v}_{i2}) \cdot \phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2 + \hat{\theta}_1\hat{v}_{i2} + \hat{\psi}_1 y_2\hat{v}_{i2}), \tag{3.16}$$

and evaluating this at chosen values for $(\mathbf{z}_1, y_2)$ (or using further averaging across the sample values). This simplification cannot be reconciled with (3.9), but it is in the spirit of adding flexibility to a standard approach and treating functional forms as approximations. As a practical matter, we can compare this with the APEs obtained from the standard Rivers-Vuong approach, and a simple test of the null hypothesis that the coefficient on y_2 is constant is $H_0 : \psi_1 = 0$ (which should account for the first step estimation of $\hat{\boldsymbol{\pi}}_2$). The null hypothesis that y_2 is exogenous is the joint test $H_0 : \theta_1 = 0, \psi_1 = 0$, and in this case no adjustment is needed for the first-stage estimation. And why stop here? If we, add, say, y_2^2 to the structural model, we might add $\hat{v}_2^2$ to the estimating equation as well. It would be very difficult to relate parameters estimated from the CF method to parameters in an underlying structural model; indeed, it would be difficult to find a structural model given rise to this particular CF approach. But if the object of interest are the average partial effects, the focus on flexible models for $E(y_1|\mathbf{z}_1, y_2, v_2)$ can be liberating (or disturbing, depending on one's point of view about "structural" parameters).

Lewbel (2000) has made some progress in estimating parameters up to scale in the model $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 > 0],$ where y_2 might be correlated with u_1 and $\mathbf{z}_1$ is a $1 \times L_1$ vector of exogenous variables. Lewbel's (2000) general approach applies to this situation as well. Let $\mathbf{z}$ be the vector of all exogenous variables uncorrelated with u_1. Then Lewbel requires a continuous element of $\mathbf{z}_1$ with nonzero coefficient – say the last element, z_{L_1}– that does not appear in $\mathrm{D}(u_1|y_2, \mathbf{z})$. (Clearly, y_2 cannot play the role of the variable excluded from $\mathrm{D}(u_1|y_2, \mathbf{z})$ if y_2 is thought to be endogenous.) When might Lewbel's exclusion restriction hold? Sufficient is $y_2 = g_2(\mathbf{z}_2) + v_2,$ where (u_1, v_2) is independent of $\mathbf{z}$ and $\mathbf{z}_2$ does not contain z_{L_1}. But this means that we have imposed an exclusion restriction on the reduced form of y_2, something usually discouraged in parametric contexts. Randomization of z_{L_1} does *not* make its exclusion from the reduced form of y_2 legitimate; in fact, one often hopes that an instrument for y_2 is effectively randomized, which means that z_{L_1} does *not* appear in the structural equation but does appear in the reduced form of y_2 – the opposite of Lewbel's assumption. Lewbel's assumption on the "special" regressor is suited to cases where a quantity that only affects the response is randomized. A randomly generated project cost presented to interviewees in a willingness-to-pay study is one possibility.

Returning to the probit response function in (3.9), we can understand the limits of the CF approach for estimating nonlinear models with discreted EEVs. The Rivers-Vuong approach, and its extension to fractional responses, cannot be expected to produce consistent estimates of the parameters or APEs for discrete y_2. The problem is that we cannot write

$$\begin{aligned} y_2 &= \mathbf{z}\boldsymbol{\pi}_2 + v_2 \\ D(v_2|\mathbf{z}) &= D(v_2) = \text{Normal}(0, \tau_2^2). \end{aligned} \tag{3.17}$$

In other words, unlike when we estimate a linear structural equation, the reduced form in the RV approach is not just a linear projection – far from it. In the extreme we have completely specified $D(y_2|\mathbf{z})$ as homoskedastic normal, which is clearly violated if y_2 is a binary or count variable, or a corner solution (commonly called a "censored" variable). Unfortunately, even just assuming independence between v_2 and $\mathbf{z}$ rules out discrete y_2, an assumption that plays an important role even in fully nonparametric approaches. The bottom line is that there are no known two-step estimation methods that allow one to estimate a probit model or fractional probit model with discrete y_2, even if we make strong distributional assumptions. And, there are some poor strategies that still linger. For example, suppose y_1 and y_2 are both binary, (3.1) holds, and

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0] \tag{3.18}$$

and we maintain joint normality of (u_1, v_2) – now both with unit variances – and, of course, independence between the errors and $\mathbf{z}$. Because $D(y_2|\mathbf{z})$ follows a standard probit, it is tempting to try to mimic 2SLS as follows: (i) Do probit of y_2 on $\mathbf{z}$ and get the fitted probabilities, $\hat{\Phi}_2 = \Phi(\mathbf{z}\hat{\boldsymbol{\delta}}_2)$. (ii) Do probit of y_1 on $\mathbf{z}_1, \hat{\Phi}_2$, that is, just replace y_2 with $\hat{\Phi}_2$. This does not work, as it requires believing that the expected value passes through nonlinear functions. Some have called prodedures like this a "forbidden regression." We could find $E(y_1|\mathbf{z}, y_2)$ as a function of the structural and reduced form parameters, insert the first-stage estimates of the RF parameters, and then use binary response estimation in the second stage. But the estimator is not probit with the fitted probabilities plugged in for y_2.Currently, the only strategy we have is maximum likelihood estimation based on $f(y_1|y_2, \mathbf{z})f(y_2|\mathbf{z})$. (The lack of options that allow some robustness to distributional assumptions on y_2 helps explain why some authors, notably Angrist (2001), have promoted the notion of just using linear probability models estimated by 2SLS. This strategy seems to provide good estimates of the average treatment effect in many applications.)

An issue that comes up occasionally is whether "bivariate" probit software be used to estimate the probit model with a binary endogenous variable. In fact, the answer is yes, and the endogenous variables can appear in any way in the model, particularly interacted with exogenous variables. The key is that the likelihood function is constructed from $f(y_1|y_2, \mathbf{x}_1)f_2(y_2|\mathbf{x}_2)$, and so its form does not change if $\mathbf{x}_1$ includes y_2. (Of course, one should have at least one exclusion restriction in the case $\mathbf{x}_1$ does depend on y_2.) MLE, of course, has all of its desirable properties, and the parameter estimates needed to compute APEs are provided directly.

If y_1 is a fractional response satisfying (3.9), y_2 follows (3.18), and (q_1, v_2) are jointly normal and independent of $\mathbf{z}$, a two-step method based on $E(y_1|\mathbf{z}, y_2)$ is possible; the expectation is not in closed form, and estimation cannot proceed by simply adding a control function to a Bernoulli QMLE. But it should not be difficult to implement. Full MLE for a fractional response is more difficult than for a binary response, particularly if y_1 takes on values at the endpoints with positive probability.

An essentially parallel discussion holds for ordered probit response models, where y_1 takes on the ordered values $\{0, 1, \ldots, J\}$. The RV procedure, and its extensions, applies immediately. In computing partial effects on the response probabilities, we simply average out the reduced

for residuals, as in equation (3.8). The comments about the forbidden regression are immediately applicable, too: one cannot simply insert, say, fitted probabilities for the binary EEV y_2 into an ordered probit model for y_1 and hope for consistent estimates of anything of interest.

Likewise, methods for Tobit models when y_1 is a corner solution, such as labor supply or charitable contributions, are analyzed in a similar fashion. If y_2 is a continuous variable, CF methods for consistent estimation can be obtained, at least under the assumptions used in the RV setup. Blundell and Smith (1986) and Wooldridge (2002, Chapter 16) contain treatments. The embellishments described above, such as letting $D(u_1|v_2)$ be a flexible normal distribution, carry over immediately to Tobit case, as do the cautions in looking for simple two-step methods when $D(y_2|\mathbf{z})$ is discrete.

3.2. Multinomial Responses

Allowing endogenous explanatory variables (EEVs) in multinomial response models is notoriously difficult, even for continuous endogenous variables. There are two basic reasons. First, multinomial probit (MNP), which mixes well well a reduced form normality assumption for $D(y_2|\mathbf{z})$, is still computationally difficult for even a moderate number of choices. Apparently, no one has undertaken a systematic treatment of MNP with EEVs, including how to obtain partial effects.

The multinomial logit (MNL), and its extensions, such as nested logit, is much simpler computationally with lots of alternatives. Unfortunately, the normal distribution does not mix well with the extreme value distribution, and so, if we begin with a structural MNL model (or conditional logit), the estimating equations obtained from a CF approach are difficult to obtain, and MLE is very difficult, too, even if we assume a normal distribution in the reduced form(s).

Recently, some authors have suggested taking a practical approach to allowing at least continuous EEVs in multinomial response. The suggestions for binary and fractional responses in the previous subsection – namely, use probit, or even logit, with flexible functions of both the observed variables and the reduced form residuals – is in this spirit.

Again it is convenient to model the source of endogeneity as an omitted variable. Let y_1 be the (unordered) multinomial response taking values $\{0, 1, \ldots, J\}$, let $\mathbf{z}$ be the vector of endogenous variables, and let $\mathbf{y}_2$ be a vector of endogenous variables. If r_1 represents omitted factors that the researcher would like to control for, then the structural model consists of specifications for the response probabilities

$$\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1), j = 0, 1, \ldots, J. \tag{3.20}$$

The average partial effects, as usual, are obtained by averaging out the unobserved heterogeneity, r_1. Assume that $\mathbf{y}_2$ follows the linear reduced form

$$\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Pi}_2 + \mathbf{v}_2. \tag{3.21}$$

Typically, at least as a first attempt, we would assume a convenient joint distribution for $(r_1, \mathbf{v}_2)$, such as multivariate normal and independent of $\mathbf{z}$. This approach has been applied when the response probabilities, conditional on r_1, have the conditional logit form. For example, Villas-Boas and Winer (1999) apply this approach to modeling brand choice, where prices are allowed to correlated with unobserved tastes that affect brand choice. In implementing the CF approach, the problem in starting with a multinomial or conditional logit model for (3.zz) is computational. Nevertheless, estimation is possible, particular if one uses simulation methods of estimation briefly mentioned in the previous subsection.

A much simpler control function approach is obtained if we skip the step of modeling $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ and jump directly to convenient models for $\mathrm{P}(y_1 = j|\mathbf{z}_{i1}, \mathbf{y}_2, \mathbf{v}_2) = \mathrm{P}(y_1 = j|\mathbf{z}, \mathbf{y}_2)$. Villas-Boas (2005) and Petrin and Train (2006) are proponents of this solution. The idea is that any parametric model for $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ is essentially arbitrary, so, if we can recover quantities of interest directly from $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, why not specify these probabilities directly? If we assume that $\mathrm{D}(r_1|\mathbf{z}, \mathbf{y}_2) = \mathrm{D}(r_1|\mathbf{v}_2)$, and that $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ can be obtained from $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ by integrating the latter with respect to $\mathrm{D}(r_1|\mathbf{v}_2)$ then we can estimate the APEs directly from $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ be averaging out across the reduced form residuals, as in previous cases.

Once we have selected a model for $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, which could be multinomial logit, conditional logit, or nested logit, we can apply a simple two-step procedure. First, estimate the reduced form for $\mathbf{y}_{i2}$ and obtain the residuals, $\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \mathbf{z}_i\hat{\boldsymbol{\Pi}}_2$. (Alternatively, we can use strictly monotonic transformations of the elements of $\mathbf{y}_{i2}$.) Then, we estimate a multinomial response model with explanatory variables $\mathbf{z}_{i1}, \mathbf{y}_{i2},$ and $\hat{\mathbf{v}}_{i2}$. As always with control function approaches, we need enough exclusion restrictions in $\mathbf{z}_{i1}$ to identify the parameters and APEs. We can include nonlinear functions of $(\mathbf{z}_{i1}, \mathbf{y}_{i2}, \hat{\mathbf{v}}_{i2})$, including quadratics and interactions for more flexibility.

Given estimates of the probabilities $p_j(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, we can estimate the average partial effects on the structural probabilities by estimating the average structural function:

$$\widehat{\text{ASF}}(\mathbf{z}_1, \mathbf{y}_2) = N^{-1} \sum_{i=1}^{N} p_j(\mathbf{z}_1, \mathbf{y}_2, \hat{\mathbf{v}}_{i2}). \tag{3.22}$$

Then, we can take derivatives or changes of $\widehat{\text{ASF}}(\mathbf{z}_1, \mathbf{y}_2)$ with respect to elements of $(\mathbf{z}_1, \mathbf{y}_2)$, as usual. While the delta method can be used to obtain analytical standard errors, the bootstrap is simpler and feasible if one uses, say, conditional logit.

In an application to choice of television service, Petrin and Train (2006) find the CF approach gives remarkably similar parameter estimates to the approach proposed by Berry, Pakes, and Levinsohn (1995), which we touch on in the cluster sample notes.

3.3. Exponential Models

Exponential models represent a middle ground between linear models and discrete response models: to allow for EEVs in an exponential model, we need to impose more assumptions than needed for standard linear models but fewer assumptions than discrete response models. Both IV approaches and CF approaches are available for exponential models, the latter having been worked out for continuous and binary EEVs. With a single EEV, write

$$E(y_1|\mathbf{z}, y_2, r_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + r_1), \tag{3.23}$$

where r_1 is the omitted variable. (Extensions to general nonlinear functions of $(\mathbf{z}_1, y_2)$ are immediate; we just add those functions with linear coefficients to (3.23). Leading cases are polynomials and interactions.) Suppose first that y_2 has a standard linear reduced form with an additive, independent error:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{3.24}$$

$$D(r_1, v_2|\mathbf{z}) = D(r_1, v_2), \tag{3.25}$$

so that (r_1, v_2) is independent of $\mathbf{z}$. Then

$$E(y_1|\mathbf{z}, y_2) = E(y_1|\mathbf{z}, v_2) = E[\exp(r_1)|v_2]\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2). \tag{3.26}$$

If (r_1, v_2) are jointly normal, then $E[\exp(r_1)|v_2] = \exp(\theta_1 v_2)$, where we set the intercept to zero, assuming $\mathbf{z}_1$ includes an intercept. This assumption can hold more generally, too. Then

$$E(y_1|\mathbf{z}, y_2) = E(y_1|\mathbf{z}, v_2) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2), \tag{3.27}$$

and this expectation immediately suggest a two-step estimation procedure. The first step, as before, is to estimate the reduced form for y_2 and obtain the residuals. Then, include $\hat{v}_2$, along with $\mathbf{z}_1$ and y_2, in nonlinear regression or, especially if y_1 is a count variable, in a Poisson QMLE analysis. Like NLS, it requires only (3.27) to hold. A t test of $H_0 : \theta_1 = 0$ is valid as a

test that y_2 is exogenous. Average partial effects on the mean are obtained from

$$\left[N^{-1} \sum_{i=1}^{N} \exp(\hat{\theta}_1 \hat{v}_{i2}) \right] \exp(\mathbf{z}_1 \hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2).$$

Proportionate effects on the expected value, that is elasticities and semi-elasticities, the expected value do not depend on the scale factor out front.

Like in the binary case, we can use a random coefficient model to suggest more flexible CF methods. For example, if we start with

$$\begin{aligned} E(y_1|\mathbf{z}, y_2, a_1, r_1) &= \exp(\mathbf{z}_1 \boldsymbol{\delta}_1 + a_1 y_2 + r_1) \\ &= \exp(\mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + d_1 y_2 + r_1) \end{aligned} \tag{3.28}$$

and assume trivariate normality of (d_1, r_1, v_2) (and independence from $\mathbf{z}$), then it can be shown that

$$\begin{aligned} E(y_1|\mathbf{z}, v_2) = \exp(&\mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2 \\ &+ (\sigma_r^2 + 2\sigma_{dr} y_2 + \sigma_d^2 y_2^2)/2). \end{aligned} \tag{3.29}$$

Therefore, the estimating equation involves a quadratic in y_2 and an interaction between y_2 and v_2. Notice that the term $(\sigma_r^2 + 2\sigma_{dr} y_2 + \sigma_d^2 y_2^2)/2$ is present even if y_2 is exogenous, that is, $\theta_1 = \psi_1 = 0$. If $\sigma_{dr} = Cov(d_1, r_1) \neq 0$ then (3.29) does not even identify $\alpha_1 = E(a_1)$ (we would have to use higher-order moments, such as a variance assumption). But (3.29) *does* identify the average structural function (and, therefore, APEs). We just absorb σ_r^2 into the intercept, combine the linear terms in y_2, and add the quadratic in y_2. So, we would estimate

$$E(y_1|\mathbf{z}, v_2) = \exp(\mathbf{z}_1 \boldsymbol{\delta}_1 + \rho_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2 + \eta_1 y_2^2) \tag{3.30}$$

using a two-step QMLE. The ASF is more complicated, and estimated as

$$\widehat{ASF}(\mathbf{z}_1, y_2) = \left[N^{-1} \sum_{i=1}^{N} \exp(\mathbf{z}_1 \hat{\boldsymbol{\delta}}_1 + \hat{\rho}_1 y_2 + \hat{\theta}_1 \hat{v}_{i2} + \hat{\psi}_1 y_2 \hat{v}_{i2} + \hat{\eta}_1 y_2^2) \right], \tag{3.31}$$

which, as in the probit example, implies that the APE with respect to y_2 need not have the same sign as α_1.

Our inability to estimate α_1 even in this very parametric setting is just one example of how delicate identification of parameters in standard index models is. Natural extensions to models with random slopes general cause even the mean heterogeneity (α_1 above) to be unidentified. Again, it must be emphasized that the loss of identification holds even if y_2 is assumed exogenous.

If y_2 is a binary model following a probit, then a CF approach due to Terza (1998) can be used. We return to the model in (3.23) where, for simplicity, we assume y_2 is not interacted with elements of $\mathbf{z}_1$; the extension is immediate.We can no longer assume (3.24) and (3.25). Instead, replace (3.24)

$$y_2 = 1[\mathbf{z}\boldsymbol{\pi}_2 + v_2 > 0] \tag{3.32}$$

and still adopt (3.25). In fact, we assume (r_1, v_2) is jointly normal. To implement a CF approach, we need to find

$$\begin{aligned} E(y_1|\mathbf{z}, y_2) &= E[E(y_1|\mathbf{z}, v_2)|\mathbf{z}, y_2] \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)E[\exp(\eta_1 + \theta_1 v_2)|\mathbf{z}, y_2] \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2, \mathbf{z}\boldsymbol{\pi}_2, \theta_1), \end{aligned} \tag{3.34}$$

where we absorb η_1 into the intercept in $\mathbf{z}_1$ without changing notation and

$$\begin{aligned} h(y_2, \mathbf{z}\boldsymbol{\pi}_2, \theta_1) = \exp(\theta_1^2/2)\{&y_2\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2) \\ &+ (1 - y_2)[1 - \Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1 - \Phi(\mathbf{z}\boldsymbol{\pi}_2)]\}, \end{aligned} \tag{3.35}$$

as shown by Terza (1998). Now, $\boldsymbol{\pi}_2$ is estimated by a first-stage probit, and then NLS or, say, Poisson QMLE can be applied to the mean function

$$\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2, \mathbf{z}\hat{\boldsymbol{\pi}}_2, \theta_1). \tag{3.36}$$

As usual, unless $\theta_1 = 0$, one must account for the estimation error in the first step when obtaining inference in the second. Terza (1998) contains analytical formulas, or one may use the bootstrap.

In the exponential case, an alternative to either of the control function approaches just presented is available – and, it produces consistent estimators regardless of the nature of y_2. Write $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, \mathbf{y}_2)$ as any function of exogenous and endogenous variables. If we start with

$$E(y_1|\mathbf{z}, \mathbf{y}_2, r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1) \tag{3.37}$$

then we can use a transformation due to Mullahy (1997) to consistently estimate $\boldsymbol{\beta}_1$ by method of moments. By definition, and assuming only that $y_1 \geq 0$, we can write

$$\begin{aligned} y_1 &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1)a_1 \\ &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1)\exp(r_1)a_1, \; E(a_1|\mathbf{z}, \mathbf{y}_2, r_1) = 1. \end{aligned}$$

If r_1 is independent of $\mathbf{z}$ then

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1|\mathbf{z}] = E[\exp(r_1)|\mathbf{z}] = E[\exp(r_1)] \; = 1, \tag{3.38}$$

where the last equality is just a normalization that defines the intercept in $\boldsymbol{\beta}_1$. Therefore, we have conditional moment conditions

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 - 1|\mathbf{z}] = 0, \tag{3.39}$$

which depends on the unknown parameters $\boldsymbol{\beta}_1$ and observable data. Any function of $\mathbf{z}$ can be used as instruments in a nonlinear GMM procedure. An important issue in implementing the procedure is choosing instruments. See Mullahy (1997) for further discussion.

4. Semiparametric and Nonparametric Approaches

Blundell and Powell (2004) show how to relax distributional assumptions on (u_1, v_2) in the model model $y_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0]$, where $\mathbf{x}_1$ can be any function of $(\mathbf{z}_1, y_2)$. The key assumption is that y_2 can be written as $y_2 = g_2(\mathbf{z}) + v_2,$ where (u_1, v_2) is independent of $\mathbf{z}$. The independence of the additive error v_2 and $\mathbf{z}$ pretty much rules out discreteness in y_2, even though $g_2(\cdot)$ can be left unspecified. Under the independence assumption,

$$\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{E}(y_1|\mathbf{z}, v_2) = H(\mathbf{x}_1\boldsymbol{\beta}_1, v_2) \tag{4.1}$$

for some (generally unknown) function $H(\cdot, \cdot)$. The average structural function is just $\mathrm{ASF}(\mathbf{z}_1, y_2) = \mathrm{E}_{v_{i2}}[H(\mathbf{x}_1\boldsymbol{\beta}_1, v_{i2})]$. We can estimate H and $\boldsymbol{\beta}_1$ quite generally by first estimating the function $g_2(\cdot)$ and then obtaining residuals $\hat{v}_{i2} = y_{i2} - \hat{g}_2(\mathbf{z}_i)$. Blundell and Powell (2004) show how to estimate H and $\boldsymbol{\beta}_1$ (up to scaled) and $G(\cdot)$, the distribution of u_1. The ASF is obtained from $G(\mathbf{x}_1\boldsymbol{\beta}_1)$. We can also estimate the ASF by averaging out the reduced form residuals,

$$\widehat{\mathrm{ASF}}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N}\hat{H}(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1, \hat{v}_{i2}); \tag{4.2}$$

derivatives and changes can be computed with respect to elements of $(\mathbf{z}_1, y_2)$.

Blundell and Powell (2003) allow $P(y_1 = 1|\mathbf{z}, y_2)$ to have the general form $H(\mathbf{z}_1, y_2, v_2)$, and then the second-step estimation is entirely nonparametric. They also allow $\hat{g}_2(\cdot)$ to be fully nonparametric. But parametric approximations in each stage might produce good estimates of the APEs. For example, y_{i2} can be regressed on flexible functions of $\mathbf{z}_i$ to obtain $\hat{v}_{i2}$. Then, one can estimate probit or logit models in the second stage that include functions of $\mathbf{z}_1$, y_2, and $\hat{v}_2$ in a flexible way – for example, with levels, quadratics, interactions, and maybe even higher-order polynomials of each. Then, one simply averages out $\hat{v}_{i2}$, as in equation (4.2). Valid standard errors and test statistics can be obtained by bootstrapping or by using the delta

method.

In certain cases, an even more parametric approach suggests itself. Suppose we have the exponential regression

$$E(y_1|\mathbf{z},y_2,r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1), \tag{4.3}$$

where r_1 is the unobservable. If $y_2 = \mathbf{g}_2(\mathbf{z})\boldsymbol{\pi}_2 + v_2$ and (r_1, v_2) is independent of $\mathbf{z}$, then

$$E(y_1|\mathbf{z}_1,y_2,v_2) = h_2(v_2)\exp(\mathbf{x}_1\boldsymbol{\beta}_1), \tag{4.4}$$

where now $h_2(\cdot)$ is an unknown function. It can be approximated using series, say, and, of course, first-stage residuals $\hat{v}_2$ replace v_2.

Blundell and Powell (2003) consider a very general setup, which starts with $y_1 = g_1(\mathbf{z}_1, \mathbf{y}_2, u_1)$, and then discuss estimation of the ASF, given by

$$ASF_1(\mathbf{z}_1,\mathbf{y}_2) = \int g_1(\mathbf{z}_1,\mathbf{y}_2,u_1)dF_1(u_1), \tag{4.5}$$

where F_1 is the distribution of u_1. The key restrictions are that $\mathbf{y}_2$ can be written as

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2, \tag{4.6}$$

where $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$. The additive, independent reduced form errors in (4.6) effectively rule out applications to discrete $\mathbf{y}_2$. Conceptually, Blundell and Powell's method is straightforward, as it is a nonparametric extenstion of parametric approaches. First, estimate $\mathbf{g}_2$ nonparametrically (which, in fact, may be done via a flexible parametric model, or kernel estimators). Obtain the residuals $\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \hat{\mathbf{g}}_2(\mathbf{z}_i)$. Next, estimate $E(y_1|\mathbf{z}_1,\mathbf{y}_2,\mathbf{v}_2) = h_1(\mathbf{z}_1,\mathbf{y}_2,\mathbf{v}_2)$ using nonparametrics, where $\hat{\mathbf{v}}_{i2}$ replaces $\mathbf{v}_2$. Identification of h_1 holds quite generally, provided we have sufficient exclusion restrictions (elements in $\mathbf{z}$ not in $\mathbf{z}_1$). BP discuss some potential pitfalls. Once we have $\hat{h}_1$, we can consistently estimate the ASF. For given $\mathbf{x}_1^o = (\mathbf{z}_1^o, \mathbf{y}_2^o)$, the ASF can always be written, using iterated expectations, as

$$E_{\mathbf{v}_2}\{E[g_1(\mathbf{x}_1^o,u_1)|\mathbf{v}_2]\}.$$

Under the assumption that $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$, $E[g_1(\mathbf{x}_1^o,u_1)|\mathbf{v}_2] = h_1(\mathbf{x}_1^o,\mathbf{v}_2)$ – that is, the regression function of y_1 on $(\mathbf{x}_1, \mathbf{v}_2)$. Therefore, a consistent estimate of the ASF is

$$N^{-1}\sum_{i=1}^{N}\hat{h}_1(\mathbf{x}_1,\hat{\mathbf{v}}_{i2}). \tag{4.7}$$

While semiparametric and parametric methods when y_2 (or, more generally, a vector $\mathbf{y}_2$) are continuous – actually, have a reduced form with an additive, independent error – they do

not currently help us with discrete EEVs.

With univariate y_2, it possible to relax the additivity of v_2 in the reduced form equation under monotonicity assumptions. Like Blundell and Powell (2003), Imbens and Newey (2006) consider the triangular system, but without additivity in the reduced form of y_2:

$$y_1 = g_1(\mathbf{z}_1, y_2, \mathbf{u}_1), \tag{4.8}$$

where $\mathbf{u}_1$ is a vector heterogeneity (whose dimension may not even be known)

$$y_2 = g_2(\mathbf{z}, e_2), \tag{4.9}$$

where $g_2(\mathbf{z}, \cdot)$ is strictly monotonic. This assumption rules out discrete y_2 but allows some interaction between the unobserved heterogeneity in y_2 and the exogenous variables. As one special case, Imbens and Newey show that, if $(\mathbf{u}_1, e_2)$ is assumed to be independent of $\mathbf{z}$, then a valid control function to be used in a second stage is $v_2 \equiv F_{y_2|\mathbf{z}}(y_2, \mathbf{z})$, where $F_{y_2|\mathbf{z}}$ is the conditional distribution of y_2 given $\mathbf{z}$. Imbens and Newey described identification of various quantities of interest, including the quantile structural function. When u_1 is a scalar and monotonically increasing in u_1, the QSF is

$$QSF_\tau(\mathbf{x}_1) = g_1(\mathbf{x}_1, \text{Quant}_\tau(u_1)), \tag{4.10}$$

where $\text{Quant}_\tau(u_1)$ is the τ^{th} of u_1. We consider quantile methods in more detail in the quantile methods notes.

5. Methods for Panel Data

We can combine methods for handling correlated random effects models with control function methods to estimate certain nonlinear panel data models with unobserved heterogeneity and EEVs. Here we provide as an illustration a parametric approach used by Papke and Wooldridge (2007), which applies to binary and fractional responses. The manipulations are routine but point to more flexible ways of estimating the average marginal effects. It is important to remember that we currently have no way of estimating, say, unobserved effects models for fractional response variables, either with or without endogenous explanatory variables. Even the approaches that treat the unobserved effects as parameters – and use large T approximations – to not allow endogenous regressors. Plus, recall from the nonlinear panel data notes that most results are for the case where the data are assumed independent across time. Jackknife approaches further assume homogeneity across time.

We write the model with time-constant unobserved heterogeneity, c_{i1}, and time-varying unobservables, v_{it1}, as

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}) = E(y_{it1}|y_{it2}, \mathbf{z}_{it1}, c_{i1}, v_{it1}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1}). \tag{5.1}$$

Thus, there are two kinds of potential omitted variables. We allow the heterogeneity, c_{i1}, to be correlated with y_{it2} and $\mathbf{z}_i$, where $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$ is the vector of strictly exogenous variables (conditional on c_{i1}). The time-varying omitted variable is uncorrelated with $\mathbf{z}_i$ – strict exogeneity – but may be correlated with y_{it2}. As an example, y_{it1} is a female labor force participation indicator and y_{it2} is other sources of income. Or, y_{it1} is a test pass rate, and the school leve, and y_{it2} is a measure of spending per student.

When we write $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$, we are assuming $\mathbf{z}_{it2}$ can be excluded from the "structural" equation (4.1). This is the same as the requirement for fixed effects two stage least squares estimation of a linear model.

To proceed, we first model the heterogeneity using a Chamberlain-Mundlak approach:

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}, a_{i1}|\mathbf{z}_i \sim \text{Normal}(0, \sigma^2_{a_1}). \tag{5.2}$$

We could allow the elements of $\mathbf{z}_i$ to appear with separate coefficients, too. Note that only exogenous variables are included in $\bar{\mathbf{z}}_i$. Plugging into (5.1) we have

$$\begin{aligned} E(y_{it1}|y_{it2}, \mathbf{z}_i, a_{i1}, v_{it1}) &= \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1} + v_{it1}) \\ &\equiv \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + r_{it1}). \end{aligned} \tag{5.3}$$

Next, we assume a linear reduced form for y_{it2}:

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\boldsymbol{\delta}_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}, t = 1, \ldots, T, \tag{5.4}$$

where, if necessary, we can allow the coefficients in (5.4) to depend on *t*. The addition of the time average of the strictly exogenous variables in (5.4) follows from the Mundlak (1978) device. The nature of endogeneity of y_{it2} is through correlation between $r_{it1} = a_{i1} + v_{it1}$ and the reduced form error, v_{it2}. Thus, y_{it2} is allowed to be correlated with unobserved heterogeneity and the time-varying omitted factor. We also assume that r_{it1} given v_{it2} is conditionally normal, which we write as

$$r_{it1} = \eta_1 v_{it2} + e_{it1}, \tag{5.5}$$

$$e_{it1}|(\mathbf{z}_i, v_{it2}) \sim \text{Normal}(0, \sigma^2_{e_1}), t = 1, \ldots, T. \tag{5.6}$$

Because e_{it1} is independent of $(\mathbf{z}_i, v_{it2})$, it is also independent of y_{it2}. Using a standard mixing property of the normal distribution,

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2}) = \Phi(\alpha_{e1} y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2}) \tag{5.7}$$

where the "*e*" subscript denotes division by $(1 + \sigma_{e_1}^2)^{1/2}$. This equation is the basis for CF estimation.

The assumptions used to obtain (5.7) would not hold for y_{it2} having discreteness or substantively limited range in its distribution. It is straightfoward to include powers of v_{it2} in (5.7) to allow greater flexibility. Following Wooldridge (2005) for the cross-sectional case, we could even model r_{it1} given v_{it2} as a heteroskedastic normal.

In deciding on estimators of the parameters in (5.7), we must note that the explanatory variables, while contemporaneous exogenous by construction, are not usually strictly exogenous. In particular, we allow y_{is2} to be correlated with v_{it1} for $t \neq s$. Therefore, generalized estimation equations, that assume strict exogeneity – see the notes on nonlinear panel data models – will not be consistent in general. We could apply method of moments procedures. A simple approach is to use use pooled nonlinear least squares or pooled quasi-MLE, using the Bernoulli log likelihood. (The latter fall under the rubric of generalized linear models.) Of course, we want to allow arbitrary serial dependence and $Var(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2})$ in obtaining inference, which means using a robust sandwich estimator.

The two step procedure is (i) Estimate the reduced form for y_{it2} (pooled across *t*, or maybe for each *t* separately; at a minimum, different time period intercepts should be allowed). Obtain the residuals, $\hat{v}_{it2}$ for all (i,t) pairs. The estimate of $\boldsymbol{\delta}_2$ is the fixed effects estimate. (ii) Use the pooled probit QMLE of y_{it1} on $y_{it2}, \mathbf{z}_{it1}, \bar{\mathbf{z}}_i, \hat{v}_{it2}$ to estimate $\alpha_{e1}, \boldsymbol{\delta}_{e1}, \psi_{e1}, \boldsymbol{\xi}_{e1}$ and η_{e1}.

Because of the two-step procedure, the standard errors in the second stage should be adjusted for the first stage estimation. Alternatively, bootstrapping can be used by resampling the cross-sectional units. Conveniently, if $\eta_{e1} = 0$, the first stage estimation can be ignored, at least using first-order asymptotics. Consequently, a test for endogeneity of y_{it2} is easily obtained as an asymptotic *t* statistic on $\hat{v}_{it2}$; it should be make robust to arbitrary serial correlation and misspecified variance. Adding first-stage residuals to test for endogeneity of an explanatory variables dates back to Hausman (1978). In a cross-sectional contexts, Rivers and Vuong (1988) suggested it for the probit model.

Estimates of average partial effects are based on the average structural function

$$E_{(c_{i1}, v_{it1})}[\Phi(\alpha_1 y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1})] \tag{5.8}$$

with respect to the elements of $(y_{t2}, \mathbf{z}_{t1})$. It can be shown that

$$E_{(\bar{\mathbf{z}}_i, v_{it2})}[\Phi(\alpha_{e1} y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2})]; \tag{5.9}$$

that is, we "integrate out" $(\bar{\mathbf{z}}_i, v_{it2})$ and then take derivatives or changes with respect to the elements of $(\mathbf{z}_{t1} y_{t2})$. Because we are not making a distributional assumption about $(\bar{\mathbf{z}}_i, v_{it2})$, we instead estimate the APEs by averaging out $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$ across the sample, for a chosen *t*:

$$N^{-1} \sum_{i=1}^{N} \Phi(\hat{\alpha}_{e1} y_{t2} + \mathbf{z}_{t1} \hat{\boldsymbol{\delta}}_{e1} + \hat{\psi}_{e1} + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_{e1} + \hat{\eta}_{e1} \hat{v}_{it2}). \tag{5.10}$$

APEs computed from (5.10) – typically with further averaging out across *t* and the values of y_{t2} and $\mathbf{z}_{t1}$ – can be compared directly with linear model estimates, particular fixed effects IV estimates.

We can use the approaches of Altonji and Matzkin (2005) and Blundell and Powell (2003) to make the analysis less parametric. For example, we might replace (5.4) with $y_{it2} = g_2(\mathbf{z}_{it}, \bar{\mathbf{z}}_i) + v_{it2}$ (or use functions in addition to $, \bar{\mathbf{z}}_i$, as in AM). Then, we could maintain

$$D(c_{i1} + v_{it1} | \mathbf{z}_i, v_{it2}) = D(c_{i1} + v_{it1} | \bar{\mathbf{z}}_i, v_{it2}).$$

In the first estimation step, $\hat{v}_{it2}$ is obtained from a nonparametric or semiparametric pooled estimation. Then the function

$$E(y_{it1} | y_{it2}, \mathbf{z}_i, v_{it2}) = h_1(\mathbf{x}_{it1} \boldsymbol{\beta}_1, \bar{\mathbf{z}}_i, v_{it2})$$

can be estimated in a second stage, with the first-stage residuals, $\hat{v}_{it2}$, inserted. Generally, identification holds because the v_{it2} varying over time separately from $\mathbf{x}_{it1}$ due to time-varying exogenous instruments $\mathbf{z}_{it2}$. The inclusion of $\bar{\mathbf{z}}_i$ requires that we have at least one time-varying, strictly exogenous instrument for y_{it2}.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 7, Tuesday, July 31th, 11.00-12.30pm

Bayesian Inference

1. Introduction

In this lecture we look at Bayesian inference. Although in the statistics literature explicitly Bayesian papers take up a large proportion of journal pages these days, Bayesian methods have had very little impact in economics. This seems to be largely for historial reasons. In many empirical settings in economics Bayesian methods appear statistically more appropriate, and computationally more attractive, than the classical or frequentist methods typically used. Recent textbooks discussing modern Bayesian methods with an applied focus include Lancaster (2004) and Gelman, Carlin, Stern and Rubin (2004).

One important consideration is that in practice frequentist and Baeysian inferences are often very similar. In a regular parametric model, conventional confidence intervals around maximum likelihood (that is, the maximum likelihood estimate plus or minus 1.96 times the estimated standard error), which formally have the property that whatever the true value of the parameter is, with probability 0.95 the confidence interval covers the true value, can in fact also be interpreted as approximate Bayesian probability intervals (that is, conditional on the data and given a wide range of prior distributions, the posterior probability that the parameter lies in the confidence interval is approximately 0.95). The formal statement of this remarkable result is known as the Bernstein-Von Mises theorem. This result does not always apply in irregular cases, such as time series settings with unit roots. In those cases there are more fundamental differences between Bayesian and frequentist methods.

Typically a number of reasons are given for the lack of Bayesian methods in econometrics. One is the difficulty in choosing prior distributions. A second reason is the need for a fully specified parametric model. A third is the computational complexity of deriving posterior distributions. None of these three are compelling.

Consider first the specification of the prior distribution. In regular cases the influence of the prior distribution disappears as the sample gets large, as formalized in the Bernstein-Von Mises theorem. This is comparable to the way in which in large samples normal approximations can be used for the finite sample distributions of classical estimators. If, on the other hand, the posterior distribution is quite sensitive to the choice of prior distribution, then it is likely that the sampling distribution of the maximum likelihood estimator is not well approximated by a normal distribution centered at the true value of the parameter in a frequentist analysis.

A conventional Bayesian analysis does require a fully specified parameter model, as well as a prior distribution on all the parameters of this model. In frequentist analyses it is often possible to specify only part of the model, and use a semi-parametric approach. This advantage is not as clear cut as it may seem. When the ultimate questions of interest do not depend on certain features of the distribution, the results of a parametric model are often robust given a flexible specification of the nuisance functions. As a result, extending a semi-parametric model to a fully parametric one by flexibly modelling the nonparametric component often works well in practice. In addition, Bayesian semi-parametric methods have been developed.

Finally, traditionally computational difficulties held back applications of Bayesian methods. Modern computational advances, in particular the development of markov chain monte carlo methods, have reduced, and in many cases eliminated, these difficulties. Bayesian analyses are now feasible in many settings where they were not twenty years ago. There are now few restrictions on the type of prior distributions that can be considered and the dimension of the models used.

Bayesian methods are especially attractive in settings with many parameters. Examples discussed in these notes include panel data with individual-level heterogeneity in multiple parameters, instrumental variables with many instruments, and discrete choice data with multiple unobserved product characteristics. In such settings, methods that attempt to estimate every parameter precisely without linking it to similar parameters, often have poor

repeated sampling properties. This shows up in Bayesian analyses in the dogmatic posterior distributions resulting from flat prior distributions. A more attractive approach, that is succesfuly applied in the aforementioned examples, can be based on hierarchical prior distributions where the parameters are assumed to be drawn independently from a common distribution with unknown location and scale. The recent computational advances make such models feasible in many settings.

2. Bayesian Inference

The formal set up is the following: we have a random variable X, which is known to have a probability density, or probability mass, function conditional on an unknown parameter θ. We are interested the value of the parameter θ, given one or more independent draws from the conditional distribution of X given θ. In addition we have prior beliefs about the value of the parameter θ. We will capture those prior beliefs in a prior probability distribution. We then combine this prior distribution and the sample information, using Bayes' theorem, to obtain the conditional distribution of the parameter given the data.

2.1 The General Case

Now let us do this more formally. There are two ingredients to a Bayesian analysis. First a model for the data given some unknown parameters, specified as a probability (density) function:

$$f_{X|\theta}(x|\theta).$$

As a function of the parameter this is called the likelihood function, and denoted by $\mathcal{L}(\theta)$ or $\mathcal{L}(\theta|x)$. Second, a prior distribution for the parameters, $p(\theta)$. This prior distribution is known to, that is, choosen by the researcher. Then, using Bayes' theorem we calculate the conditional distribution of the parameters given the data, also known as the posterior distribution,

$$p(\theta|x) = \frac{f_{X,\theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\theta}(x|\theta)\cdot p(\theta)}{\int f_{X|\theta}(x|\theta)\cdot p(\theta)d\theta}.$$

In this step we often use a shortcut. First note that, as a function of θ, the conditional

density of θ given X is proportional to

$$p(\theta|x) \propto f_{X|\theta}(x|\theta) \cdot p(\theta) = \mathcal{L}(\theta|x) \cdot p(\theta).$$

Once we calculate this product, all we have to do is found the constant that makes this expression integrate out to one as a function of the parameter. At that stage it is often easy to recognize the distribution and figure out through that route what the constant is.

2.2 A Normal Example with Unknown Mean and Known Variance, and a Single Observation

Let us look at a simple example. Suppose the conditional distribution of X given the parameter μ is normal with mean μ and variance 1, denoted by $\mathcal{N}(\mu, 1)$. Suppose we choose the prior distribution for μ to be normal with mean zero and variance 100, $\mathcal{N}(0, 100)$. What is the posterior distribution of μ given $X = x$? The posterior distribution is proportional to

$$\begin{aligned} f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2\cdot 100}\mu^2\right) \\ &= \exp -\frac{1}{2}\Big(x^2 - 2x\mu + \mu^2 + \mu^2/100\Big) \\ &\propto \exp\Big(-\frac{1}{2(100/101)}\Big(\mu - (100/101)x\Big)^2\Big). \end{aligned}$$

This implies that the conditional distribution of μ given $X = x$ is normal with mean $(100/101) \cdot x$ and variance $100/101$, or $\mathcal{N}(x \cdot 100/101, 100/101)$.

In this example the model was a normal distribution for X given the unknown mean μ, and we choose a normal prior distribution. This was a very convenient choice, leading the posterior distribution to be normal as well. In this case the normal prior distribution is a conjugate prior distribution, implying that the posterior distribution is in the same family of distributions as the prior distribution, allowing for analytic calculations. If we had choosen a different prior distribution it would typically not have been possible to obtain an analytic expression for the posterior distribution.

Let us continue the normal distribution example, but generalize the prior distribution. Suppose that given μ the random variable X has a normal distribution with mean μ and

known variance σ^2, or $\mathcal{N}(\mu,\sigma^2)$. The prior distribution for μ is normal with mean μ_0 and variance τ^2, or $\mathcal{N}(\mu_0,\tau^2)$. Then the posterior distribution is proportional to:

$$\begin{aligned}
f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\cdot\exp\left(-\frac{1}{2\cdot\tau^2}(\mu-\mu_0)^2\right)\\
&\propto \exp-\frac{1}{2}\left(\frac{x^2}{\sigma^2}-\frac{2x\mu}{\sigma^2}+\frac{\mu^2}{\sigma^2}+\frac{\mu^2}{\tau^2}-\frac{2\mu\mu_0}{\tau^2}+\frac{\mu_0^2}{\tau^2}\right)\\
&\propto \exp-\frac{1}{2}\left(\mu^2\frac{\sigma^2+\tau^2}{\tau^2\sigma^2}-\mu\frac{2x\tau^2+2\mu_0\sigma^2}{\tau^2\cdot\sigma^2}\right)\\
&\propto \exp-\frac{1}{2(1/(1/\tau^2+1/\sigma^2))}\left((\mu-(x/\sigma^2+\mu_0/\tau^2)/(1/\sigma^2+1/\tau^2)\right)\\
&\sim \mathcal{N}\left(\frac{x/\sigma^2+\mu_0/\tau^2}{1/\sigma^2+1/\tau^2},\frac{1}{1/\tau^2+1/\sigma^2)}\right).
\end{aligned}$$

The result is quite intuitive: the posterior mean is a weighted average of the prior mean μ_0 and the observation x with weights adding up to one and proportional to the precision (defined as one over the variance), $1/\sigma^2$ for x and $1/\tau^2$ for μ_0:

$$\mathbb{E}[\mu|X=x]=\frac{\frac{x}{\sigma^2}+\frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}.$$

The posterior precision is obtained by adding up the precision for each component

$$\frac{1}{\mathbb{V}(\mu|X)}=\frac{1}{\sigma^2}+\frac{1}{\tau^2}.$$

So, what you expect *ex post*, $\mathbb{E}[\mu|X]$, that is, after seeing the data, is a weighted average of what you expected before seeing the data, $\mathbb{E}[\mu]=\mu_0$, and the observation, X, with the weights determined by their respective variances.

There are a number of insights obtained by studying this example more carefully. Suppose we are really sure about the value of μ before we conduct the experiment. In that case we would set τ^2 small and the weight given to the observation would be small, and the posterior distribution would be close to the prior distribution. Suppose on the other hand we are very unsure about the value of μ. What value for τ should we choose? Obviously a large value, but what is the limit? We can in fact let τ go to infinity. Even though the prior distribution is not a proper distribution anymore if $\tau^2=\infty$, the posterior distribution is

perfectly well defined, namely $\mu|X \sim \mathcal{N}(X, \sigma^2)$. In that case we have an improper prior distribution. We give equal prior weight to any value of μ. That would seem to capture pretty well the idea that a priori we are ignorant about μ. This is the idea of looking for an relatively uninformative prior distribution. This is not always easy, and the subject of a large literature. For example, a flat prior distribution is not always uninformative about particular functions of parameters.

2.3 A Normal Example with Unknown Mean and Known Variance and Multiple Observations

Now suppose we have N independent draws from a normal distribution with unknown mean μ and known variance σ^2. Suppose we choose, as before, the prior distribution to be normal with mean μ_0 and variance τ^2, or $\mathcal{N}(\mu_0, \tau^2)$.

The likelihood function is

$$\mathcal{L}(\mu|\sigma^2, x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right),$$

so that with a normal (μ_0, τ^2) prior distribution the posterior distribution is proportional to

$$p(\mu|x_1, \ldots, x_N) \propto \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right) \cdot \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right).$$

Thus, with N observations $x_1, \ldots, x_N$ we find, after straightforward calculations,

$$\mu|X_1, \ldots, X_N \sim \mathcal{N}\left(\frac{\mu_0/\tau^2 + \sum x_i/\sigma^2}{1/\tau^2 + N/\sigma^2}, \frac{1}{1/\tau^2 + N/\sigma^2}\right).$$

2.4 The Normal Distribution with Known Mean and Unknown Variance

Let us also briefly look at the case of a normal model with known mean and unknown variance. Thus,

$$X_i|\sigma^2 \sim \mathcal{N}(0, \sigma^2),$$

and $X_1, \ldots, X_N$ independent given σ^2. The likelihood function is

$$\mathcal{L}(\sigma^2) = \prod_{i=1}^{N} \sigma^{-N} \exp\left(-\frac{1}{2\sigma^2} X_i^2\right).$$

Now suppose that the prior distribution for σ^2 is, for some fixed S_0^2 and K_0, such that the distribution of $\sigma^{-2} \cdot S_0^2 \cdot K_0$ is chi-squared with K_0 degrees of freedom. In other words, the prior distribution of σ^{-2} is $(S_0^2 \cdot K_0)^{-1}$ times a chi-squared distribution with K_0 degrees of freedom. Then the posterior distribution of σ^{-2} is $(S_0^2 \cdot K_0 + \sum_i X_i^2)^{-1}$ times a chi-squared distribution with $K_0 + N$ degrees of freedom, so this is a conjugate prior distribution.

3. The Bernstein-Von Mises Theorem

Let us go back to the normal example with N observations, and unknown mean and known variance. In that case with a normal $\mathcal{N}(\mu_0, \tau^2)$ prior distribution the posterior for μ is

$$\mu|x_1, \ldots, x_N \sim \mathcal{N}\left(\overline{x} \cdot \frac{1}{1+\sigma^2/(N\cdot\tau^2)} + \mu_0 \cdot \frac{\sigma^2/(N\tau^2)}{1+\sigma^2/(N\tau^2)}, \frac{\sigma^2/N}{1+\sigma^2/(N\tau^2)}\right).$$

When N is very large, the distribution of $\sqrt{N}(\mu - \bar{x})$ conditional on the data is approximately

$$\sqrt{N}(\bar{x} - \mu)|x_1, \ldots, x_N \sim \mathcal{N}(0, \sigma^2).$$

In other words, in large samples the influence of the prior distribution disappears, unless the prior distribution is choosen particularly badly, e.g., τ^2 equal to zero. This is true in general, i.e., for different models and different prior distributions. Moreover, in a frequentist analysis we would find that in large samples (and in this specific normal example even in finite samples),

$$\sqrt{N}(\bar{x} - \mu)|\mu \sim \mathcal{N}(0, \sigma^2).$$

Let us return to the Bernoulli example to see the same point. Suppose that conditional on the parameter $P = p$, the random variables $X_1, X_2, \ldots, X_N$ are independent with Bernoulli distributions with probability p. Let the prior distribution of P be Beta with parameters α and β, or $\mathcal{B}(\alpha, \beta)$. Now consider the conditional distribution of P given $X_1, \ldots, X_N$:

$$f_{P|X_1,\ldots,X_N}(p|x) \propto p^{\alpha-1+\sum_{i=1}^N X_i} \cdot (1-p)^{\beta-1+N-\sum_{i=1}^N X_i},$$

which is a Beta distribution, $\mathcal{B}(\alpha - 1 + \sum_{i=1}^N X_i, \beta - 1 + N - \sum_{i=1}^N X_i)$, with mean and

variance

$$\mathbb{E}[P|X_1,\ldots,X_N] = \frac{\alpha + \sum_{i=1}^N X_i}{\alpha+\beta+N}, \qquad \text{and} \quad \mathbb{V}(P) = \frac{(\alpha + \sum_{i=1}^N X_i)(\beta + N - \sum_{i=1}^N X_i)}{(\alpha+\beta+N)^2(\alpha+\beta+1+N)}.$$

What happens if N gets large? Let $\hat{p} = \sum_i X_i/N$ be the relative frequency of success (which is the maximum likelihood estimator for p). Then the mean and variance converge to

$$\mathbb{E}[P|X_1,\ldots,X_N] \approx \hat{p},$$

and

$$\mathbb{V}(P) \approx 0.$$

As the sample size gets larger, the posterior distribution becomes concentrated at a value that does not depend on the prior distribution. This in fact can be taken a step further. In this example, the limiting distribution of $\sqrt{N}\cdot(P-\hat{p})$ conditional on the data, can be shown to be

$$\sqrt{N}(\hat{p} - P|x_1,\ldots,x_N \xrightarrow{d} \mathcal{N}(0,\hat{p}(1-\hat{p})),$$

again irrespective of the choice of α and β. The interpretation of this result is very important: in large sample the choice of prior distribution is not important in the sense that the information in the prior distribution gets dominated by the sample information. That is, unless your prior beliefs are so strong that they cannot be overturned by evidence (i.e., the prior distribution is zero over some important range of the parameter space), at some point the evidence in the data outweighs any prior beliefs you might have started out with.

This is known as the Bernstein-von Mises Theorem. Here is a general statement for the scalar case. Let the information matrix $\mathcal{I}_\theta$ at θ:

$$\mathcal{I}_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ln f_X(x|\theta)\right] = -\int \frac{\partial^2}{\partial\theta\partial\theta'}\ln f_X(x|\theta) f_X(x|\theta)dx,$$

and let σ^2 be the inverse at a fixed value θ_0.

$$\sigma^2 = \mathcal{I}_{\theta_0}^{-1}.$$

Let $p(\theta)$ be the prior distribution, and $p_{\theta|X_1,\ldots,X_N}(\theta|X_1,\ldots,X_N)$ be the posterior distribution. Now let us look at the distribution of a transformation of θ, $\gamma = \sqrt{N}(\theta-\theta_0)$, with density $p_{\gamma|X_1,\ldots,X_N}(\gamma|X_1,\ldots,X_N) = p_{\theta|X_1,\ldots,X_N}(\theta_0+\sqrt{N}\cdot\gamma|X_1,\ldots,X_N)/\sqrt{N}$. Now let us look at the posterior distribution for θ if in fact the data were generated by $f(x|\theta)$ with $\theta=\theta_0$. In that case the posterior distribution of γ converges to a normal distribution with mean zero and variance equal to σ^2 in the sense that

$$\sup_{\gamma}\left|p_{\gamma|X_1,\ldots,X_N}(\gamma|X_1,\ldots,X_N) - \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\gamma^2\right)\right| \longrightarrow 0.$$

See Van der Vaart (2001), or Ferguson (1996). At the same time, if the true value is θ_0, then the mle $\hat{\theta}_{mle}$ also has a limiting distribution with mean zero and variance σ^2:

$$\sqrt{N}(\hat{\theta}_{ml}-\theta_0) \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

The implication is that we can interpret confidence intervals as approximate probability intervals from a Bayesian perspective. Specifically, let the 95% confidence interval be $[\hat{\theta}_{ml} - 1.96\cdot\hat{\sigma}/\sqrt{N}, \hat{\theta}_{ml}+1.96\cdot\hat{\sigma}/\sqrt{N}]$. Then, approximately,

$$\Pr\left(\hat{\theta}_{ml}-1.96\cdot\hat{\sigma}/\sqrt{N} \leq \theta \leq \hat{\theta}_{ml}+1.96\cdot\hat{\sigma}/\sqrt{N}\,\middle|\, X_1,\ldots,X_N\right) \longrightarrow 0.95.$$

There are important cases where this result does not hold, typically when convergence to the limit distribution is not uniform. One is the unit-root setting. In a simple first order autoregressive example it is still the case that with a normal prior distribution for the autoregressive parameter the posterior distribution is normal (see Sims and Uhlig, 1991). However, if the true value of the autoregressive parameter is unity, the sampling distribution is not normal even in large samples. In that case one has to take a more principled stand whether one wants to make subjective probability statements, or frequentist claims.

4. Markov Chain Monte Carlo Methods

Are we really restricted to choosing the prior distributions in these conjugate families as we did in the examples so far? No. The posterior distributions are well defined irrespective of conjugacy. Conjugacy only simplifies the computations. If you are outside the conjugate

families, you typically have to resort to numerical methods for calculating posterior moments. Recently many methods have been developed that make this process much easier, including *Gibbs* sampling, *Data Augmentation*, and the *Metropolis-Hastings* algorithm. All three are examples of *M*arkov-Chain-Monte-Carlo or MCMC methods.

The general idea is to construct a chain, or sequence of values, $\theta_0, \theta_1, \ldots$, such that for large k θ_k can be viewed as a draw from the posterior distribution of θ given the data. This is implemented through an algorithm that, given a current value of the parameter vector θ_k, and given the data $X_1, \ldots, X_N$ draws a new value θ_{k+1} from a distribution $f(\cdot)$ indexed by θ_k and the data:

$$\theta_{k+1} \sim f(\theta|\theta_k, \text{data}),$$

in such a way that if the original θ_k came from the posterior distribution, then so does θ_{k+1} (although θ_k and θ_{k+1} in general will not be independent draws)

$$\theta_k|\text{data} \sim p(\theta|\text{data}), \qquad \text{then} \quad \theta_{k+1}|\text{data} \sim p(\theta|\text{data}).$$

Even if we have such an algorithm, the problem is that in principle we would need a starting value θ_0 that such that

$$\theta_0 \sim p(\theta|\text{data}).$$

However, in many cases, irrespective of where we start, that is, irrespective of θ_0, as $k \longrightarrow \infty$, it will be the case that the distribution of the parameter conditional only on the data converges to the posterior distribution:

$$\theta_k|\text{data} \xrightarrow{d} p(\theta|\text{data}),$$

as $k \longrightarrow \infty$.

If that is true, then we can just pick a θ_0, run the chain for a long time, collect the values $\theta_0, \ldots, \theta_K$ for a large value of K, and approximate the posterior distribution by the distribution of $\theta_{K_0}, \ldots, \theta_K$. For example, the mean and standard deviation of the posterior

distribution would be estimated as

$$\hat{\mathbb{E}}[\theta|\text{data}] = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} \theta_k,$$

and

$$\hat{\mathbb{V}}[\theta|\text{data}] = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} \Big(\theta_k - \hat{\mathbb{E}}[\theta|\text{data}]\Big)^2.$$

The first $K_0 - 1$ iterations are discarded to let algorithm converge to the stationary distribution, or "burn in."

4.1 Gibbs Sampling

The idea being the Gibbs sampler is to partition the vector of parameters θ into two (or more) parts, $\theta' = (\theta_1', \theta_2')$. Instead of sampling θ_{k+1} directly from a conditional distribution of

$$f(\theta|\theta_k, \text{data}),$$

we first sample $\theta_{1,k+1}$ from the conditional distribution of

$$p(\theta_1|\theta_{2,k}, \text{data}),$$

and then sample $\theta_{2,k+1}$ from the conditional distribution of

$$p(\theta_2|\theta_{1,k+1}, \text{data}).$$

It is clear that if $(\theta_{1,k}, \theta_{2,k})$ is from the posterior distribution, then so is $(\theta_{1,k}, \theta_{2,k})$.

4.2 Data Augmentation

This is best illustrated with an example. Suppose we are interested in estimating the parameters of a censored regression or Tobit model. There is a latent variable

$$Y_i^* = X_i'\beta + \varepsilon_i,$$

with $\varepsilon_i|X_i \sim \mathcal{N}(0,1)$. (I assume the variance is known for simplicity. This is not essential). We observe

$$Y_i = \max(0, Y_i^*),$$

and the regressors X_i. Suppose the prior distribution for β is normal with some mean μ, and some covariance matrix Ω.

The posterior distribution for β does not have a closed form expression. This is not due to an awkward choice for the prior distribution, because there is no conjugate family for this problem. There is however a simple way of obtaining draws from the posterior distribution using data augmentation in combination with the Gibbs sampler. The first key insight is to view both the vector $\mathbf{Y}^* = (Y_1^*, \ldots, Y_N^*)$ and β as unknown random variables. The Gibbs sampler consists of two steps. First we draw all the missing elements of $\mathbf{Y}^*$ given the current value of the parameter β, say β_k. This involves drawing a series of truncated univariate normal random variables:

$$Y_i^* | \beta, \text{data} \sim \mathcal{TN}\left(X_i'\beta, 1, 0\right),$$

if observation i is truncated, where $\mathcal{TN}(\mu, \sigma^2, c)$ denotes a truncated normal distribution with mean and variance (for the not truncated normal distribution) μ and σ^2, and truncation point c (truncated from above). (Note that we do not need to draw the observed values of Y_i^*.) Second, we draw a new value for the parameter, β_{k+1} given the data and given the (partly drawn) $\mathbf{Y}^*$. The latter is easy given the normal prior distribution: the posterior is normal:

$$p\left(\beta | \text{data}, \mathbf{Y}^*\right) \sim \mathcal{N}\left(\left(\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1} \cdot \left(\mathbf{X}'\mathbf{Y} + \Omega^{-1}\mu\right), \left(\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1}\right).$$

In this example it would still have been feasible to do evaluate the likelihood function directly using numerical integration. Another example where the computational advantages of using data augmentation are even more striking is the multinomial probit model with an unrestricted covariance matrix. See Rossi, Allenby and McCulloch (2005).

4.3 Metropolis-Hastings

We are again interested in $p(\theta|\text{data})$. In this case $p(\theta|\text{data})$ is assumed to be easy to evaluate, but difficult to draw from. Suppose we have a current value θ_k. Then we draw a new candidate value for the chain from a candidate distribution $q(\theta|\theta_k, \text{data})$. This distribution

may (but need not) depend on θ_k. Denote the candidate value by θ. We will either accept the new value, or keep the old value. Then we calculate the ratio

$$r(\theta_k,\theta)=\frac{p(\theta|\text{data})\cdot q(\theta_k|\theta,\text{data})}{p(\theta_k|\text{data})\cdot q(\theta|\theta_k,\text{data})}.$$

The probability that the new draw θ is accepted is

$$\rho(\theta_k,\theta)=\min\left(1,r(\theta_k,\theta)\right),$$

so that

$$\Pr\left(\theta_{k+1}=\theta\right)=\rho(\theta_k,\theta),\qquad\text{and}\quad\Pr\left(\theta_{k+1}=\theta_k\right)=1-\rho(\theta_k,\theta).$$

The optimal choice for the candidate distribution is

$$q^*(\theta|\theta_k,\text{data})=p(\theta|\text{data}),$$

so that $\rho(\theta_k,\theta)=1$ and every draw will get accepted. The trouble is that it is difficult to draw from this distribution. In practice you want to have a relatively dispersed distribution as the candidate distribution, so that the ratio $r(\theta_k,\theta)$ does not get too large.

5. Examples

Here we discuss a number of applications of Bayesian methods. All models contain parameters that are difficult to estimate consistently, and in all cases numerical methods are required to obtain draws from the posterior distribution. The first two are about random coeffiecients. In that case Bernstein-Von Mises would only apply to the individual level parameters if the number of observations per individual would get large.

5.1 Demand Models with Unobserved Heterogeneity in Preferences

Rossi, McCulloch, and Allenby (1996, RMA) are interested in the optimal design of coupon policies. Supermarkets can choose to offer identical coupons for a particular product (tuna cans is the example they use). Alternatively, they may choose to offer differential coupons based on consumer's fixed characteristics. Taking this ever further, they could make the value of the coupon a function of the purchase history of the individual, for example

tailoring the amount of the discount offered in the coupon to the evidence for price sensitivity contained in purchase patterns. RMA set out to estimate the returns to various coupon policies. It is clear that for this investigation to be meaningful one needs to allow for household-level heterogeneity in taste parameters and price elasticities. Even with large amounts of data available, there will be many households for whom these parameters cannot be estimated precisely. RMA therefore use a hieararchical or random coefficients model.

RMA model households choosing the product with the highest utility, where utility for household i, product j, $j = 0, 1, \ldots, J$, at purchase time t is

$$U_{ijt} = X_{it}'\beta_i + \epsilon_{ijt},$$

with the ϵ_{ijt} independent accross households, products and purchase times, and normally distributed with product-specific variances σ_j^2 (and σ_0^2 normalized to one). The X_{it} are observed choice characteristics that in the RMA application include price, some marketing variables, as well as brand dummies. All choice characteristics are assumed to be exogenous, although that assumption may be questioned for the price and marketing variables. Because for some households we have few purchases, it is not possible to accurately estimate all β_i parameters. RMA therefore assume that the household-specific taste parameters are random draws from a normal distribution centered at $Z_i'\Gamma$:

$$\beta_i = Z_i'\Gamma + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \Sigma).$$

Now Gibbs sampling can be used to obtain draws from the posterior distribution of the β_i. To be a little more precise, let us describe the steps in the Gibbs sampler for this example. For more details see RMA. RMA use a normal prior distribution for Γ, a Wishart prior distribution for Σ^{-1}, and inverse Gamma prior distributions for the σ_j^2. To implement the Gibbs sampler, the key is to treat the unobserved utilities as parameters.

The first step is to draw the household parameters β_i given the utilities U_{ijt} and the common parameters Γ, Σ, and σ_j^2. This is straightforward, because we have a standard normal linear model for the utilities, with a normal prior distribution for β_i with parameters

$Z_i'\Gamma$ and variance Σ, and T_i observations. We can draw from this posterior distribution for each household i.

In the second step we draw the σ_j^2 using the results for the normal distribution with known mean and unknown variance.

The third step is to draw from the posterior of Γ and Σ, given the β_i. This again is just a normal linear model, now with unknown mean and unknown variance.

The fourth step is to draw the unobserved utilities given the β_i and the data. Doing this one household/choice at a time, conditioning on the utilities for the other choices, this merely involves drawing from a truncated normal distribution, which is simple and fast.

For some households, those with many recorded purchases and sufficient variation in product characteristics, the posterior distribution will be tight, whereas for others there may be little information in the data and the posterior distribution, conditional on the data as well as Γ and Σ, will essentially be the prior distribution for β_i, which is $\mathcal{N}(Z_i'\Gamma, \Sigma)$.

To think about optimal coupon policies given a particular information set it is useful to think first about the posterior distribution of the household specific parameters β_i. If a supermarket had full information about the household parameters β_i, there would be no additional value in the household characteristics or the purchase history. When we therefore compare a blanket coupon policy (where every household would receive a coupon with the same value) with one that depends on a larger information set that household demographics, or one that also includes purchase histories, the key question is how much precision the information adds about the household level parameters. Specifically, how does the marginal distribution of the household parameters compare with the conditional distribution given purchase histories or given demographics. To make this specific, suppose that the there is only one choice characteristic, price, with household parameter β_i.

The starting point is the case with no household information whatsoever. We can simulate draws from this distribution by drawing from the conditional distribution of β_i given the data for randomly selected households. In the second case we allow conditioning on the household

demographic characteristics Z_i. This leads to less dispersed posterior distributions for the price coefficients. In the third case we also condition on purchase histories. Figure 1, taken from RMA shows for ten households the boxplots of the posterior distribution of the price coefficient under these information sets, one can see the increased precision that results from conditioning on the purchase histories.

5.2 Panel Data Models with Multiple Individual Specific Parameters

Chamberlain and Hirano (1999, CH), see also Hirano (2002), are interested in deriving predictive distributions for earnings using longitudinal data. They are particularly interested in allowing for unobserved individual-level heterogeneity in earnings variances. The specific model they use assumes that log earnings Y_{it} follow the process

$$Y_{it} = X_i'\beta + V_{it} + \alpha_i + U_{it}/h_i.$$

The key innovation in the CH study is the individual variation in the conditional variance, captured by h_i. In this specification $X_i'\beta$ is a systematic component of log earnings, similar to that in specifications used in Abowd and Card () (CH actually use a more general non-linear specification, but the simpler one suffices for the points we make here.) The second component in the model, V_{it}, is a first order autoregressive component,

$$V_{it} = \gamma \cdot V_{it-1} + W_{it},$$

where

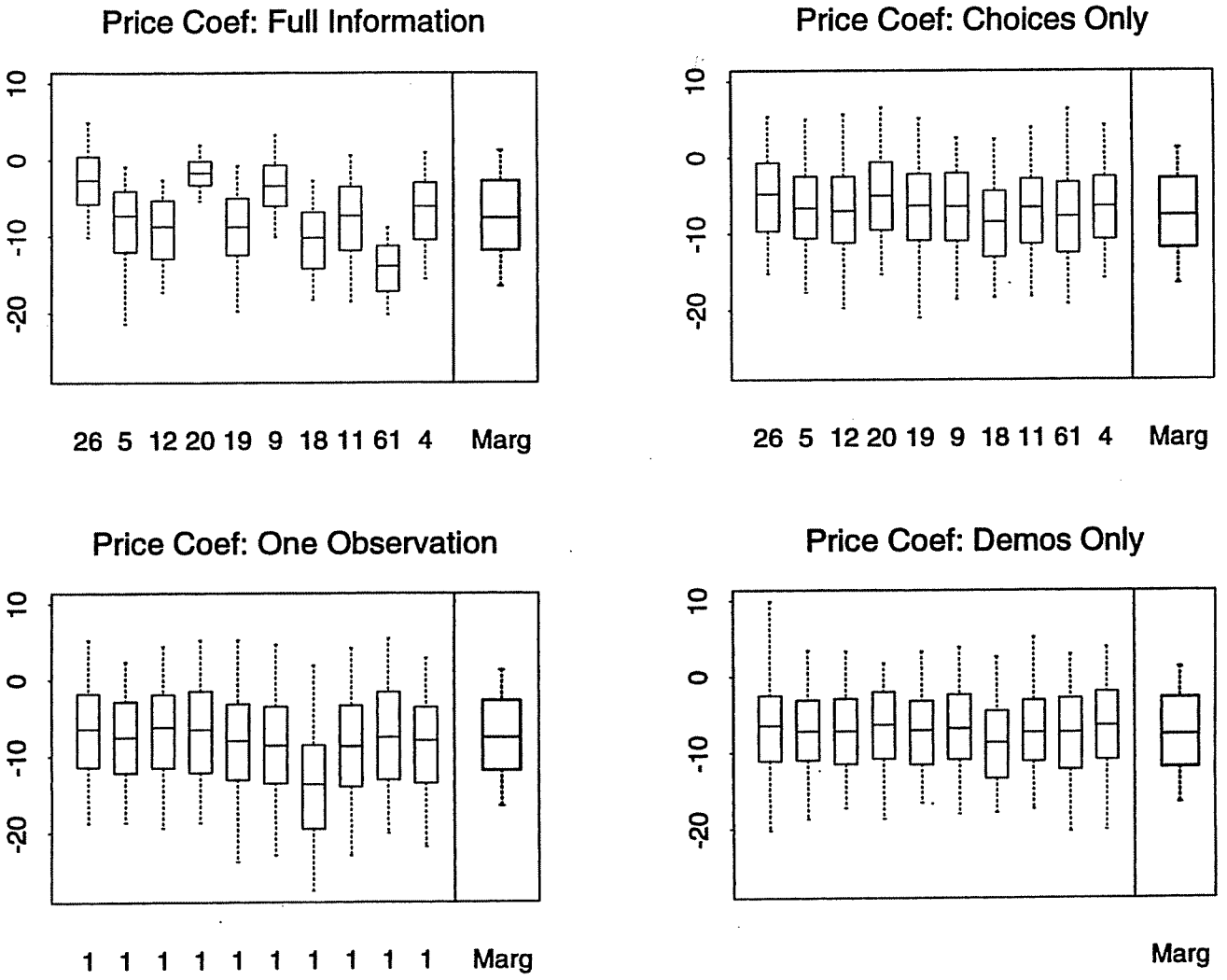
$$V_{i1} \sim \mathcal{N}(0, \sigma_v^2), \qquad W_{it} \sim \mathcal{N}(0, \sigma_w^2).$$

The first factor in the last component has a standard normal distribution,

$$U_{it} \sim \mathcal{N}(0, 1).$$

Analyzing this model by attempting to estimate the α_i and h_i directly would be misguided. From a Bayesian perspective this corresponds to assuming a flat prior distribution on a high-dimensional parameter space.

Figure 2 **Boxplots of posterior distributions of household price coefficients. Various information sets. 10 selected households with the number of purchase occasions indicated along the *X* axis below each boxplot. The boxplot labelled "Marg" is the predictive distribution for a representative household from the model heterogeneity distribution. Note that these are the 11–20th households as ordered in our dataset.**



To avoid such pitfalls CH model α_i and h_i through a random effects specification.

$$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2). \qquad \text{and} \quad h_i \sim \mathcal{G}(m/2, \tau/2).$$

In their empirical application using data from the Panel Study of Income Dynamics (PSID), CH find strong evidence of heterogeneity in conditional variances. Some of this heterogeneity is systematically associated with observed characteristics of the individual such as education, with higher educated individuals experiences lower levels of volatility. Much of the heterogeneity, however, is within groups homogenous in observed characteristics.

The following table, from CH, presents quantiles of the predictive distribution of the conditional standard deviation $1/\sqrt{h_i}$ for different demographic groups: Up to here one

Table 1: QUANTILES OF THE PREDICTIVE DISTRIBUTION OF THE CONDITIONAL STANDARD DEVIATION

	Quantile						
Sample	0.05	0.10	0.25	0.50	0.75	0.90	0.95
All (N=813)	0.04	0.05	0.07	0.11	0.20	0.45	0.81
High School Dropouts (N=37)	0.06	0.08	0.11	0.16	0.27	0.49	0.79
High School Graduates (N=100)	0.04	0.05	0.06	0.11	0.21	0.49	0.93
College Graduates (N=122)	0.03	0.04	0.05	0.09	0.18	0.40	0.75

could have done essentially the same using frequentist methods. One could estimate first the common parameters of the model, β, σ_v^2, σ_w^2, m, τ, and σ_α^2 by maximum likelihood given the specification of the model. Conditional on the covariates one could for each demographic group write the quantiles of the conditional standard deviation in terms of these parameters and obtain point estimates for them.

However, CH wish to go beyond this and infer individual-level predictive distributions for earnings. Taking a particular individual, one can derive the posterior distribution of α_i, h_i, β, σ_v^2, and σ_w^2, given that individual's earnings as well as other earnings, and predict future earnings. To illustrate this CH report earnings predictions for a number of individuals.

Taking two of their observations, one an individual with a sample standard deviation of log earnings of 0.07 and one an individual with a sample standard deviation of 0.47, they report the difference between the 0.90 and 0.10 quantile for the log earnings distribution for these individuals 1 and 5 years into the future.

Table 2:

individual	sample std	0.90-0.10 quantile 1 year out	0.90-0.10 quantile 5 years out
321	0.07	0.32	0.60
415	0.47	1.29	1.29

The variation reported in the CH results may have substantial importance for variation in optimal savings behavior by individuals.

5.3 Instrumental Variables with Many Instruments

In Chamberlain and Imbens (1995, CI) analyze the many instrument problem from a Bayesian perspective. CI use the reduced form for years of education,

$$X_i = \pi_0 + Z_i'\pi_1 + \eta_i,$$

combined with a linear specification for log earnings,

$$Y_i = \alpha + \beta \cdot Z_i'\pi_1 + \varepsilon_i.$$

CI assume joint normality for the reduced form errors,

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}(0, \Omega).$$

This gives a likelihood function

$$\mathcal{L}(\beta, \alpha, \pi_0, \pi_1, \Omega | \text{data}).$$

The focus of the CI paper is on inference for β, and the sensitivity of such inferences to the choice of prior distribution in settings with large numbers of instruments. In that case the dimension of the parameter space is high. Hence a flat prior distribution may in fact be a poor choice. One way to illustrate see this is that a flat prior on π_1 leads to a prior on the sum $\sum_{k=1}^{K} \pi_{ik}^2$ that puts most probability mass away from zero. If in fact the concern is that collectively, the instruments are all weak, one should allow for this possibility in the prior distribution. To be specific, if the prior distribution for the π_{1k} is dispersed, say $\mathcal{N}(0, 100^2)$, then the prior distribution for the $\sum_i \pi_{1k}^2$ is 100 times a chi-squared random variable with degrees of freedom equal to K, implying that *a priori* the concentration parameter is known to be large.

CI then show that the posterior distribution for β, under a flat prior distribution for π_1 provides an accurate approximation to the sampling distribution of the TSLS estimator, providing both a further illustration of the lack of appeal of TSLS in settings with many instruments, and the unattractiveness of the flat prior distribution.

As an alternative CI suggest a hierarchical prior distribution with

$$\pi_{1k} \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2).$$

In the Angrist-Krueger 1991 compulsory schooling example there is in fact a substantive reason to believe that σ_π^2 is small. If the π_{1k} represent the effect of the differences in the amount of required schooling, one would expect the magnitude of the π_{1k} to be less than the amount of variation in the compulsory schooling. The latter is less than one year. Since any distribution with support on $[0, 1]$ has a variance less than or equal to $1/12$, the standard deviation of the first stage coefficients should not be more than $\sqrt{1/12} = 0.289$. Using the Angrist-Krueger data CI find that the posterior distribution for σ_π is concentrated close to zero, with the posterior mean and median equal to 0.119.

5.4 Binary Response with Endogenous Discrete Regressors

Geweke, Gowrisankaran, and Town (2003, GGT) are interested in estimating the effect of hospital quality on mortality, taking into account possibly non-random selection of patients

into hospitals. Patients can choose from 114 hospitals. Given their observed individual characteristics Z_i, latent mortality is

$$Y_i^* = \sum_{j=1}^{113} C_{ij}\beta_j + Z_i'\gamma + \epsilon_i,$$

where C_{ij} is an indicator for patient i going to hospital j. The focus is on the hospital effects on mortality, β_j. Realized mortality is

$$Y_i = 1\{Y_i^* \geq 0\}.$$

The concern is about selection into the hospitals, and the possibility that this is related to unobserved components of latent mortality GGT model latent the latent utility for patient i associated with hospital j as

$$C_{ij}^* = X_{ij}'\alpha + \eta_{ij},$$

where the X_{ij} are hospital-individual specific characteristics, including distance to hospital. Patient i then chooses hospital j if

$$C_{ij}^* \geq C_{ik}, \quad \text{for } k = 1, \ldots, 114.$$

The endogeneity is modelled through the potential correlation between η_{ij} and ϵ_i. Specifically, GGT asssume that as

$$\epsilon_i = \sum_{j=1}^{113} \eta_{ij} \cdot \delta_j + \zeta_i,$$

where the ζ_i is a standard normal random variable, independent of the other unobserved components. GGT model the η_{ij} as standard normal, independent across hospitals and across individuals. This is a very strong assumption, implying essentially the independence of irrelevant alternatives property. One may wish to relax this by allowing for random coefficients on the hospital characteristics.

Given these modelling decisions GGT have a fully specified joint distribution of hospital choice and mortality given hospital and individual characteristics. The log likelihood

function is highly nonlinear, and it is unlikely it can be well approximated by a quadratic function. GGT therefore use Bayesian methods, and in particular the Gibbs sampler to obtain draws from the posterior distribution of interest. In their empirical analysis GGT find strong evidence for non-random selection. They find that higher quality hospitals attract sicker patients, to the extent that a model based on exogenous selection would have led to misleading conclusions on hospital quality.

5.5 Discrete Choice Models with Unobserved Choice Characteristics

Athey and Imbens (2007, AI) study discrete choice models, allowing both for unobserved individual heterogeneity in taste parameters as well as for multiple unobserved choice characteristics. In such settings the likelihood function is multi-modal, and frequentist approximations based on quadratic approximations to the log likelihood function around the maximum likelihood estimator are unlikely to be accurate. The specific model AI use assumes that the utility for individual i in market t for choice j is

$$U_{ijt} = X_{it}'\beta_i + \xi_j'\gamma_i + \epsilon_{ijt},$$

where X_{it} are market-specific observed choice characteristics, ξ_j is a vector of unobserved choice characteristics, and ϵ_{ijt} is an idiosyncratic error term, independent accross market, choices, and individuals, with a normal distribution centered at zero, and with the variance normalized to unity. The individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Delta Z_i, \Omega),$$

with the Z_i observed individual characteristics.

AI specify a prior distribution on the common parameters, Δ, and Ω, and on the values of the unobserved choice characteristics ξ_j. Using gibbs sampling and data augmentation with the unobserved utilities as unobserved random variables makes sampling from the posterior distribution conceptually straightforward even in cases with more than one unobserved choice characteristic. In contrast, earlier studies using multiple unobserved choice characteristics

(Elrod and Keane, 1995; Goettler and Shachar, 2001), using frequentist methods, faced much heavier computational burdens.

References

Athey, S., and G. Imbens, (2007), "Discrete Choice Models with Multiple Unobserved Product Characteristics," *International Economic Review*, forthcoming.

Box, G., and G. Tiao, (1973), *Bayesian Inference in Statistical Analysis,* Wiley, NY.

Chamberlain, G., and K. Hirano, (1996), "Hirearchical Bayes Models with Many Instrumental Variables," NBER Technical Working Paper 204.

Chamberlain, G., and G. Imbens, (1999), "Predictive Distributions based on Longitudinal Eearnings Data," *Annales d'Economie et de Statistique*, 55-56, 211-242.

Elrod, T., and M. Keane, (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data, "*Journal of Marketing Research*, Vol. XXXII, 1-16.

Ferguson, T., (1996), *A Course in Large Sample Theory*, Chapman and Hall, new York, NY.

Gelman, A., J. Carlin, H. Stenr, and D. Rubin, (2004), *Bayesian Data Analysis,* Chapman and Hall, New York, NY.

Gelman, A., and J. Hill, (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Cambridge University Press.

Geweke, J., G. Gowrisankaran, and R. Town, (2003), "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 71(4), 1215-1238.

Geweke, J., (1997), "Posterior Simulations in Econometrics," in *Advances in Economics and Econometrics: Theory and Applications*, Vol III, Kreps and Wallis (eds.), Cambridge University Press.

Gilks, W. S. Richardson and D. Spiegelhalter, (1996), *Markvo Chain Monte Carlo in Practice,* Chapman and Hall, New York, NY.

Goettler, J., and R. Shachar (2001), "Spatial Competition in the Network Televi-

sion Industry," *RAND Journal of Economics*, Vol. 32(4), 624-656.

Lancaster, T., (2004), *An Introduction to Modern Bayesian Econometrics,* Blackwell Publishing, Malden, MA.

Rossi, P., R. McCulloch, and G. Allenby, (1996), "The Value of Purchasing History Data in Target Marketing," *Marketing Science*, Vol 15(4), 321-340.

Rossi, P., G. Allbeny, and R. McCulloch, (2005), *Bayesian Statistics and Marketing*, Wiley, Hoboken, NJ.

Sims, C., and H. Uhlig, (1991), "Understanding Unit Rotters: A Helicopter View," *Econometrica*, 59(6), 1591-1599.

What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 8, **Tuesday**, **July 31st**, **2.00-3.00 pm**

Cluster and Stratified Sampling

These notes consider estimation and inference with cluster samples and samples obtained by stratifying the population. The main focus is on true cluster samples, although the case of applying cluster-sample methods to panel data is treated, including recent work where the sizes of the cross section and time series are similar. Wooldridge (2003, extended version 2006) contains a survey, but some recent work is discussed here.

1. THE LINEAR MODEL WITH CLUSTER EFFECTS

This section considers linear models estimated using cluster samples (of which a panel data set is a special case). For each group or cluster g, let $\{(y_{gm}, x_g, z_{gm}) : m = 1, \ldots, M_g\}$ be the observable data, where M_g is the number of units in cluster g, y_{gm} is a scalar response, x_g is a $1 \times K$ vector containing explanatory variables that vary only at the group level, and z_{gm} is a $1 \times L$ vector of covariates that vary within (as well as across) groups.

1.1 Specification of the Model

The linear model with an additive error is

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma + v_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.1}$$

Our approach to estimation and inference in equation (1.1) depends on several factors, including whether we are interested in the effects of aggregate variables (β) or individual-specific variables (γ). Plus, we need to make assumptions about the error terms. In the context of pure cluster sampling, an important issue is whether the v_{gm} contain a common group effect that can be separated in an additive fashion, as in

$$v_{gm} = c_g + u_{gm}, m = 1, \ldots, M_g, \tag{1.2}$$

where c_g is an unobserved cluster effect and u_{gm} is the idiosyncratic error. (In the statistics literature, (1.1) and (1.2) are referred to as a "hierarchical linear model.") One important issue is whether the explanatory variables in (1.1) can be taken to be appropriately exogenous. Under (1.2), exogeneity issues are usefully broken down by separately considering c_g and u_{gm}.

Throughout we assume that the sampling scheme generates observations that are independent across g. This assumption can be restrictive, particularly when the clusters are large geographical units. We do not consider problems of "spatial correlation" across clusters, although, as we will see, fixed effects estimators have advantages in such settings.

We treat two kinds of sampling schemes. The simplest case also allows the most flexibility

for robust inference: from a large population of relatively small clusters, we draw a large number of clusters (G), where cluster g has M_g members. This setup is appropriate, for example, in randomly sampling a large number of families, classrooms, or firms from a large population. The key feature is that the number of groups is large enough relative to the group sizes so that we can allow essentially unrestricted within-cluster correlation. Randomly sampling a large number of clusters also applies to many panel data sets, where the cross-sectional population size is large (say, individuals, firms, even cities or counties) and the number of time periods is relatively small. In the panel data setting, G is the number of cross-sectional units and M_g is the number of time periods for unit g.

A different sampling scheme results in data sets that also can be arranged by group, but is better interpreted in the context of sampling from different populations are different strata within a population. We stratify the population into into $G \geq 2$ nonoverlapping groups. Then, we obtain a random sample of size M_g from each group. Ideally, the group sizes are large in the population, hopefully resulting in large M_g. This is the perspective for the "small G" case in Section 1.3.

1.2. Large Group Asymptotics

In this section I review methods and estimators justified when the asymptotic approximations theory is with The theory with $G \to \infty$ and the group sizes, M_g, fixed is well developed; see, for example, White (1984), Arellano (1987), and Wooldridge (2002, Chapters 10, 11). Here, the emphasis is on how one might wish to use methods robust to cluster sampling even when it is not so obvious.

First suppose that the covariates satisfy

$$E(v_{gm}|x_g, z_{gm}) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.3}$$

For consistency, we can, of course, get by with zero correlation assumptions, but we use (1.3) for convenience because it meshes well with assumptions concerning conditional second moments. Importantly, the exogeneity in (1.3) only requires that z_{gm} and v_{gm} are uncorrelated. In particular, it does not specify assumptions concerning v_{gm} and z_{gp} for $m \neq p$. As we saw in the linear panel data notes, (1.3) is called the "contemporaneous exogeneity" assumption when m represents time. Allowing for correlation between v_{gm} and $z_{gp}, m \neq p$ is useful for some panel data applications and possibly even cluster samples (if the covariates of one unit can affect another unit's response). Under (1.3) and a standard rank condition on the covariates, the pooled OLS estimator, where we regress y_{gm} on $1, x_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G,$ is

consistent for $\lambda \equiv (\alpha, \beta', \gamma')'$ (as $G \to \infty$ with M_g fixed) and $\sqrt{G}$-asymptotically normal.

Without more assumptions, a robust variance matrix is needed to account for correlation within clusters or heteroskedasticity in $Var(v_{gm}|x_g, z_{gm})$, or both. When v_{gm} has the form in (1.2), the amount of within-cluster correlation can be substantial, which means the usual OLS standard errors can be very misleading (and, in most cases, systematically too small). Write W_g as the $M_g \times (1 + K + L)$ matrix of all regressors for group g. Then the $(1 + K + L) \times (1 + K + L)$ variance matrix estimator is

$$\widehat{Avar}(\hat{\lambda}_{POLS}) = \left(\sum_{g=1}^{G} W_g' W_g\right)^{-1} \left(\sum_{g=1}^{G} W_g' \hat{v}_g \hat{v}_g' W_g\right) \left(\sum_{g=1}^{G} W_g' W_g\right)^{-1} \tag{1.4}$$

where $\hat{v}_g$ is the $M_g \times 1$ vector of pooled OLS residuals for group g. This asymptotic variance is now computed routinely using "cluster" options.

Pooled OLS estimation of the parameters in (1.1) ignores the within-cluster correlation of the v_{gm}; even if the procedure is consistent (again, with $G \to \infty$ and the M_g fixed), the POLS estimators can be very inefficient. If we strengthen the exogeneity assumption to

$$E(v_{gm}|x_g, Z_g) = 0, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.5}$$

where Z_g is the $M_g \times L$ matrix of unit-specific covariates, then we can exploit the presence of c_g in (1.2) in a generalized least squares (GLS) analysis. With true cluster samples, (1.5) rules out the covariates from one member of the cluster affecting the outcomes on another, holding own covariates fixed. In the panel data case, (1.5) is the strict exogeneity assumption on $\{z_{gm} : m = 1, \ldots, M_g\}$ that we discussed in the linear panel data notes The standard random effects approach makes enough assumptions so that the $M_g \times M_g$ variance-covariance matrix of $v_g = (v_{g1}, v_{g2}, \ldots, v_{g,M_g})'$ has the so-called "random effects" form,

$$Var(v_g) = \sigma_c^2 j_{M_g}' j_{M_g} + \sigma_u^2 I_{M_g}, \tag{1.6}$$

where j_{M_g} is the $M_g \times 1$ vector of ones and I_{M_g} is the $M_g \times M_g$ identity matrix. In the standard setup, we also make the "system homoskedasticity" assumption,

$$Var(v_g|x_g, Z_g) = Var(v_g). \tag{1.7}$$

It is important to understand the role of assumption (1,7): it implies that the conditional variance-covariance matrix is the same as the unconditional variance-covariance matrix, but it does not restrict $Var(v_g)$; it can be any $M_g \times M_g$ matrix under (1.7). The particular random effects structure on $Var(v_g)$ is given by (1.6). Under (1.6) and (1.7), the resulting GLS

estimator is the well-known random effects (RE) estimator.

The random effects estimator $\hat{\lambda}_{RE}$ is asymptotically more efficient than pooled OLS under (1.5), (1.6), and (1.7) as $G \to \infty$ with the M_g fixed. The RE estimates and test statistics are computed routinely by popular software packages. Nevertheless, an important point is often overlooked in applications of RE: one can, and in many cases should, make inference completely robust to an unknown form of $Var(v_g|x_g, Z_g)$.

The idea in obtaining a fully robust variance matrix of RE is straightforward and we essentially discussed it in the notes on nonlinear panel data models. Even if $Var(v_g|x_g, Z_g)$ does not have the RE form, the RE estimator is still consistent and $\sqrt{G}$-asymptotically normal under (1.5), and it is likely to be more efficient than pooled OLS. Yet we should recognize that the RE second moment assumptions can be violated without causing inconsistency in the RE estimator. For panel data applications, making inference robust to serial correlation in the idiosyncratic errors, especially with more than a few time periods, can be very important. Further, within-group correlation in the idiosyncratic errors can arise for cluster samples, too, especially if underlying (1.1) is a random coefficient model,

$$y_{gm} = \alpha + x_g\beta + z_{gm}\gamma_g + v_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{1.8}$$

By estimating a standard random effects model that assumes common slopes γ, we effectively include $z_{gm}(\gamma_g - \gamma)$ in the idiosyncratic error; this generally creates within-group correlation because $z_{gm}(\gamma_g - \gamma)$ and $z_{gp}(\gamma_g - \gamma)$ will be correlated for $m \neq p$, conditional on Z_g. Also, the idiosyncratic error will have heteroskedasticity that is a function of z_{gm}. Nevertheless, if we assume $E(\gamma_g|X_g, Z_g) = E(\gamma_g) \equiv \gamma$ along with (1.5), the random effects estimator still consistently estimates the average slopes, γ. Therefore, in applying random effects to panel data or cluster samples, it is sensible (with large *G*) to make the variance estimator of random effects robust to arbitrary heteroskedasticity and within-group correlation.

One way to see what the robust variance matrix looks like for $\hat{\lambda}_{RE}$ is to use the pooled OLS characterization of RE on a transformed set of data. For each *g*, define $\hat{\theta}_g = 1 - \{1/[1 + M_g(\hat{\sigma}_c^2/\hat{\sigma}_u^2)]\}^{1/2}$, where $\hat{\sigma}_c^2$ and $\hat{\sigma}_u^2$ are estimators of the variances of c_g and u_{gm}, respectively. Then the RE estimator is identical to the pooled OLS estimator of

$$y_{gm} - \hat{\theta}_g\bar{y}_g \text{ on } (1 - \hat{\theta}_g), (1 - \hat{\theta}_g)x_g, z_{gm} - \hat{\theta}_g\bar{z}_g, m = 1, \ldots, M_g; g = 1, \ldots, G; \tag{1.9}$$

see, for example, Hsiao (2003). For fully robust inference, we can just apply the fully robust variance matrix estimator in (1.4) but on the transformed data.

With panel data, it may make sense to estimate an unrestricted version of $Var(v_g)$, especially if G is large. Even in that case, it makes sense to obtain a variance matrix robust to $Var(v_{gm}|x_g, Z_g) \neq Var(v_g)$, as the GEE literature does. One can also specify a particular structure, such as an AR(1) model for the idiosyncratic errors. In any case, fully robust inference is still a good idea.

In summary, with large G and relatively small M_g, it makes sense to compute fully robust variance estimators even if we apply a GLS procedure that allows $Var(v_g)$ to be unrestricted. Nothing ever guarantees $Var(v_{gm}|x_g, Z_g) = Var(v_g)$. Because RE imposes a specific structure on $Var(v_g)$, there is a strong case for making RE inference fully robust. When c_g is in the error term, it is even more critical to use robust inference when using pooled OLS because the usual standard errors ignore within-cluster correlation entirely.

If we are only interested in estimating γ, the "fixed effects" (FE) or "within" estimator is attractive. The within transformation subtracts off group averages from the dependent variable and explanatory variables:

$$y_{gm} - \bar{y}_g = (z_{gm} - \bar{z}_g)\gamma + u_{gm} - \bar{u}_g, m = 1, \ldots, M_g; g = 1, \ldots, G, \tag{1.10}$$

and this equation is estimated by pooled OLS. (Of course, the x_g get swept away by the within-group demeaning.) Under a full set of "fixed effects" assumptions – which, unlike pooled OLS and random effects, allows arbitrary correlation between c_g and the z_{gm} – inference is straightforward using standard software. Nevertheless, analogous to the random effects case, it is often important to allow $Var(u_g|Z_g)$ to have an arbitrary form, including within-group correlation and heteroskedasticity. For panel data, the idiosyncratic errors can always have serial correlation or heteroskedasticity, and it is easy to guard against these problems in inference. Reasons for wanting a fully robust variance matrix estimator for FE applied to cluster samples are similar to the RE case. For example, if we start with the model (1.8) then $(z_{gm} - \bar{z}_g)(\gamma_g - \gamma)$ appears in the error term. As we discussed in the linear panel data notes, the FE estimator is still consistent if $E(\gamma_g|z_{g1} - \bar{z}_g, \ldots, z_{g,M_g} - \bar{z}_g) = E(\gamma_g) = \gamma$, an assumption that allows γ_g to be correlated with $\bar{z}_g$. Nevertheless, u_{gm}, u_{gp} will be correlated for $m \neq p$. A fully robust variance matrix estimator is

$$\widehat{Avar}(\hat{\gamma}_{FE}) = \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g\right)^{-1} \left(\sum_{g=1}^{G} \ddot{Z}_g' \hat{u}_g \hat{u}_g' \ddot{Z}_g\right) \left(\sum_{g=1}^{G} \ddot{Z}_g' \ddot{Z}_g\right)^{-1}, \tag{1.11}$$

where $\ddot{Z}_g$ is the matrix of within-group deviations from means and $\hat{u}_g$ is the $M_g \times 1$ vector of

fixed effects residuals. This estimator is justified with large-G asymptotics.

One benefit of a fixed effects approach, especially in the standard model with constant slopes but c_g in the composite error term, is that no adjustments are necessary if the c_g are correlated across groups. When the groups represent different geographical units, we might expect correlation across groups close to each other. If we think such correlation is largely captured through the unobserved effect c_g, then its elimination via the within transformation effectively solves the problem. If we use pooled OLS or a random effects approach, we would have to deal with spatial correlation across g, in addition to within-group correlation, and this is a difficult problem.

The previous discussion extends immediately to instrumental variables versions of all estimators. With large G, one can afford to make pooled two stage least squares (2SLS), random effects 2SLS, and fixed effects 2SLS robust to arbitrary within-cluster correlation and heteroskedasticity. Also, more efficient estimation is possible by applying generalized method of moments (GMM); again, GMM is justified with large G.

1.3. Should we Use the "Large" G Formulas with "Large" M_g?

Until recently, the standard errors and test statistics obtained from pooled OLS, random effects, and fixed effects were known to be valid only as $G \to \infty$ with each M_g fixed. As a practical matter, that means one should have lots of small groups. Consider again formula (1.4), for pooled OLS, when the cluster effect, c_g, is left in the error term. With a large number of groups and small group sizes, we can get good estimates of the within-cluster correlations – technically, of the cluster correlations of the cross products of the regressors and errors – even if they are unrestricted, and that is why the robust variance matrix is consistent as $G \to \infty$ with M_g fixed. In fact, in this scenario, one loses nothing in terms of asymptotic local power (with local alternatives shrinking to zero at the rate $G^{-1/2}$) if c_g is not present. In other words, based on first-order asymptotic analysis, there is no cost to being fully robust to any kind of within-group correlation or heteroskedasticity. These arguments apply equally to panel data sets with a large number of cross sections and relatively few time periods, whether or not the idiosyncratic errors are serially correlated.

What if one applies robust inference in scenarios where the fixed M_g, $G \to \infty$ asymptotic analysis not realistic? Hansen (2007) has recently derived properties of the cluster-robust variance matrix and related test statistics under various scenarios that help us more fully understand the properties of cluster robust inference across different data configurations. First

consider how his results apply to true cluster samples. Hansen (2007, Theorem 2) shows that, with G and M_g both getting large, the usual inference based on (1.4) is valid with arbitrary correlation among the errors, v_{gm}, within each group. Because we usually think of v_{gm} as including the group effect c_g, this means that, with large group sizes, we can obtain valid inference using the cluster-robust variance matrix, provided that G is also large. So, for example, if we have a sample of $G = 100$ schools and roughly $M_g = 100$ students per school, and we use pooled OLS leaving the school effects in the error term, we should expect the inference to have roughly the correct size. Probably we leave the school effects in the error term because we are interested in a school-specific explanatory variable, perhaps indicating a policy change.

Unfortunately, pooled OLS with cluster effects when G is small and group sizes are large fall outside Hansen's theoretical findings: the proper asymptotic analysis would be with G fixed, $M_g \to \infty$, and persistent within-cluster correlation (because of the presence of c_g in the error). Hansen (2007, Theorem 4) is aimed at panel data where the time series dependence satisfies strong mixing assumptions, that is, where the correlation within each group g is weakly dependent. Even in this case, the variance matrix in (1.4) must be multiplied by $G/(G-1)$ and inference based on the t_{G-1} distribution. (Conveniently, this adjustment is standard in Stata's calculation of cluster-robust variance matrices.) Interestingly, Hansen finds, in simulations, that with $G = 10$ and $M_g = 50$ for all g, using the adjusted robust variance matrix estimator with critical values from the t_{G-1} distribution produces fairly small size distortions. But the simulation study is special (one covariate whose variance is as large as the variance of the composite error).

We probably should not expect good properties of the cluster-robust inference with small groups and very large group sizes when cluster effects are left in the error term. As an example, suppose that $G = 10$ hospitals have been sampled with several hundred patients per hospital. If the explanatory variable of interest is exogenous and varies only at the hospital level, it is tempting to use pooled OLS with cluster-robust inference. But we have no theoretical justification for doing so, and reasons to expect it will not work well. In the next section we discuss other approaches available with small G and large M_g.

If the explanatory variables of interest vary within group, FE is attractive for a couple of reasons. The first advantage is the usal one about allowing c_g to be arbitrarily correlated with the z_{gm}. The second advantage is that, with large M_g, we can treat the c_g as parameters to

estimate – because we can estimate them precisely – and then assume that the observations are independent across m (as well as g). This means that the usual inference is valid, perhaps with adjustment for heteroskedasticity. Interestingly, the fixed G, large M_g asymptotic results in Theorem 4 of Hansen (2007) for cluster-robust inference apply in this case. But using cluster-robust inference is likely to be very costly in this situation: the cluster-robust variance matrix actually converges to a random variable, and t statistics based on the adjusted version of (1.11) – that is, multiplied by $G/(G-1)$ – have an asymptotic t_{G-1} distribution. Therefore, while the usual or heteroskedasticity-robust inference can be based on the standard normal distribution, the cluster-robust inference is based on the t_{G-1} distribution (and the cluster-robust standard errors may be larger than the usual standard errors). With small G, inference based on cluster-robust statistics could be very conservative when it need not be. (Also, Hansen's Theorem 4 is not completely general, and may not apply with heterogeneous sampling across groups.)

In summary, for true cluster sample applications, cluster-robust inference using pooled OLS delivers statistics with proper size when G and M_g are both moderately large, but they should probably be avoided with large M_g and small G. When cluster fixed effects are included, the usual inference is often valid, perhaps made robust to heteroskedasticity, and is likely to be much more powerful than cluster-robust inference.

For panel data applications, Hansen's (2007) results, particularly Theorem 3, imply that cluster-robust inference for the fixed effects estimator should work well when the cross section (N) and time series (T) dimensions are similar and not too small. If full time effects are allowed in addition to unit-specific fixed effects – as they often should – then the asymptotics must be with N and T both getting large. In this case, any serial dependence in the idiosyncratic errors is assumed to be weakly dependent. The simililations in Bertrand, Duflo, and Mullainathan (2004) and Hansen (2007) verify that the fully robust cluster-robust variance matrix works well.

There is some scope for applying the fully robust variance matrix estimator when N is small relative to T when unit-specific fixed effects are included. Unlike in the true cluster sampling case, it makes sense to treat the idiosyncratic errors as correlated with only weakly dependent. But Hansen's (2007, Theorem 4) does not allow time fixed effects (because the asymptotics is with fixed N and $T \to \infty$, and so the inclusion of time fixed effects means adding more and more parameters without more cross section data to estimate them). As a practical

matter, it seems dangerous to rely on omitting time effects or unit effects with panel data. Hansen's result that applies in this case requires *N* and *T* both getting large.

2. Estimation with a Small Number of Groups and Large Group Sizes

We can summarize the findings of the previous section as follows: fully robust inference justified for large *G* (*N*) and small M_g (*T*) can also be relied on when M_g (*T*) is also large, provided *G* (*N*) is also reasonably large. However, whether or not we leave cluster (unobserved) effects in the error term, there are good reasons not to rely on cluster-robust inference when *G* (*N*) is small and M_g (*T*) is large.

In this section, we describe approaches to inference when *G* is small and the M_g are large. These results apply primarily to the true cluster sample case, although we will draw on them for difference-in-differences frameworks using pooled cross sections in a later set of notes.

In the large *G*, small M_g case, it often makes sense to think of sampling a large number of groups from a large population of clusters, where each cluster is relatively small. When *G* is small while each M_g is large, this thought experiment needs to be modified. For most data sets with small *G*, a stratified sampling scheme makes more sense: we have defined *G* groups in the population, and we obtain our data by randomly sampling from each group. As before, M_g is the sample size for group *g*. Except for the relative dimensions of *G* and M_g, the resulting data set is essentially indistinguishable from that described in Section 1.2.

The problem of proper inference when M_g is large relative to *G* was brought to light by Moulton (1990), and has been recently studied by Donald and Lang (2007). DL focus on applications that seem well described by the stratified sampling scheme, but their approach seems to imply a different sampling experiment. In particular, they treat the parameters associated with the different groups as outcomes of random draws. One way to think about the sampling in this case is that a small number of groups is drawn from a (large) population of potential groups; therefore, the parameters common to all members of the group can be viewed as random. Given the groups, samples are then obtained via random sampling within each group.

To illustrate the issues considered by Donald and Lang, consider the simplest case, with a single regressor that varies only by group:

$$y_{gm} = \alpha + \beta x_g + c_g + u_{gm} \tag{2.1}$$

$$= \delta_g + \beta x_g + u_{gm}, \; m = 1, \ldots, M_g; g = 1, \ldots, G. \tag{2.2}$$

Notice how (2.2) is written as a model with common slope, β, but intercept, δ_g, that varies across g. Donald and Lang focus on (2.1), where c_g is assumed to be independent of x_g with zero mean. They use this formulation to highlight the problems of applying standard inference to (2.1), leaving c_g as part of the composite error term, $v_{gm} = c_g + u_{gm}$. We know this is a bad idea even in the large G, small M_g case, as it ignores the persistent correlation in the errors within each group. Further, from the discussion of Hansen's (2007) results, using cluster-robust inference when G is small is likely to produce poor inference.

One way to see the problem with the usual inference in applying standard inference is to note that when $M_g = M$ for all $g = 1, \ldots, G$, the pooled OLS estimator, $\hat{\beta}$, is identical to the "between" estimator obtained from the regression

$$\bar{y}_g \text{ on } 1, x_g, g = 1, \ldots, G. \tag{2.3}$$

Conditional on the x_g, $\hat{\beta}$ inherits its distribution from $\{\bar{v}_g : g = 1, \ldots, G\}$, the within-group averages of the composite errors $v_{gm} \equiv c_g + u_{gm}$. The presence of c_g means new observations within group do not provide additional information for estimating β beyond how they affect the group average, $\bar{y}_g$. In effect, we only have G useful pieces of information.

If we add some strong assumptions, there is a solution to the inference problem. In addition to assuming $M_g = M$ for all g, assume $c_g | x_g \sim \text{Normal}(0, \sigma_c^2)$ and assume $u_{gm} | x_g, c_g \sim \text{Normal}(0, \sigma_u^2)$. Then $\bar{v}_g$ is independent of x_g and $\bar{v}_g \sim \text{Normal}(0, \sigma_c^2 + \sigma_u^2/M)$ for all g. Because we assume independence across g, the equation

$$\bar{y}_g = \alpha + \beta x_g + \bar{v}_g, g = 1, \ldots, G \tag{2.4}$$

satisfies the classical linear model assumptions. Therefore, we can use inference based on the t_{G-2} distribution to test hypotheses about β, provided $G > 2$. When G is very small, the requirements for a significant t statistic using the t_{G-2} distribution are much more stringent then if we use the $t_{M_1+M_2+\ldots+M_G-2}$ distribution – which is what we would be doing if we use the usual pooled OLS statistics. (Interestingly, if we use cluster-robust inference and apply Hansen's results – even though they do not apply – we would use the t_{G-1} distribution.)

When x_g is a $1 \times K$ vector, we need $G > K + 1$ to use the t_{G-K-1} distribution for inference. [In Moulton (1990), $G = 50$ states and x_g contains 17 elements]

As pointed out by DL, performing the correct inference in the presence of c_g is *not* just a matter of correcting the pooled OLS standard errors for cluster correlation – something that does not appear to be valid for small G, anyway – or using the RE estimator. In the case of

common group sizes, there is only estimator: pooled OLS, random effects, and the between regression in (2.4) all lead to the *same* $\hat{\beta}$. The regression in (2.4), by using the t_{G-K-1} distribution, yields inference with appropriate size.

We can apply the DL method without normality of the u_{gm} if the common group size M is large: by the central limit theorem, $\bar{u}_g$ will be approximately normally distributed very generally. Then, because c_g is normally distributed, we can treat $\bar{v}_g$ as approximately normal with constant variance. Further, even if the group sizes differ across g, for very large group sizes $\bar{u}_g$ will be a negligible part of $\bar{v}_g$: $Var(\bar{v}_g) = \sigma_c^2 + \sigma_u^2/M_g$. Provided c_g is normally distributed and it dominates $\bar{v}_g$, a classical linear model analysis on (2.4) should be roughly valid.

The broadest applicability of DL's setup is when the average of the idiosyncratic errors, $\bar{u}_g$, can be ignored – either because σ_u^2 is small relative to σ_c^2, M_g is large, or both. In fact, applying DL with different group sizes or nonnormality of the u_{gm} is identical to ignoring the estimation error in the sample averages, $\bar{y}_g$. In other words, it is as if we are analyzing the simple regression $\mu_g = \alpha + \beta x_g + c_g$ using the classical linear model assumptions (where $\bar{y}_g$ is used in place of the unknown group mean, μ_g). With small G, we need to further assume c_g is normally distributed.

If z_{gm} appears in the model, then we can use the averaged equation

$$\bar{y}_g = \alpha + x_g\beta + \bar{z}_g\gamma + \bar{v}_g, g = 1, \ldots, G, \tag{2.5}$$

provided $G > K + L + 1$. If c_g is independent of $(x_g, \bar{z}_g)$ with a homoskedastic normal distribution and the group sizes are large, inference can be carried out using the $t_{G-K-L-1}$ distribution.

The DL solution to the inference problem with small G is pretty common as a strategy to check robustness of results obtained from cluster samples, but often it is implemented with somewhat large G (say, $G = 50$). Often with cluster samples one estimates the parameters using the disaggregated data and also the averaged data. When some covariates that vary within cluster, using averaged data is generally inefficient. But it does mean that standard errors need not be made robust to within-cluster correlation. We now know that if G is reasonably large and the group sizes not too large, the cluster-robust inference can be acceptable. DL point out that with small G one should think about simply using the group averages in a classical linear model analysis.

For small G and large M_g, inference obtained from analyzing (2.5) as a classical linear

model will be very conservative in the absense of a cluster effect. Perhaps in some cases it is desirable to inject this kind of uncertainty, but it rules out some widely-used staples of policy analysis. For example, suppose we have two populations (maybe men and women, two different cities, or a treatment and a control group) with means $\mu_g, g = 1,2$, and we would like to obtain a confidence interval for their difference. In almost all cases, it makes sense to view the data as being two random samples, one from each subgroup of the population. Under random sampling from each group, and assuming normality and equal population variances, the usual comparison-of-means statistic is distributed exactly as $t_{M_1+M_2-2}$ under the null hypothesis of equal population means. (Or, we can construct an exact 95% confidence interval of the difference in population means.) With even moderate sizes for M_1 and M_2, the $t_{M_1+M_2-2}$ distribution is close to the standard normal distribution. Plus, we can relax normality to obtain approximately valid inference, and it is easy to adjust the *t* statistic to allow for different population variances. With a controlled experiment the standard difference-in-means analysis is often quite convincing. Yet we cannot even study this estimator in the DL setup because $G = 2$. The problem can be seen from (2.2): in effect, we have three parameters, δ_1, δ_2, and β, but only two observations.

DL criticize Card and Krueger (1994) for comparing mean wage changes of fast-food workers across two states because Card and Krueger fail to account for the state effect (New Jersery or Pennsylvania), c_g, in the composite error, v_{gm}. But the DL criticism in the $G = 2$ case is no different from a common question raised for any difference-in-differences analyses: How can we be sure that any observed difference in means is due entirely to the policy change? To characterize the problem as failing to account for an unobserved group effect is not necessarily helpful.

To further study the $G = 2$ case, recall that c_g is independent of x_g with mean zero. In other words, the expected deviation in using the simple comparison-of-means estimator is zero. In effect, it estimates

$$\mu_2 - \mu_1 = (\delta_2 + \beta) - \delta_1 = (\alpha + c_2 + \beta) - (\alpha + c_1) = \beta + (c_2 - c_1). \tag{2.6}$$

Under the DL assumptions, $c_2 - c_1$ has mean zero, and so estimating it should not bias the analysis. DL work under the assumption that β is the parameter of interest, but, if the experiment is properly randomized – as is maintained by DL – it is harmless to include the c_g in the estimated effect.

Consider now a case where the DL approach can be applied. Assume there are $G = 4$

groups with groups one and two control groups ($x_1 = x_2 = 0$) and two treatment groups ($x_3 = x_4 = 1$). The DL approach would involve computing the averages for each group, $\bar{y}_g$, and running the regression $\bar{y}_g$ on $1, x_g$, $g = 1, \ldots, 4$. Inference is based on the t_2 distribution. The estimator $\hat{\beta}$ in this case can be written as

$$\hat{\beta} = (\bar{y}_3 + \bar{y}_4)/2 - (\bar{y}_1 + \bar{y}_2)/2. \tag{2.7}$$

(The pooled OLS regression using the disaggregated data results in the weighted average $(p_3\bar{y}_3 + p_4\bar{y}_4) - (p_1\bar{y}_1 + p_2\bar{y}_2)$, where $p_1 = M_1/(M_1 + M_2)$, $p_2 = M_2/(M_1 + M_2)$, $p_3 = M_3/(M_3 + M_4)$, and $p_4 = M_4/(M_3 + M_4)$ are the relative proportions within the control and treatment groups, respectively.) With $\hat{\beta}$ written as in (2.7), we are left to wonder why we need to use the t_2 distribution for inference. Each $\bar{y}_g$ is usually obtained from a large sample – $M_g = 30$ or so is usually sufficient for approximate normality of the standardized mean – and so $\hat{\beta}$, when properly standardized, has an approximate standard normal distribution quite generally.

In effect, the DL approach rejects the usual inference based on group means from large sample sizes because it may not be the case that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. In other words, the control group may be heterogeneous as might be the treatment group. But this in itself does not invalidate standard inference applied to (2.7).

Equation (2.7) hints at a different way to view the small G, large M_g setup. In this particular application, we estimate two parameters, α and β, given four moments that we can estimate with the data. The OLS estimates from (2.4) in this case are minimum distance estimates that impose the restrictions $\mu_1 = \mu_2 = \alpha$ and $\mu_3 = \mu_4 = \alpha + \beta$. If we use the 4×4 identity matrix as the weight matrix, we get $\hat{\beta}$ as in (2.7) and $\hat{\alpha} = (\bar{y}_1 + \bar{y}_2)/2$. Using the MD approach, we see there are two overidentifying restrictions, which are easily tested. But even if we reject them, it simply implies at least one pair of means within each of the control and treatment groups is different.

With large group sizes, and whether or not G is especially large, we can put the probably generally into an MD framework, as done, for example, by Loeb and Bound (1996), who had $G = 36$ cohort-division groups and many observations per group. For each group g, write

$$y_{gm} = \delta_g + z_{gm}\gamma_g + u_{gm}, m = 1, \ldots, M_g, \tag{2.8}$$

where we assume random sampling within group and independent sampling across groups. We make the standard assumptions for OLS to be consistent (as $M_g \to \infty$) and

$\sqrt{M_g}$-asymptotically normal; see, for example, Wooldridge (2002, Chapter 4). The presence of group-level variables x_g in a "structural" model can be viewed as putting restrictions on the intercepts, δ_g, in the separate group models in (2.8). In particular,

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G, \tag{2.9}$$

where we now think of x_g as fixed, observed attributes of heterogeneous groups. With K attributes we must have $G \geq K + 1$ to determine α and β. If M_g is large enough to estimate the δ_g precisely, a simple two-step estimation strategy suggests itself. First, obtain the $\hat{\delta}_g$, along with $\hat{\gamma}_g$, from an OLS regression within each group. If $G = K + 1$ then, typically, we can solve for $\hat{\theta} \equiv (\hat{\alpha}, \hat{\beta}')'$ uniquely in terms of the $G \times 1$ vector $\hat{\delta}$:. $\hat{\theta} = X^{-1}\hat{\delta}$, where X is the $(K + 1) \times (K + 1)$ matrix with g^{th} row $(1, x_g)$. If $G > K + 1$ then, in a second step, we can use a minimum distance approach, as described in Wooldridge (2002, Section 14.6). If we use as the weighting matrix I_G, the $G \times G$ identity matrix, then the minimum distance estimator can be computed from the OLS regression

$$\hat{\delta}_g \text{ on } 1, x_g, g = 1, \ldots, G. \tag{2.10}$$

Under asymptotics such that $M_g = \rho_g M$ where $0 < \rho_g \leq 1$ and $M \to \infty$, the minimum distance estimator $\hat{\theta}$ is consistent and $\sqrt{M}$-asymptotically normal. Still, this particular minimum distance estimator is asymptotically inefficient except under strong assumptions. Because the samples are assumed to be independent, it is not appreciably more difficult to obtain the efficient minimum distance (MD) estimator, also called the "minimum chi-square" estimator.

First consider the case where z_{gm} does not appear in the first stage estimation, so that the $\hat{\delta}_g$ is just $\bar{y}_g$, the sample mean for group g. Let $\hat{\sigma}_g^2$ denote the usual sample variance for group g. Because the $\bar{y}_g$ are independent across g, the efficient MD estimator uses a diagonal weighting matrix. As a computational device, the minimum chi-square estimator can be computed by using the weighted least squares (WLS) version of (2.10), where group g is weighted by $M_g/\hat{\sigma}_g^2$ (groups that have more data and smaller variance receive greater weight). Conveniently, the reported t statistics from the WLS regression are asymptotically standard normal as the group sizes M_g get large. (With fixed G, the WLS nature of the estimation is just a computational device; the standard asymptotic analysis of the WLS estimator has $G \to \infty$.). The minimum distance approach works with small G provided $G \geq K + 1$ and each M_g is large enough so that normality is a good approximation to the distribution of the (properly scaled) sample average within each group.

If z_{gm} is present in the first-stage estimation, we use as the minimum chi-square weights the inverses of the asymptotic variances for the g intercepts in the separate G regressions. With large M_g, we might make these fully robust to heteroskedasticity in $E(u_{gm}^2|z_{gm})$ using the White (1980) sandwich variance estimator. At a minimum we would want to allow different σ_g^2 even if we assume homoskedasticity within groups. Once we have the $\widehat{Avar}(\hat{\delta}_g)$ – which are just the squared reported standard errors for the $\hat{\delta}_g$ – we use as weights $1/\widehat{Avar}(\hat{\delta}_g)$ in the computationally simple WLS procedure. We are still using independence across g in obtaining a diagonal weighting matrix in the MD estimation.

An important by-product of the WLS regression is a minimum chi-square statistic that can be used to test the $G - K - 1$ overidentifying restrictions. The statistic is easily obtained as the weighted sum of squared residuals, say SSR_w. Under the null hypothesis in (2.9), $SSR_w \overset{a}{\sim} \chi^2_{G-K-1}$ as the group sizes, M_g, get large. If we reject H_0 at a reasonably small significance level, the x_g are not sufficient for characterizing the changing intercepts across groups. If we fail to reject H_0, we can have some confidence in our specification, and perform inference using the standard normal distribution for t statistics for testing linear combinations of the population averages.

We might also be interested in how one of the slopes in γ_g depends on the group features, x_g. Then, we simple replace $\hat{\delta}_g$ with, say $\hat{\gamma}_{g1}$, the slope on the first element of z_{gm}. Naturally, we would use $1/\widehat{Avar}(\hat{\gamma}_{g1})$ as the weights in the MD estimation.

The minimum distance approach can also be applied if we impose $\gamma_g = \gamma$ for all g, as in the original model (1). Obtaining the $\hat{\delta}_g$ themselves is easy: run the pooled regression

$$y_{gm} \text{ on } d1_g, d2_g, \ldots, dG_g, z_{gm}, m = 1, \ldots, M_g; g = 1, \ldots, G \tag{2.11}$$

where $d1_g, d2_g, \ldots, dG_g$ are group dummy variables. Using the $\hat{\delta}_g$ from the pooled regression (2.11) in MD estimation is complicated by the fact that the $\hat{\delta}_g$ are no longer asymptotically independent; in fact, $\hat{\delta}_g = \bar{y}_g - \bar{z}_g\hat{\gamma}$, where $\hat{\gamma}$ is the vector of common slopes, and the presence of $\hat{\gamma}$ induces correlation among the intercept estimators. Let $\hat{V}$ be the $G \times G$ estimated (asymptotic) variance matrix of the $G \times 1$ vector $\hat{\delta}$. Then the MD estimator is $\hat{\theta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\hat{\delta}$ and its estimated asymptotic variance is $(X'\hat{V}^{-1}X)^{-1}$. If the OLS regression (2.10) is used, or the WLS version, the resulting standard errors will be incorrect because they ignore the across group correlation in the estimators. (With large group sizes the errors might be small; see the next section.)

Intermediate approaches are available, too. Loeb and Bound (1996) (LB for short) allow different group intercepts and group-specific slopes on education, but impose common slopes on demographic and family background variable. The main group-level covariate is the student-teacher ratio. Thus, LB are interested in seeing how the student-teach ratio affects the relationship between test scores and education levels. LB use both the unweighted estimator and the weighted estimator and find that the results differ in unimportant ways. Because they impose common slopes on a set of regressors, the estimated slopes on education (say $\hat{\gamma}_{g1}$) are not asymptotically independent, and perhaps using a nondiagonal estimated variance matrix $\hat{V}$ (which would be 36×36 in this case) is more appropriate; but see Section 3.

If we reject the overidentifying restrictions, we are essentially concluding that $\delta_g = \alpha + x_g\beta + c_g,$ where c_g can be interpreted as the deviation from the restrictions in (2.9) for group g. As G increases relative to K, the likelihood of rejecting the restrictions increases. One possibility is to apply the Donald and Lang approach, which is to analyze the OLS regression in (2.10) in the context of the classical linear model (CLM), where inference is based on the t_{G-K-1} distribution. Why is a CLM analysis justified? Since $\hat{\delta}_g = \delta_g + O_p(M_g^{-1/2})$, we can ingore the estimation error in $\hat{\delta}_g$ for large M_g (Recall that the same "large M_g" assumption underlies the minimum distance approach.) Then, it is as if we are estimating the equation $\delta_g = \alpha + x_g\beta + c_g, g = 1, \ldots, G$ by OLS. If the c_g are drawn from a normal distribution, classical analysis is applicable because c_g is assumed to be independent of x_g. This approach is desirable when one cannot, or does not want to, find group-level observables that completely determine the δ_g. It is predicated on the assumption that the other factors in c_g are not systematically related to x_g, a reasonable assumption if, say, x_g is a randomly assigned treatment at the group level, a case considered by Angrist and Lavy (2002).

Beyond the treatment effect case, the issue of how to define parameters of interest appears complicated, and deserves further research.

3. What if G and M_g are Both "Large"?

Section 1 reviewed methods appropriate for a large number of groups and relatively small group sizes. Section 2 considered two approaches appropriate for large group sizes and a small number of groups. The DL and minimum distance approaches use the large group sizes assumption differently: in its most applicable setting, DL use the large M_g assumption to ignore the first-stage estimation error entirely, with all uncertainty coming from unobserved group effects, while the asymptotics underlying the MD approach is based on applying the

central limit theorem within each group. Not surprisingly, more flexibility is afforded if G and M_g are both "large."

For example, suppose we adopt the DL specification (with an unobserved cluster effect),

$$\delta_g = \alpha + x_g\beta + c_g, g = 1, \ldots, G, \tag{3.1}$$

and $G = 50$ (say, states in the U.S.). Further, assume first that the group sizes are large enough (or the cluster effects are so strong) that the first-stage estimation error can be ignored. Then, it matters not whether we impose some common slopes or run separate regressions for each group (state) in the first stage estimation. In the second step, we can treat the group-specific intercepts, $\hat{\delta}_g, g = 1, \ldots, G$, as independent "observations" to be used in the second stage. This means we apply regression (2.10) and apply the usual inference procedures. The difference now is that with $G = 50$, the usual t statistics have some robustness to nonnormality of the c_g, assuming the CLT approximation works well With small G, the exact inference was based on normality of the c_g.

Loeb and Bound (1996), with $G = 38$, essentially use regression (2.10), but with estimated slopes as the dependent variable in place of estimated intercepts. As mentioned in Section 2, LB impose some common slopes across groups, which means the estimated parameters are dependent across group. The minimum distance approach without cluster effects is one way to account for the dependence. Alternatively, one can simply adopt the DL perspective and just assume the estimation error is swamped by c_g; then standard OLS analysis is approximately justfied.

4. NONLINEAR MODELS

Many of the issues for nonlinear models are the same as for linear models. The biggest difference is that, in many cases, standard approaches require distributional assumptions about the unobserved group effects. In addition, it is more difficult in nonlinear models to allow for group effects correlated with covariates, especially when group sizes differ. For the small G case, we offer extensions of the Donald and Lang (2007) approach (with large group sizes) and the minimum distance approach.

Rather than using a general, abstract setting, the issues for nonlinear models are easily illustrated with the probit model. Wooldridge (2006) considers other models (which are also covered in the nonlinear panel data notes).

4.1. Large Group Asymptotics

We can illustrate many issues using an unobserved effects probit model. Let y_{gm} be a

binary response, with x_g and z_{gm}, $m = 1, \ldots, M_g, g = 1, \ldots, G$ defined as in Section 1. Assume that

$$y_{gm} = 1[\alpha + x_g\beta + z_{gm}\gamma + c_g + u_{gm} \geq 0] \tag{4.1}$$

$$u_{gm}|x_g, Z_g, c_g \sim Normal(0, 1) \tag{4.2}$$

(where $1[\cdot]$ is the indicator function). Equations (4.1) and (4.2) imply

$$P(y_{gm} = 1|x_g, z_{gm}, c_g) = P(y_{gm} = 1|x_g, Z_g, c_g) = \Phi(\alpha + x_g\beta + z_{gm}\gamma + c_g), \tag{4.3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf). We assume throughout that only z_{gm} affects the response probability of y_{gm} conditional on x_g and c_g; the outcomes of z_{gp} for $p \neq m$ are assumed not to matter. This is captured in (4.3). For pooled methods we could relax this restriction (as in the linear case), but, with the presence of c_g, this affords little generality in practice.

As in nonlinear panel data models, the presence of c_g in (4.3) raises several important issues, including how we estimate quantities of interest. As in the panel data case, we have some interest in estimating average partial or marginal effects. For example, if the first element of x_g is continuous,

$$\frac{\partial P(y_{gm} = 1|x_g, z_{gm}, c_g)}{\partial x_{g1}} = \beta_1\phi(\alpha + x_g\beta + z_{gm}\gamma + c_g), \tag{4.4}$$

where $\phi(\cdot)$ is the standard normal density function. If

$$c_g|x_g, Z_g \sim Normal(0, \sigma_c^2), \tag{4.5}$$

where the zero mean is without loss of generality because (4.1) contains an intercept, α, then the APEs are obtained from the function

$$P(y_{gm} = 1|x_g, Z_g) = \Phi[(\alpha + x_g\beta + z_{gm}\gamma)/(1 + \sigma_c^2)^{1/2}] \equiv \Phi(\alpha_c + x_g\beta_c + z_{gm}\gamma_c), \tag{4.6}$$

where $\alpha_c = \alpha/(1 + \sigma_c^2)^{1/2}$, and so on. Conveniently, the scaled coefficients are exactly the coefficients estimated by using a simple pooled probit procedure. So, for estimating the average partial effects, pooled probit is perfectly acceptable. With large G and small group sizes, we can easily make the standard errors and test statistics robust to arbitarary within group correlation using standard sandwich variance estimators (robust to within-cluster correlation).

Some authors prefer to call procedures such as pooled probit applied to cluster samples *pseudo maximum likelihood*. Unfortunately, this term is used in contexts where only the

conditional mean is correctly specified in the context of the linear exponential family. Wooldridge (2002, Chapter 13) calls such methods *partial maximum likelihood* to emphasize that we have partially specified a distribution, namely the marginal distribution of y_{gm} given (x_g, Z_m), without specifying a joint distribution $(y_{g1}, \ldots, y_{g,M_g})$ conditional on (x_g, Z_g).

If we supplement (4.1),.(4.2), and (4.5) with

$$\{u_{g1}, \ldots, u_{g,M_g}\} \text{ are independent conditional on } (x_g, Z_g, c_g) \tag{4.7}$$

then we have the so-called *random effects probi*t model. Under the RE probit assumptions, α, β, γ and σ_c^2 are all identified, and estimable by MLE, which means we can estimate the APEs as well as the partial effects evaluated at the mean of c_g, which is zero. We can also compute partial effects at other values of c_g that we might select from the normal distribution with estimated standard deviation σ_c. The details for random effects probit in the balanced panel data case are given in Wooldridge (2002, Chapter 15). The unbalanced case is similar.

As we discussed in the nonlinear panel data notes, minimum distance estimator or generalized estimating equations can be used to obtain estimators (of the scaled coefficients) more efficient than pooled probit but just as robust. (Remember, the RE probit estimator has no known robustness properties to violation of assumption (4.7).)

A very challenging task, and one that appears not to have gotten much attention for true cluster samples, is allowing correlation between the unobserved heterogeneity, c_g, and the covariates that vary within group, z_{gm}. (For notational simplicity, we assume there are no group-level controls in the model, but these can always be added.) For linear models, we know that the within or fixed effects estimator allows arbitrary correlation, and does not restrict the within-cluster dependence of $\{u_{g1}, \ldots, u_{g,M_g}\}$. Unfortunately, allowing correlation between c_g and $(z_{g1}, z_{g2}, \ldots, z_{gM})$ is much more difficult in nonlinear models. In the balanced case, where the group sizes M_g are the same for all g, the Chamberlain (1980) device can be used:

$$c_g | Z_g \sim Normal(\eta + \bar{z}_g \xi, \sigma_a^2), \tag{4.8}$$

where σ_a^2 is the conditional variance $Var(c_g|Z_g)$. If we use all random effects probit assumptions but with (4.8) in place of (4.5), then we obtain a simple extension of the RE probit model: simply add the group averages, $\bar{z}_g$, as a set of additional explanatory variables. This is identical to the balanced panel case we covered earlier. The marginal distributions are

$$P(y_{gm} = 1|Z_g) = \Phi[(\eta + z_{gm}\gamma + \bar{z}_g\xi)/(1 + \sigma_a^2)^{1/2}] \equiv \Phi(\eta_a + z_{gm}\gamma_a + \bar{z}_g\xi_a) \tag{4.9}$$

where now the coefficients are scaled by a function of the conditional variance. This is just as

in the case of a balanced panel, and all calculations, including those for APEs, follow immediately.

The Chamberlain-Mundlak needs to be modified for the unbalanced case. [One possibility is to discard observations and balance the cluster sample under the assumption that the cluster sizes are exogenous, and that might be desirable if there is not much variation in the cluster sizes.] An alternative is to use the cluster setup and assuming a kind of exchangeability assumption concerning the correlation between the cluster effect and the covariates. At a minimum, (4.8) should be modified to allow the variances to depend on the cluster size, M_g. Under restrictive assumptions, such as joint normality of $(c_g, z_{g1}, \ldots, z_{g,M_g})$, with the z_{gm} independent and identically distributed within a cluster, one can derive $Var(c_g|Z_g)$. But these are strong assumptions. We might just assume

$$c_g|(z_{g1}, \ldots, z_{g,M_g}) \sim \text{Normal}(\eta + \bar{z}_g\xi, \sigma^2_{a,M_g}), \tag{4.10}$$

where σ^2_{a,M_g} denotes a different variance for each group size, M_g. Then the marginal distributions are

$$P(y_{gm} = 1|Z_g) = \Phi[(\eta + z_{gm}\gamma + \bar{z}_g\xi)/(1 + \sigma^2_{a,M_g})^{1/2}]. \tag{4.11}$$

Equation (4.11) can be estimated by pooled probit, allowing for different group variances. (A normalization is also required.) A simpler approach is to estimate a different set of parameters, $(\eta_{M_g}, \xi_{M_g}, \gamma_{M_g})$, for each group size, and then to impose the restrictions in (4.11) by minimum distance estimation. With very large G and little variation in M_g, we might just use the unrestricted estimates $(\hat{\eta}_{M_g}, \hat{\xi}_{M_g}, \hat{\gamma}_{M_g})$, estimate the APEs for each group size, and then average these across group size. But more work needs to be done to see if such an approach loses too much in terms of efficiency.

The methods of Altonji and Matzkin (2005) – see also Wooldridge (2005) – can be applied. A completely nonparametric approach is based on

$$P(y_{gm} = 1|Z_g, c_g) = P(y_{gm} = 1|z_{gm}, c_g) \equiv F(z_{gm}, c_g) \tag{4.12}$$

and

$$D(c_g|z_{g1}, z_{g2}, \ldots, z_{g,M_g}) = D(c_g|\bar{z}_g). \tag{4.13}$$

Define $H_g(z_{gm}, \bar{z}_g) = P(y_{gm} = 1|z_{gm}, \bar{z}_g)$. As discussed in the nonlinear panel data notes, under (4.12) and (4.13), it can be show that the APEs are obtained from

$$E_{\bar{z}_g}[H_g(z, \bar{z}_g)]. \tag{4.14}$$

If the group sizes differ, $H_g(\cdot,\cdot)$ generally depends on g. If there are relatively few group sizes, it makes sense to estimate the $H_g(\cdot,\cdot)$ separately for each group size M_g. Then, the APEs can be estimated from

$$G^{-1}\sum_{g=1}^{G}\hat{H}_g(z,\bar{z}_g). \tag{4.15}$$

As discussed before, as a practical matter we might just use flexible parametric models, such as probit with flexible functional forms.

Other strategies are available for estimating APEs. We can apply "fixed effects probit" to cluster samples just as with panel data and treat the c_g as parameters to estimate in

$$P(y_{gm} = 1|Z_g, c_g) = P(y_{gm} = 1|z_{gm}, c_g) = \Phi(z_{gm}\gamma + c_g). \tag{4.16}$$

The same issues arise as in the panel data case, except with true cluster samples the conditional indepdence assumption likely is more reasonable than in the panel data case. With small group sizes M_g (say, siblings or short panel data sets), treating the c_g as parameters to estimate creates an incidental parameters problem. As before, we might use

$$G^{-1}\sum_{g=1}^{G}\Phi(z\hat{\gamma} + \hat{c}_g), \tag{4.17}$$

to estimate the APEs.

The logit conditional MLE can be applied to cluster samples essentially without change, which means we can estimate the parameters, γ, without restricting $D(c_g|Z_g)$. This is especially convenient in the unbalanced case.

4.2. A Small Number of Groups and Large Group Sizes

Unlike in the linear case, for nonlinear models exact inference is unavailable even under the strongest set of assumptions. Nevertheless, if the group sizes M_g are reasonably large, we can extend the DL approach to nonlinear models and obtain approximate inference. In addition, the the minimum distance approach carries over essentially without change.

We can apply the methods to any nonlinear model that has an index structure – which includes all of the common ones, and many other models besides, but we again consider the probit case. With small G and random sampling of $\{(y_{gm}, z_{gm}) : m = 1,\ldots,M_g\}$ within each g, write

$$P(y_{gm} = 1|z_{gm}) = \Phi(\delta_g + z_{gm}\gamma_g), m = 1,\ldots,M_g \tag{4.18}$$

$$\delta_g = \alpha + x_g\beta, g = 1, \ldots, G. \tag{4.19}$$

As with the linear model, we assume the intercept, δ_g in (4.18), is a function of the group features x_g. With the M_g moderately large, we can get good estimates of the δ_g. The $\hat{\delta}_g, g = 1, \ldots, G$, are easily obtained by estimating a separate probit for each group. Or, we can impose common γ_g and just estimate different group intercepts (sometimes called "group fixed effects").

Under (4.19), we can apply the minimum distance approach just as before. Let $\widehat{Avar}(\hat{\delta}_g)$ denote the estimated asymptotic variances of the $\hat{\delta}_g$ (so these shrink to zero at the rate $1/M_g$). If the $\hat{\delta}_g$ are obtained from G separate probits, they are independent, and the $\widehat{Avar}(\hat{\delta}_g)$ are all we need. As in the linear case, if a pooled method is used, the $G \times G$ matrix $\widehat{Avar}(\hat{\delta})$ should be obtained as the weighting matrix. For binary response, we use the usual MLE estimated variance. If we are using fractional probit for a fractional response, these would be from a sandwich estimate of the asymptotic variance. In the case where the $\hat{\delta}_g$ are obtained from separate probits, we can obtain the minimum distance estimates as the WLS estimates from

$$\hat{\delta}_g \text{ on } 1, x_g, g = 1, \ldots, G \tag{4.20}$$

using weights $1/\widehat{Avar}(\hat{\delta}_g)$ are used as the weights. This is the efficient minimum distance estimator and, conveniently, the proper asymptotic standard errors are reported from the WLS estimation (even though we are doing large M_g, not large G, asymptotics.) Generally, we can write the MD estimator as in the linear case, $\hat{\theta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\hat{\delta}$, where $\hat{\delta}$ is the $G \times 1$ vector of $\hat{\delta}_g$ and $\hat{V} = \widehat{Avar}(\hat{\delta})$. The overidentification test is obtained exactly as in the linear case: there are $G - K - 1$ degrees-of-freedom in the chi-square distribution.

The same cautions about using the overidentification test to reject the minimum distance approach apply here as well. In particular, in the treatment effect setup, where x_g is zero or one, we might reject a comparision of means across multiple groups simply because the means within the control or within the treatment group differ, or both. It might make sense to define the treatment effect as the difference in averages between treatment and control, or use weighted averages, without worrying about whether the means are the same. (We consider an alternative, namely, using data to choose a synthetic control from a set of potentil control groups, the the notes on difference-in-differences.)

If we reject the overidentification restrictions, we can adapt Donald and Lang (2007) and

treat

$$\hat{\delta}_g = \alpha + x_g\beta + error_g, g = 1, \ldots, G \tag{4.21}$$

as approximately satisfying the classical linear model assumptions, provided $G > K + 1$, just as before. As in the linear case, this approach is justified if $\delta_g = \alpha + x_g\beta + c_g$ with c_g independent of x_g and c_g drawn from a homoskedastic normal distribution. It assumes that we can ignore the estimation error in $\hat{\delta}_g$, based on $\hat{\delta}_g = \delta_g + O(1/\sqrt{M_g})$. Because the DL approach ignores the estimation error in $\hat{\delta}_g$, it is unchanged if one imposes some constant slopes across the groups, as with the linear model.

Once we have estimated α and β, the estimated effect on the response probability can be obtained by averaging the response probability for a given x:

$$G^{-1}\sum_{g=1}^{G}\left(M_g^{-1}\sum_{m=1}^{M_g}\Phi(\hat{\alpha} + x\hat{\beta} + z_{gm}\hat{\gamma}_g)\right), \tag{4.22}$$

where derivatives or differences with respect to the elements of x can be computed. Here, the minimum distance approach has an important advantage over the DL approach: the finite sample properties of (4.22) are virtually impossible to obtain, whereas the large-M_g asymptotics underlying minimum distance would be straightforward using the delta method. How the bootstrap might work in this situation is an interesting question.

Particularly with binary response problems, the two-step methods described here are problematical when the response does not vary within group. For example, suppose that x_g is a binary treatment – equal to one for receiving a voucher to attend college – and y_{gm} is an indicator of attending college. Each group is a high school class, say. If some high schools have all students attend college, one cannot use probit (or logit) of y_{gm} on $z_{gm}, m = 1, \ldots, M_g$. A linear regression returns zero slope coefficients and intercept equal to unity. Of course, if randomization occurs at the group level – that is, x_g is independent of group attributes – then it is not necessary to control for the z_{gm}. Instead, the within-group averages can be used in a simple minimum distance approach. In this case, as y_{gm} is binary, the DL approximation will not be valid, as the CLM assumptions will not even approximately hold in the model $\bar{y}_g = \alpha + x_g\beta + e_g$ (because $\bar{y}_g$ is always a fraction regardless of the size of M_g).

4.3. Large G and Large M_g

As in the linear case, more flexibility is afforded if G is somewhat large along with large M_g. If we can ignore the estimation error in the $\hat{\delta}_g$, then, in implementing the DL approach –

with or without common slopes imposed in the first stage – one gains robustness to nonnormality of c_g if G is large enough so that $G^{-1/2}\sum_{g=1}^{G} c_g$ and $G^{-1/2}\sum_{g=1}^{G} x_g c_g$ are approximately normally distributed. The second step is the same as in the linear model case: $\hat{\delta}_g$ is regressed on 1, x_g, $g = 1,\ldots,G$; one can use heteroskedasticity-robust inference with large G to partly account for the estimation error in the $\hat{\delta}_g$.

A version of the method proposed by Berry, Levinsohn, and Pakes (1995) for estimating structural models using both individual-level and product-level data, or market-level data, or both can be treated in the large G, large M_g framework, where g indexes good or market and m indexes individuals within a market. BLP's original application was where g indexes different automobile models. Petrin and Train (2002) cover the case of about 170 television markets and four TV services. To handle this case, assume that H products are available in each market. Therefore, we now think of δ_g as an H-vector for each g, and so is c_g. The main difference here with the previous setup is that, for reasons discussed in BLP and Petrin and Train, we must allow the c_{gh} to be correlated with the x_{gh} (which contains the price of good j in market g, in addition to product/market attributes). BLP propose a two-step estimation strategy. In the first step, a choice model, such as multinomial logit, is estimated using the individual-level data pooled across markets. The key estimates are what we call the $\hat{\delta}_g$ – the market "fixed effects." Typically, most or all of the "slope" parameters in the multinomial logit estimation are assumed to be constant across g, although, with many individuals per market, that is not necessary.

In the second step, the $\hat{\delta}_{gh}$ are used in place of δ_{gh} in the market/good-level equation

$$\delta_{gh} = \alpha + x_{gh}\beta + c_{gh}, \mathrm{h} = 1,\ldots,H;\ g = 1,\ldots,G, \tag{4.23}$$

where, say, w_g is a set of instruments for x_{gh}. (Typically, w_g varies only by market, g, and not by good, h.) This allows for market/good-specific unobservables in the individual choice equations to be correlated with prices. If we could observe the δ_{gh}, then (4.23) would be a standard problem in IV estimation for a cross section system of equations, provided G is large enough to invoke the law of large numbers and central limit theorem. Replacing δ_g with $\hat{\delta}_g$ is justified if the M_g are large because the variance of c_g will dominate that of the $\hat{\delta}_g$. Further, any correlation induced in the $\hat{\delta}_g$ by pooling in the first-stage estimation shrinks to zero at the rate $1/M$, where we can think of M as the average group size. In other words, we just apply, say, 2SLS in the second step.

Ignoring the estimation in $\hat{\delta}_g$, efficient estimation is obtained by writing the system of equations as

$$\hat{\delta}_g \approx X_g\theta + c_g \tag{4.24}$$

where X_g is the $J \times (K+1)$ matrix of attributes (including an intercept and prices). Because (4.24) is a system of equations with instruments $I_J \otimes w_g$, we can use the 3SLS estimator or GMM to efficiently account for the correlation across $\{c_{gh} : h = 1,\ldots,H\}$.

5. Estimation of Population Parameters with Stratified Samples

We now provide a brief, modern treatment of estimation with stratified samples. The emphasis here is in estimation parameters from a population that has been stratified. Typically, with stratified sampling, some segments of the population are over- or underrepresented by the sampling scheme. Fortunately, if we know enough information about the stratification scheme, we can often modify standard econometric methods and consistently estimate population parameters.

There are two common types of stratified sampling, standard stratified (SS) sampling and variable probability (VP) sampling. A third type of sampling, typically called multinomial sampling, is practically indistinguishable from SS sampling, but it generates a random sample from a modified population (thereby simplifying certain theoretical analyses). See Cosslett (1993), Imbens (1992), Imbens and Lancaster (1996), and Wooldridge (1999) for further discussion. We focus on SS and VP sampling here.

SS sampling begins by partitioning the sample space (set of possible outcomes), say W, into G non-overlapping, exhaustive groups, $\{W_g : g = 1,\ldots G\}$. Then, a random sample is taken from each group g, say $\{w_{gi} : i = 1,\ldots,N_g\}$, where N_g is the number of observations drawn from stratum g and $N = N_1 + N_2 + \ldots + N_G$ is the total number of observations. If w is a random vector representing the population, and taking values in W, then each random draw from stratum g has the same distribution as w conditional on w belonging to W_g:

$$D(w_{gi}) = D(w|w \in W_g),\ i = 1,\ldots,N_g.$$

Therefore, the resulting sample across all strata consists of independent but not identically distributed observations. Unless we are told, we have no way of knowing that our data came from SS sampling.

What if we want to estimate the mean of w from an SS sample? It turns out we cannot get an unbiased or consistent estimator of unless we have some additional information. Typically,

the information comes in the form of population frequencies for each of the strata. Specifically, let $\pi_g = P(w \in W_g)$ be the probability that w falls into stratum g; the π_g are often called the "aggregate shares."

If we know the π_g (or can consistently estimate them), then $\mu_w = E(w)$ is identified by a weighted average of the expected values for the strata:

$$\mu_w = \pi_1 E(w|w \in W_1) + \ldots + \pi_G E(w|w \in W_G). \tag{5.1}$$

Because we can estimate each of the conditional means using the random sample from the appropriate stratum, an unbiased estimator of is simply

$$\hat{\mu}_w = \pi_1 \bar{w}_1 + \pi_2 \bar{w}_2 \ldots + \pi_G \bar{w}_G, \tag{5.2}$$

where $\bar{w}_g$ is the sample average from stratum g. As the strata sample sizes grow, $\hat{\mu}_w$ is also a consistent estimator of μ_w. The variance of $\hat{\mu}_w$ is easily obtained because of independence withing and between strata:

$$Var(\hat{\mu}_w) = \pi_1^2 Var(\bar{w}_1) + \ldots + \pi_G^2 Var(\bar{w}_G). \tag{5.3}$$

Because $Var(\bar{w}_g) = \sigma_g^2/N_f$, each of the variances can be estimated in an unbiased fashion by using the usual unbiased variance estimator,

$$\hat{\sigma}_g^2 = (N_g - 1)^{-1} \sum_{i=1}^{N_g} (w_{gi} - \bar{w}_g)^2. \tag{5.4}$$

Sometimes it is useful to have a formula for $\hat{\mu}_w$ that shows it is a weighted average across all observations:

$$\begin{aligned} \hat{\mu}_w &= (\pi_1/h_1) N^{-1} \sum_{i=1}^{N_1} w_{1i} + \ldots + (\pi_G/h_G) N^{-1} \sum_{i=1}^{N_G} w_{Gi} \\ &= N^{-1} \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) w_i \end{aligned} \tag{5.5}$$

where $h_g = N_g/N$ is the fraction of observations in stratum g and in (5.5) we drop the strata index on the observations and use the stratum for observation i, g_i, to pick out the appropriate weight, π_{g_i}/h_{g_i}. Formula (1.5) is useful because the sampling weights associated with SS samples are reported as (π_{g_i}/h_{g_i}), and so applying these weights in averaging across all N produces an unbiased, consistent estimator. Nevertheless, the large sample properties of estimators from SS samples are properly derived from (5.2) and its extensions.

A different sampling scheme is usually called *variable probability (VP) sampling*, which is

more convenient for telephone or email surveys, where little, if anything, is known ahead of time about those being contacted. With VP sampling, each stratum g is assigned a nonzero sampling probability, p_g. Now, a random draw w_i is taking from the population, and it is kept with probability p_g if $w_i \in W_g$. With VP sampling, the population is sampled N times. Often N is not reported with VP samples (although, as we discuss latter, knowing how many times each stratum was sampled can improve efficiency). Instead, we know how many data points were kept, and we call this M. Because of the randomness in whether an observation is kept, M is properly viewed as a random variable. With VP sampling, it is handy for each draw from the population to have a selection indicator, s_i, which is one of observation i is kept (and then its stratum is also known). Then $M = \sum_{i=1}^{N} s_i$. Let z_i be a G-vector of stratum indicators, and let $p(z_i) = p_1 z_{i1} + \ldots + p_G z_{iG}$ be the function that delivers the sampling probability for any random draw i.

A key assumption for VP sampling is that conditional on being in stratum g, the chance of keeping an observation is p_g. Statistically, this means that, conditional on z_i, s_i and w_i are independent. Using this assumption, we can show, just as in the treatment effect case,

$$E[(s_i/p(z_i))w_i] = E(w_i); \tag{5.6}$$

that is, weighting a selected observation by the inverse of its sampling probability allows us to recover the population mean. (We will use a more general version of this result when we discuss missing data general. Like estimating counterfactual means in program evaluation, VP sampling is, in effect, a missing data problem. But it is usually treated along with other stratified sampling schemes.) It follows that

$$N^{-1} \sum_{i=1}^{N} (s_i/p(z_i))w_i \tag{5.7}$$

is a consistent estimator of $E(w_i)$. We can also write this as

$$(M/N)M^{-1} \sum_{i=1}^{N} (s_i/p(z_i))w_i; \tag{5.8}$$

if we define weights as $\hat{v}_i = \hat{\rho}/p(z_i)$ where $\hat{\rho} = M/N$ is the fraction of observations retained from the sampling scheme, then (5.8) is $M^{-1} \sum_{i=1}^{M} \hat{v}_i w_i$, where only the observed points are included in the sum. Thus, like in the SS case, we can write the esimator for the mean under VP sampling as a weighted average of the observed data points. In the VP case, the weight is (an estimate of) the probability of keeping an observation, $\rho = P(s_i = 1)$, over the probability

that an observation in its stratum is kept. If $p_g < \rho$, the observations for stratum g are underpresented in the eventual sample (asymptotically), and they receive weight greater than one.

In both the SS and VP cases, one may replace the number of observed data points in the average with the sum of the weights described in each case.

Virtually any estimation method can be used with SS or VP sampled data. Wooldridge (1999, 2001) covers M-estimation for the VP and SS cases, respectively. This includes a wide variety of estimators, including least squares, MLE, and quasi-MLE. There are several interesting findings concerning asymptotic efficiency and estimating the asymptotic variances. Consider the problem of linear regression for simplicity; analogous claims hold for MLE, NLS, and many other estimators. The model in the population is

$$y = \mathbf{x}\boldsymbol{\beta} + u, \tag{5.9}$$

where $\boldsymbol{\beta}$ may index the conditional mean, but consistency follows from $E(\mathbf{x}'u) = \mathbf{0}$. Consider SS sampling. Then a consistent estimator $\hat{\boldsymbol{\beta}}$ is obtained from the "weighted" least squares problem

$$\min_{\mathbf{b}} \sum_{i=1}^{N} v_i(y_i - \mathbf{x}_i\mathbf{b})^2, \tag{5.10}$$

where $v_i = \pi_{g_i}/h_{g_i}$ is the weight for observation i. Remember, the weighting used here is not to solve any heteroskedasticity problem; it is to reweight the sample in order to consistently estimate the population parameter $\boldsymbol{\beta}$.

One possibility for performing inference on $\hat{\boldsymbol{\beta}}$ is to use the White (1980) robust sandwich estimator and associated statistics. This estimator is routinely reported by regression packages when sampling weights are included. In fact, sometimes this estimator is consistent for $Avar\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. There are two assumptions that imply consistency of this widely used variance matrix estimator: (i) $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, so that we are actually estimating a conditional mean; and (ii) the strata are determined by the explanatory variables, $\mathbf{x}$. It turns out that when the White estimator is not consistent, it is actually conservative. In other words, the White estimator converges to a matrix that is larger, in the matrix sense, than the correct asymptotic variance.

To obtain the correct asymptotic variance, we need to use a more detailed formulation of the estimation problem, which is

$$\min_{\mathbf{b}} \left\{ \sum_{g=1}^{G} \pi_g \left[N_g^{-1} \sum_{i=1}^{N} (y_{gi} - \mathbf{x}_{gi}\mathbf{b})^2 \right] \right\} \tag{5.11}$$

so that we are minimizing the a weighted average sum of squared residuals. Using this formulation – actually, the M-estimator version of it – Wooldridge (2001) showed that a consistent estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \widehat{Avar(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} = & \left[\sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \\ & \cdot \left\{ \sum_{g=1}^{G} (\pi_g/h_g)^2 \left[\sum_{i=1}^{N_g} (\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})(\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})' \right] \right\} \\ & \cdot \left[\sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) \mathbf{x}_i' \mathbf{x}_i \right]^{-1}. \end{aligned} \tag{5.12}$$

This formula looks a bit daunting, but, it can be seen that the outer parts of the sandwich are identical to the usual White sandwich estimator. The difference is in the middle. The usual estimator ignores the information on the strata of the observations, which is the same as dropping the within-strata averages, $\overline{\mathbf{x}_g' \hat{u}_g}$. Because a smaller sum of squared residuals (in a matrix sense) is obtained by subtracting off the same average – rather than centering around zero – the matrix in (5.12) is smaller than the usual White matrix. That happens asymptotically, too, provided the means $E(\mathbf{x}'u|\mathbf{w} \in W_g)$, where $\mathbf{w} = (\mathbf{x}, y)$, are nonzero. So, it is the difference between subtracting off within-strata averages and not that produces the more precise inference with stratified sampled data. Econometrics packages, such as Stata, will compute (5.12) provided strata membership is included along with the weights. If only the weights are provided, the larger asymptotic variance is computed.

One case where there is no gain from subtracting within-strata means is when $E(u|\mathbf{x}) = 0$ and $\mathbf{w} \in W_g$ is the same as $\mathbf{x} \in X_g$ for some partition of the regressor space. In fact, if we add the homoskedasticity assumption $Var(u|\mathbf{x}) = \sigma^2$, then the weighted estmator is less efficient than the unweighted estimator, which, of course, is also consistent because $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ and stratification is based on $\mathbf{x}$. So, the cost to weighting when the classical linear model assumptions hold and stratification is exogenous is in terms of efficiency loss.

Some argue that even if stratification is based on $\mathbf{x}$, one should use the weighted estimator. The argument is based on consistently estimating the linear projection, $L(y|\mathbf{x})$, even if the

conditional mean is not linear. If we can only assume $L(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, then the weighted estimator consistently estiimates $\boldsymbol{\beta}$ whether or not the stratification is based on $\mathbf{x}$. The unweighted estimator does not consistently estimate $\boldsymbol{\beta}$ in either case.

The previous discussion applies to nonlinear least squares and maximum likelihood problems, and others. Now, to exploit the stratification, strata means should be subtracted from the gradient of the objective function when computing the asymptotic variance. This requires knowing the stratum and its weight for each observation. A conservative estimate is obtained by the Huber-White sandwich form for misspecified MLE – but with sampling weights. This is the proper formula for, say, MLE if the conditional density $f(y|\mathbf{x},\boldsymbol{\theta})$ is correctly specified and stratification is based on $\mathbf{x}$. But in that case the unweighted MLE is fully efficient, and the usual variance matrix estimators can be used. The weighted estimator does consistently estimate the solution to the population problem $\min_{\mathbf{t}} E[\log f(y|\mathbf{x},\mathbf{t})]$ if the density is misspecified, and that is valuable in some situations.

The above findings have analogs for VP sampling. One interesting finding is that while the Huber-White sandwich matrix applied to the weighted objective function (weighted by the $1/p_g$) *is* consistent when the known p_g are used, an asymptotically more efficient estimator is available when the retention frequencies, $\hat{p}_g = M_g/N_g$, are observed, where M_g is the number of observed data points in stratum g and N_g is the number of times stratum g was sampled. We always know M_g if we are given a stratum indicator with each observation. Generally, N_g might not be known. If it is, we should use the $\hat{p}_g$ in place of p_g. Results such as this are discussed in Imbens (1992), Imbens and Lancaster (1996), and Wooldridge (1999, 2007). The VP sampling example in Wooldridge (2007) can be used to show that the following matrix is valid:

$$\widehat{Avar(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} = \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i} \right]^{-1} \cdot \left\{ \sum_{g=1}^{G} \hat{p}_g^{-2} \left[\sum_{i=1}^{M_g} (\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})(\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})' \right] \right\} \cdot \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i} \right]^{-1}, \tag{5.13}$$

where, remember, M_g is the number of observed data points in stratum g, and the above sums are over the observed data points. This formula is essentially the same as (5.12) in that the

quantities are weighted so that the sample represents the population and $\mathbf{x}_{gi}'\hat{u}_{gi}$ are centered about the within-strata means. If we use the known sampling weights, we drop $\mathbf{x}_g'\hat{u}_g$ from (5.13). If $E(u|\mathbf{x}) = 0$ and the sampling is exogenous, we also drop this term because $E(\mathbf{x}'u|\mathbf{w} \in W_g) = \mathbf{0}$ for all g, and this is whether or not we estimate the p_g. See Wooldridge (2007) for how these same claims carry over to general nonlinear models and estimation methods.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 9, Tuesday, July 31th, 3.15-4.15pm

Partial Identification

1. Introduction

Traditionally in constructing statistical or econometric models researchers look for models that are *(point-)identified*: given a large (infinite) data set, one can infer without uncertainty what the values are of the objects of interest, the estimands. Even though the fact that a model is identified does not necessarily imply that we do well in finite samples, it would appear that a model where we cannot learn the parameter values even in infinitely large samples would not be very useful. Traditionally therefore researchers have stayed away from models that are not (point-)identified, often adding assumptions beyond those that could be justified using substantive arguments. However, it turns out that even in cases where we cannot learn the value of the estimand *exactly* in large samples, in many cases we can still learn a fair amount, even in finite samples. A research agenda initiated by Manski (an early paper is Manski (1990), monographs include Manski (1995, 2003)), referred to as *partial identification*, or earlier as *bounds*, and more recently adopted by a large number of others, notably Tamer in a series papers (Haile and Tamer, 2003, Ciliberto and Tamer, 2007; Aradillas-Lopez and Tamer, 2007), has taken this perspective. In this lecture we focus primarily on a number of examples to show the richness of this approach. In addition we discuss some of the theoretical issues connected with this literature, and some practical issues in implementation of these methods.

The basic set up we adopt is one where we have a random sample of units from some population. For the typical unit, unit i, we observe the value of a vector of variables Z_i. Sometimes it is useful to think of there being in the background a latent variable variable W_i. We are interested in some functional θ of the joint distribution of Z_i and W_i, but, not observing W_i for any units, we may not be able to learn the value of θ even in infinite samples because the estimand cannot be written as a functional of the distribution of Z_i alone. The

three key questions are (i) what we can learn about θ in large samples (identification), (ii) how do we estimate this (estimation), and (iii) how do we quantify the uncertainty regarding θ (inference).

The solution to the first question will typically be a set, the *identified set*. Even if we can characterize estimators for these sets, computing them can present serious challenges. Finally, inference involves challenges concerning uniformity of the coverage rates, as well as the question whether we are interested in coverage of the entire identified set or only of the parameter of interest.

There are a number of cases of general interest. I will discuss two leading cases in more detail. In the first case the focus is on a scalar, with the identified set equal to an interval with lower and upper bound a smooth, $\sqrt{N}$-estimable functional of the data. A second case of interest is that where the information about the parameters can be characterized by moment restrictions, often arising from revealed preference comparisons between utilities at actions taken and actions not taken. I refer to this as the generalized inequality restrictions (GIR) setting. This set up is closely related to the generalized method of moments framework.

2. Partial Identification: Examples

Here we discuss a number of examples to demonstrate the richness of the partial identification approach.

2.1 Missing Data

This is a basic example, see e.g., Manski (1990), and Imbens and Manski (2004). It is substantively not very interesting, but it illustrates a lot of the basic issues. Suppose the observed variable is the pair $Z_i = (D_i, D_i \cdot Y_i)$, and the unobserved variable is $W_i = Y_i$. D_i is a binary variable. This corresponds to a missing data case. If $D_i = 1$, we observe Y_i, and if $D_i = 0$ we do not observe Y_i. We always observe the missing data indicator D_i. We assume the quantity of interest is the population mean $\theta = \mathbb{E}[Y_i]$.

In large samples we can learn $p = \mathbb{E}[D_i]$ and $\mu_1 = \mathbb{E}[Y_i|D_i = 1]$. The data contain no

information about $\mu_0 = \mathbb{E}[Y_i|D_i = 0]$. It can be useful, though not always possible, to write the estimand in terms of parameters that are point-identified and parameters that the data are not informative about. In this case we can do so:

$$\theta = p \cdot \mu_1 + (1-p) \cdot \mu_0.$$

Since even in large samples we learn nothing about μ_0, it follows that without additional information there is no limit on the range of possible values for θ. Even if p is very close to 1, this small probability that $D_i = 0$ combined with the possibility that μ_0 is very large or very small allows for a wide range of values for θ.

Now suppose we know that the variable of interest is binary: $Y_i \in \{0, 1\}$. Then natural (not data-informed) lower and upper bounds for μ_0 are 0 and 1 respectively. This implies bounds on θ:

$$\theta \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = [p \cdot \mu_1, p \cdot \mu_1 + (1-p)] \,.$$

These bounds are *sharp*, in the sense that without additional information we can not improve on them. Formally, for all values θ in $[\theta_{\text{LB}}, \theta_{\text{UB}}]$, we can find a joint distribution of (Y_i, W_i) that is consistent with the joint distribution of the observed data and with θ. Even if Y is not binary, but has some natural bounds, we can obtain potentially informative bounds on θ.

We can also obtain informative bounds if we modify the object of interest a little bit. Suppose we are interested in quantiles of the distribution of Y_i. To make this specific, suppose we are interested in the median of Y_i, $\theta_{0.5} = \text{med}(Y_i)$. The largest possible value for the median arises if all the missing value of Y_i are large. Define $q_\tau(Y_i|D_i = d)$ to be the τ quantile of the conditional distribution of Y_i given $D_i = d$. Then the median cannot be larger than $q_{1/(2p)}(Y_i|D_i = 1)$ because even if all the missing values were large, we know that at least $p \cdot (1/(2p)) = 1/2$ of the units have a value less than or equal to $q_{1/(2p)}(Y_i|D_i = 1)$. Similarly, the smallest possible value for the median corrresponds to the case where all the

missing values are small, leading to a lower bound of $q_{(2p-1)/(2p)}(Y_i|D_i=1)$. Then, if $p>1/2$, we can infer that the median must satisfy

$$\theta_{0.5}\in[\theta_{\rm LB},\theta_{\rm UB}]=\left[q_{(2p-1)/(2p)}(Y_i|D_i=1),q_{1/(2p)}(Y_i|D_i=1)\right],$$

and we end up with a well defined, and, depending on the data, more or less informative identified interval for the median. If fewer than 50% of the values are observed, or $p<1/2$, then we cannot learn anything about the median of Y_i without additional information (for example, a bound on the values of Y_i), and the interval is $(-\infty,\infty)$. More generally, we can obtain bounds on the τ quantile of the distribution of Y_i, equal to

$$\theta_\tau\in[\theta_{\rm LB},\theta_{\rm UB}]=\left[q_{(\tau-(1-p))/p}(Y_i|D_i=1),q_{\tau/p}(Y_i|D_i=1)\right].$$

which is bounded if the probability of Y_i being missing is less than $\min(\tau,1-\tau)$.

2.2 Returns to Schooling

Manski and Pepper (2000, MP) are interested in estimating returns to schooling. They start with an individual level response function $Y_i(w)$, where $w\in\{0,1,\ldots,20\}$ is years of schooling. Let

$$\Delta(s,t)=\mathbb{E}[Y_i(t)-Y_i(s)],$$

be the difference in average outcomes (log earnings) given t rather than s years of schooling. Values of $\Delta(s,t)$ at different combinations of (s,t) are the object of interest. Let W_i be the actual years of school, and $Y_i=Y_i(W_i)$ be the actual log earnings. If one makes an unconfoundedness type assumption that

$$Y_i(w)\ \perp\!\!\!\perp\ W_i\ \Big|\ X_i,$$

for some set of covariates, one can estimate $\Delta(s,t)$ consistently given some support conditions. MP relax this assumption. Dropping this assumption entirely without additional

assumptions one can derive the bounds using the missing data results in the previous section. In this case most of the data would be missing, and the bounds would be wide. More interestingly MP focus on a number of alternative, weaker assumptions, that do not allow for point-identification of $\Delta(s,t)$, but that nevertheless may be able to narrow the range of values consistent with the data to an informative set. One of their assumptions requires that increasing education does not lower earnings:

Assumption 1 (Monotone Treatment Response)
If $w' \geq w$, *then* $Y_i(w') \geq Y_i(w)$.

Another assumption states that, on average, individuals who choose higher levels of education would have higher earnings at each level of education than individuals who choose lower levels of education.

Assumption 2 (Monotone Treatment Selection)
If $w'' \geq w'$, *then for all* w, $\mathbb{E}[Y_i(w)|W_i = w''] \geq \mathbb{E}[Y_i(w)|W_i = w']$.

Both assumptions are consistent with many models of human capital accumulation. They also address the main concern with the exogenous schooling assumption, namely that higher ability individuals who would have had higher earnings in the absence of more schooling, are more likely to acquire more schooling.

Under these two assumptions, the bound on the average outcome given w years of schooling is

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \geq w) + \sum_{v<w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v)$$

$$\leq \mathbb{E}[Y_i(w)] \leq$$

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \leq w) + \sum_{v>w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v).$$

Using data from the National Longitudinal Study of Youth MP a point estimator for the upper bound on the the returns to four years of college, $\Delta(12,16)$ to be 0.397, with a 0.95 upper quantile of 0.450. Translated into an average yearl returns this gives us 0.099, which is in fact lower than some estimates that have been reported in the literature. This analysis suggests that the upper bound is in this case reasonably informative, given a remarkably weaker set of assumptions.

2.3 Changes in Inequality and Selection

There is a large literature on the changes in the wage distribution and the role of changes in the returns to skills that drive these changes. One concern is that if one compares the wage distribution at two points in time, any differences may be partly or wholly due to differences in the composition of the workforce. Blundell, Gosling, Ichimura, and Meghir (2007, BGHM) investigate this using bounds. They study changes in the wage distribution in the United Kingdom for both men and women. Even for men at prime employment ages employment in the late nineties is less than 0.90, down from 0.95 in the late seventies. The concern is that the 10% who do not work are potentially different, both from those who work, as well as from those who did not work in the seventies, corrupting comparisons between the wage distributions in both years. Traditionally such concerns may have been ignored by implicitly assuming that the wages for those not working are similar to those who are working, possibly conditional on some observed covariates, or they may have been addressed by using selection models. The type of selection models used ranges from very parametric models of the type originally developed by Heckman (1978), to semi- and non-parametric versions of this (Heckman, 1990). The concern that BGHM raise is that those selection models rely on assumptions that are difficult to motivate by economic theory. They investigate what can be learned about the changes in the wage distributions without the final, most controversial assumptions of those selection models.

BGHM focus on the interquartile range as their measure of dispersion in the wage distribution. As discussed in Section 2.1, this is convenient, because bounds on quantiles often exist in the presence of missing data. Let $F_{Y|X}(y|x)$ be the distribution of wages condi-

tional on some characteristics X. This is assumed to be well defined irrespective of whether an individual works or not. However, if an individual does not work, Y_i is not observed. Let D_i be an indicator for employment. Then we can estimate the conditional wage distribution given employment, $F_{Y|X,D}(y|x,d=1)$, as well as the probability of employment, $p(x) = \text{pr}(D_i = 1|X_i = x)$. This gives us tight bounds on the (unconditional on employment) wage distribution

$$F_{Y|X,D}(y|x,d=1)\cdot p(x) \leq F_{Y|X,D}(y|x,d=1) \leq F_{Y|X,D}(y|x,d=1)\cdot p(x) + (1-p(x)).$$

We can convert this to bounds on the τ quantile of the conditional distribution of Y_i given $X_i = x$, denoted by $q_\tau(x)$:

$$q_{(\tau-(1-p(x)))/p(x)}(Y_i|D_i=1) \leq q_\tau(x) \leq q_{\tau/p(x)}(Y_i|D_i=1),$$

Then this can be used to derive bounds on the interquartile range $q_{0.75}(x) - q_{0.25}(x)$:

$$q_{(0.75-(1-p(x)))/p(x)}(Y_i|D_i=1) - q_{0.25/p(x)}(Y_i|D_i=1)$$

$$\leq q_{0.75}(x) - q_{0.25}(x) \leq$$

$$q_{(0.25-(1-p(x)))/p(x)}(Y_i|D_i=1) - q_{0.75/p(x)}(Y_i|D_i=1).$$

So far this is just an application of the missing data bounds derived in the previous section. What makes this more interesting is the use of additional information short of imposing a full selection model that would point identify the interquartile range. The first assumption BGHM add is that of stochastic dominance of the wage distribution for employed individuals:

$$F_{Y|X,D}(y|x,d=1) \leq F_{Y|X,D}(y|x,d=0).$$

One can argue with this stochastic dominance assumption, but within groups homogenous in background characteristics including education, it may be reasonable. This assumption tightens the bounds on the distribution function to:

$$F_{Y|X,D}(y|x,d=1) \leq F_{Y|X,D}(y|x,d=1) \leq$$

$$F_{Y|X,D}(y|x,d=1)\cdot p(x) + (1-p(x)).$$

Another assumption BGHM consider is a modification of an instrumental variables assumption that an observed covariate Z is excluded from the wage distribution:

$$F_{Y|X,Z}(y|X=x,Z=z) = F_{Y|X,Z}(y|X=x,Z=z'), \qquad \text{for all } x,z,z'.$$

This changes the bounds on the distribution function to:

$$\max_z F_{Y|X,Z,D}(y|x,z,d=1)\cdot p(x,z)$$

$$\leq F_{Y|X,D}(y|x) \leq$$

$$\min_z F_{Y|X,Z,D}(y|x,z,d=1)\cdot p(x) + (1-p(x)).$$

(An alternative weakening of the standard instrumental variables assumption is in Hotz, Mullin and Sanders (1997), where a valid instrument exists, but is not observed directly.)

Such an instrument may be difficult to find, and BGHM argue that it may be easier to find a covariate that affects the wage distribution in one direction, using a monotone instrumental variables restriction suggested by Manski and Pepper (2000):

$$F_{Y|X,Z}(y|X=x,Z=z) \leq F_{Y|X,Z}(y|X=x,Z=z'), \qquad \text{for all } x,z<z'.$$

This discussion is somewhat typical of what is done in empirical work in this area. A number of assumptions are considered, with the implications for the bounds investigated. The results lay out part of the mapping between the assumptions and the bounds.

2.4 Random Effects Panel Data Models with Initial Condition Problems

Honoré and Tamer (2006) study dynamic random effects panel data models. We observe $(X_{i1}, Y_{i1}, \ldots, X_{iT}, Y_{iT})$, for $i = 1, \ldots, N$. The time dimension T is small relative to the cross-section dimension N. Large sample approximations are based on fixed T and large N. Inference would be standard if we specified a parametric model for the (components of the) conditional distribution of $(Y_{i1}, \ldots, Y_{iT})$ given $(X_{i1}, \ldots, X_{iT})$. In that case we could use maximum likelihood methods. However, it is difficult to specify this conditional distribution directly. Often we start with a model for the evolution of Y_{it} in terms of the present and past covariates and its lags. As an example, consider the model

$$Y_{it} = 1\{X_{it}'\beta + Y_{it-1}\gamma + \alpha_i + \epsilon_{it} \geq 0\},$$

with the ϵ_{it} independent over time and individuals, and normally distributed, $\epsilon_{it} \sim \mathcal{N}(0,1)$. The object of interest is the parameter governing the dynamics, γ. This model gives us the conditional distribution of $Y_{i2}, \ldots, Y_{iT}$ given Y_{i1}, α_i and given $X_{i1}, \ldots, X_{iT}$. Suppose we also postulate a parametric model for the random effects α_i:

$$\alpha | X_{i1}, \ldots, X_{iT} \sim G(\alpha|\theta),$$

(so in this case α_i is independent of the covariates). Then the model is (almost) complete, in the sense that we can almost write down the conditional distribution of $(Y_{i1}, \ldots, Y_{iT})$ given $(X_{i1}, \ldots, X_{iT})$. All that is missing is the conditional distribution of the initial condition:

$$p(Y_{i1}|\alpha_i, X_{i1}, \ldots, X_{iT}).$$

This is a difficult distribution to specify. One could directly specify this distribution, but one might want it to be internally consistent across different number of time periods, and that makes it awkward to choose a functional form. See for general discussions of this initial conditions problem Wooldridge (2002). Honoré and Tamer investigate what can be learned about γ without making parametric assumptions about this distribution. From the literature

it is known that in many cases γ is not point-identified (for example, the case with $T \leq 3$, no covariates, and a logistic distribution for ϵ_{it}). Nevertheless, it may be that the range of values of γ consistent with the data is very small, and it might reveal the sign of γ.

Honoré and Tamer study the case with a discrete distribution for α, with a finite and known set of support points. They fix the support to be $-3, -2.8, \ldots, 2.8, 3$, with unknown probabilities. Given that the ϵ_{it} are standard normal, this is very flexible. In a computational exercise they assume that the true probabilities make this discrete distribution mimic the standard normal distribution. In addition they set $\Pr(Y_{i1} = 1|\alpha_i) = 1/2$. In the case with $T = 3$ they find that the range of values for γ consistent with the data generating process (the identified set) is very narrow. If γ is in fact equal to zero, the width of the set is zero. If the true value is $\gamma = 1$, then the width of the interval is approximately 0.1. (It is largest for γ close to, but not equal to, -1.) See Figure 1, taken from Honoré and Tamer (2006).

The Honoré-Tamer analysis, in the context of the literature on initial conditions problems, shows very nicely the power of the partial identification approach. A problem that had been viewed as essentially intractable, with many non-identification results, was shown to admit potentially precise inferences despite these non-identification results.

2.5 Auction Data

Haile and Tamer (2003, HT from hereon), in what is one of the most influential applications of the partial identification approach, study English or oral ascending bid auctions. In such auctions bidders offer increasingly higher prices until only one bidder remains. HT focus on a symmetric independent private values model. In auction t, for $t = 1, \ldots, T$, bidder i has a value ν_{it}, drawn independently from the value for bidder j. Large sample results refer to the number of auctions getting large. HT assume that the value distribution is the same in each auction (after adjusting for observable auction characteristics). A key object of interest, is the value distribution. Given that one can derive other interesting objects, such as the optimal reserve price.

One can imagine a set up where the researcher observes, as the price increases, for each

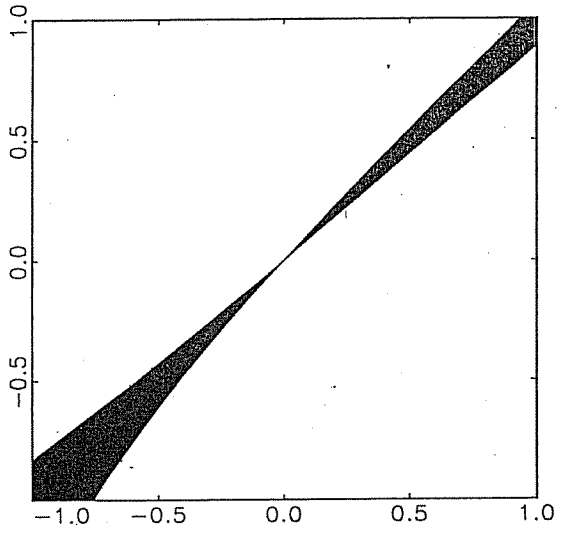


FIGURE 1.—Identified region for γ as a function of its true value.

bidder whether that bidder is still participating in the auction. (Milgrom and Weber (1982) assume that each bidder continuously confirms their participation by holding down a button while prices rise continuously.) In that case one would be able to infer for each bidder their valuation, and thus directly estimate the value distribution.

This is not what is typically observed. Instead of prices rising continuously, there are jumps in the bids, and for each bidder we do not know at any point in time whether they are still participating unless they subsequently make a higher bid. HT study identification in this, more realistic, setting. They assume that no bidder ever bids more than their valuation, and that no bidder will walk away and let another bidder win the auction if the winning bid is lower than their own valuation. Under those two assumptions, HT show that one can derive bounds on the value distribution.

One set of bounds they propose is as follows. Let the highest bid for participant i in auction t be b_{it}. The number of participants in auction t is n_t. Ignoring any covariates, let the distribution of the value for individual i, ν_{it}, be $F_\nu(v)$. This distribution function is the same for all auctions. Let $F_b(b) = \Pr(b_{it} \leq b)$ be the distribution function of the bids (ignoring variation in the number of bidders by auction). This distribution can be estimated because the bids are observed. The winning bid in auction t is $B_t = \max_{i=1,\ldots,n_t} b_{it}$. First HT derive an upper bound on the distribution function $F_\nu(v)$. Because no bidder ever bids more than their value, it follows that $b_{it} \leq \nu_{it}$. Hence, without additional assumptions,

$$F_\nu(v) \leq F_b(v), \qquad \text{for all } v.$$

For a lower bound on the distribution function one can use the fact that the second highest of the values among the n participants in auction t must be less than or equal to the winning bid. This follows from the assumption that no participant will let someone else win with a bid below their valuation. Let $F_{\nu,m:n}(v)$ denote the mth order statistic in a random sample of size n from the value distribution, and let $F_{B,n:n}(b)$ denote the distribution of the

winning bid in auctions with n participants. Then

$$F_{B,n:n}(v) \leq F_{\nu,n-1:n}(v).$$

The distribution of the any order statistic is monotonically related to the distribution of the parent distribution, and so a lower bound on $F_{\nu,n-1:n}(v)$ implies a lower bound on $F_\nu(v)$.

HT derive tighter bounds based on the information in other bids and the inequalities arising from the order statistics, but the above discussion illustrates the point that outside of the Milgrom-Weber button auction model one can still derive bounds on the value distribution in an English auction even if one cannot point-identify the value distribution. If in fact the highest bid for each individual was equal to their value (other than for the winner for whom the bid is equal to the second highest value), the bounds would collaps and point-identification would be obtained.

2.6 Entry Models and Inequality Conditions

Recently a number of papers has studied entry models in settings with multiple equilibria. In such settings traditionally researchers have added *ad hoc* equilbrium selection mechanisms. In the recent literature a key feature is the avoidance of such assumptions, as these are often difficult to justify on theoretical grounds. Instead the focus is on what can be learned in the absence of such assumptions. In this section I will discuss some examples from this literature. An important feature of these models is that they often lead to inequality restrictions, where the parameters of interest θ satisfy

$$\mathbb{E}[\psi(Z,\theta)] \geq 0,$$

for known $\psi(z,\theta)$. This relates closely to the standard (Hansen, 1983) generalized method of moments (GMM) set up where the functions $\psi(Z,\theta)$ would have expectation equal to zero at the true values of the parameters. We refer to this as the generalized inequality restrictions (GIR) form. These papers include Pakes, Porter, Ho, and Ishii (2006), Cilberto and Tamer (2004, CM from hereon), Andrews, Berry and Jia (2004). Here I will discuss a simplified

version of the CM model. Suppose two firms, A and B, contest a set of markets. In market m, $m = 1, \ldots, M$, the profits for firms A and B are

$$\pi_{Am} = \alpha_A + \delta_A \cdot d_{Bm} + \varepsilon_{Am}, \qquad \text{and} \quad \pi_{Bm} = \alpha_B + \delta_B \cdot d_{Am} + \varepsilon_{Bm}.$$

where $d_{Fm} = 1$ if firm F is present in market m, for $F \in \{A, B\}$, and zero otherwise. The more realistic model CM consider also includes observed market and firm characteristics. Firms enter market m if their profits in that market are positive. Firms observe all components of profits, including those that are unobserved to the econometrician, $(\varepsilon_{Am}, \varepsilon_{Bm})$, and so their decisions satisfy:

$$d_{Am} = 1\{\pi_{Am} \geq 0\}, \qquad d_{Bm} = 1\{\pi_{Bm} \geq 0\}. \tag{1}$$

(Pakes, Porter, Ho, and Ishii allow for incomplete information where expected profits are at least as high for the action taken as for actions not taken, given some information set.) The unobserved (to the econometrician) components of profits, ε_{Fm}, are independent accross markets and firms. For ease of exposition we assume here that they have a normal $\mathcal{N}(0, 1)$ distribution. (Note that we only observe indicators of the sign of profits, so the scale of the unobserved components is not relevant for predictions.) The econometrician observes in each market only the pair of indicators d_A and d_B. We focus on the case where the effect of entry of the other firm on a firm's profits, captured by the parameters δ_A and δ_B is negative, which is the case of most economic interest.

An important feature of this model is that given the parameters $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, for a given set of $(\varepsilon_{Am}, \varepsilon_{Bm})$ there is not necessarily a unique solution (d_{Am}, d_{Bm}). For pairs of values $(\varepsilon_{Am}, \varepsilon_{Bm})$ such that

$$-\alpha_A < \varepsilon_A \leq -\alpha_A - \delta_A, \qquad -\alpha_B < \varepsilon_B \leq -\alpha_B - \delta_B,$$

both $(d_A, d_B) = (0, 1)$ and $(d_A, d_B) = (1, 0)$ satisfy the profit maximization condition (1). In the terminology of this literature, the model is not *complete*. It does not specify the

outcomes given the inputs. Figure 1, adapted from CM, shows the different regions in the $(\varepsilon_{Am}, \varepsilon_{Bm})$ space.

The implication of this is that the probability of the outcome $(d_{Am}, d_{Bm}) = (0, 1)$ cannot be written as a function of the parameters of the model, $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, even given distributional assumptions on $(\varepsilon_{Am}, \varepsilon_{Bm})$. Instead the model implies a lower and upper bound on this probability:

$$H_{L,01}(\theta) \leq \Pr\left((d_{Am}, d_{Bm}) = (0, 1)\right) \leq H_{U,01}(\theta).$$

Inspecting Figure 1 shows that

$$H_{L,01}(\theta) = \Pr(\varepsilon_{Am} < -\alpha_A, -\alpha_B < \varepsilon_{Bm})$$

$$+\Pr(-\alpha_A \leq \varepsilon_{Am} < -\alpha_A - \delta_A, -\alpha_B - \delta_B < \varepsilon_{Bm}),$$

and

$$H_{U,01}(\theta) = \Pr(\varepsilon_{Am} < -\alpha_A, \alpha_B < \varepsilon_{Bm})$$

$$+\Pr(-\alpha_A \leq \varepsilon_{Am} < -\alpha_A - \delta_A, -\alpha_B - \delta_B < \varepsilon_{Bm}),$$

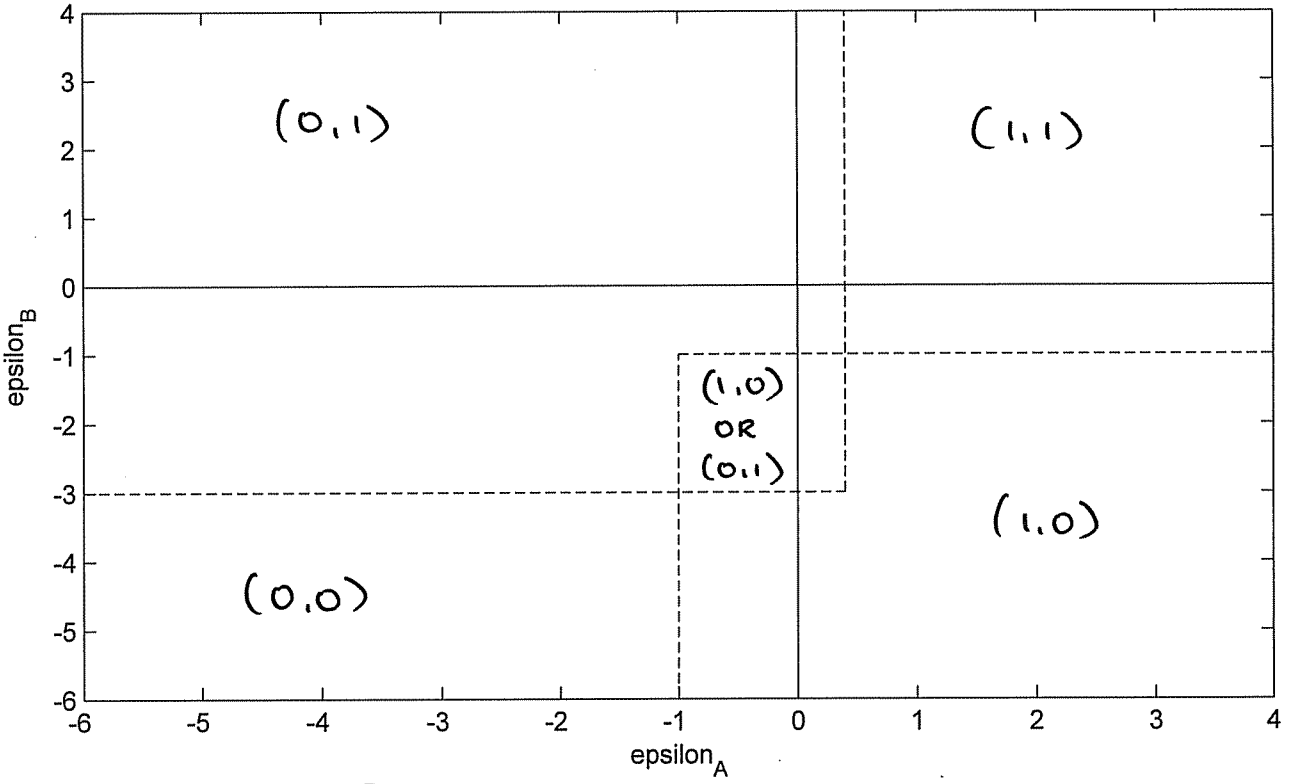
$$+\Pr(-\alpha_A \leq \varepsilon_{Am} < -\alpha_A - \delta_A, -\alpha_B < \varepsilon_{Bm} < -\alpha_B - \delta_B),$$

Similar expressions can be derived for the probability $\Pr\left((d_{Am}, d_{Bm}) = (1, 0)\right)$. Thus in general we can write the information about the parameters in large samples as

$$\begin{pmatrix} H_{L,00}(\theta) \\ H_{L,01}(\theta) \\ H_{L,10}(\theta) \\ H_{L,11}(\theta) \end{pmatrix} \leq \begin{pmatrix} \Pr\left((d_{Am}, d_{Bm}) = (0, 0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (0, 1)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1, 0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1, 1)\right) \end{pmatrix} \leq \begin{pmatrix} H_{U,00}(\theta) \\ H_{U,01}(\theta) \\ H_{U,11}(\theta) \\ H_{U,11}(\theta) \end{pmatrix}.$$

(For $(d_A, d_B) = (0, 0)$ or $(d_A, d_B) = (1, 1)$ the lower and upper bound coincide, but for ease of exposition we treat all four configurations symmetrically.) The $H_{L,ij}(\theta)$ and $H_{U,ij}(\theta)$ are

Figure 1 (d_A, d_B)



$\alpha_A = 1$ $\delta_A = -1.4$

$\alpha_B = 3$ $\delta_B = -2$

known functions of θ. The data allow us to estimate the foru probabilities, which contain only three separate pieces of information because the probabilities add up to one. Given these probabilities, the identified set is the set of all θ that satisfy all eight inequalities. In the simple model above, there are four parameters. Even in the case with the lower and upper bounds for the probabilities coinciding, these would in general not be identified.

We can write this in the GIR form by defining

$$\psi(d_A, d_B|\alpha_A, \alpha_B, \delta_A, \delta_B) = \begin{pmatrix} H_{U,00}(\theta) - (1-d_A)\cdot(1-d_B) \\ (1-d_A)\cdot(1-d_B) - H_{L,00}(\theta) \\ H_{U,01}(\theta) - (1-d_A)\cdot d_B \\ (1-d_A)\cdot d_B - H_{L,01}(\theta) \\ H_{U,10}(\theta) - d_A\cdot(1-d_B) \\ d_A\cdot(1-d_B) - H_{L,10}(\theta) \\ H_{U,11}(\theta) - d_A\cdot d_B \\ d_A\cdot d_B - H_{L,11}(\theta) \end{pmatrix},$$

so that the model implies that at the true values of the parameters

$$\mathbb{E}\left[\psi(d_A, d_B|\alpha_A, \alpha_B, \delta_A, \delta_B)\right] \geq 0.$$

3. Estimation

Chernozhukov, Hong, and Tamer (2007, CHT) consider, among other things, the case with moment inequality conditions,

$$\mathbb{E}[\psi(Z,\theta)] \geq 0,$$

where $\psi(z,\theta)$ is a known vector of functions, of dimension M, and the unknown parameter θ is of dimension K. Let Θ be the parameter space, a subset of $\mathbb{R}^K$.

Define for a vector x the vector $(x)_+$ to be the component-wise non-negative part, and $(x)_-$ to be the component-wise non-positive part, so that for all x, $x = (x)_- + (x)_+$. For a given $M \times M$ non-negative definite weight matrix W, CHT consider the population objective function

$$Q(\theta) = \mathbb{E}[\psi(Z,\theta)]'_- W \mathbb{E}[\psi(Z,\theta)]_-.$$

For all θ in the identified set, denoted by $\Theta_I \subset \Theta$, we have $Q(\theta) = 0$.

The sample equivalent to this population objective function is

$$Q_N(\theta) = \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)_-' W \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)_- .$$

We cannot simply estimate the identified set as

$$\tilde{\Theta}_I = \{ \theta \in \Theta \,|\, Q_N(\theta) = 0 \} ,$$

The reason is that even for θ in the identified set $Q_N(\theta)$ may be positive with high probability. A simple way to see that is to consider the standard GMM case with equalities and over-identification. If $\mathbb{E}[\psi(Z, \theta)] = 0$, the objective function will not be zero in finite samples in the case with over-identification. As a result, $\tilde{\Theta}_I$ can be empty when Θ_I is not, even in large samples.

This is the reason CHT estimate the set Θ_I as

$$\hat{\Theta}_I = \{ \theta \in \Theta \,|\, Q_N(\theta) \leq a_N \} ,$$

where $a_N \to 0$ at the appropriate rate. In most regular problems $a_N = c/N$, leading to an estimator $\hat{\Theta}_I$ that is consistent for Θ_I, by which we mean that the two sets get close to each other, in the Hausdorf sense that

$$\sup_{\theta \in \Theta_I} \inf_{\theta' \in \hat{\Theta}_I} d(\theta, \theta') \longrightarrow 0, \qquad \text{and} \quad \sup_{\theta' \in \hat{\Theta}_I} \inf_{\theta \in \Theta_I} d(\theta, \theta') \longrightarrow 0,$$

where $d(\theta, \theta') = ((\theta - \theta)'(\theta - \theta'))^{1/2}$.

3. Inference: General Issues

There is a rapidly growing literature concerned with developing methods for inference in partially identified models, including Beresteanu and Molinari (2006), Chernozhukov, Hong, and Tamer (2007), Imbens and Manski (2004), Rosen (2006), and Romano and Shaikh

(2007ab). In many cases the partially identified set itself is difficult to characterize. In the scalar case this can be much simpler. There it often is an interval, $[\theta_{\rm LB}, \theta_{\rm UB}]$. There are by now a number of proposals for constructing confidence sets. They differ in implementation as well as in their goals. One issue is whether one wants a confidence set that includes each element of the identified set with fixed probability, or the entire identified set with that probability. Formally, the first question looks for a confidence set $\mathrm{CI}_\alpha^\theta$ that satisfies

$$\inf_{\theta\in[\theta_{\rm LB},\theta_{\rm UB}]} \Pr\left(\theta \in \mathrm{CI}_\alpha^\theta\right) \geq \alpha.$$

In the second case we look for a set $\mathrm{CI}_\alpha^{[\theta_{\rm LB},\theta_{\rm UB}]}$ such that

$$\Pr\left([\theta_{\rm LB}, \theta_{\rm UB}] \subset \mathrm{CI}_\alpha^\theta\right) \geq \alpha.$$

The second requirement is stronger than the first, and so generally $\mathrm{CI}_\alpha^\theta \subset \mathrm{CI}_\alpha^{[\theta_{\rm LB},\theta_{\rm UB}]}$. Here we follow Imbens and Manski (2004) and Romano and Shaikh (2007a) who focus on the first case. This seems more in line with teh traditional view of confidence interval in that they should cover the true value of the parameter with fixed probability. It is not clear why the fact that the object of interest is not point-identified should change the definition of a confidence interval. CHT and Romano and Shaikh (2007b) focus on the second case.

Next we discuss two specific examples to illustrate some of the issues that can arise, in particular the uniformity of confidence intervals.

3.1 Inference: A Missing Data Problem

Here we continue the missing data example from Section 2.1. We have a random sample of $(W_i, W_i \cdot Y_i)$, for $i = 1, \ldots, N$. Y_i is known to lie in the interval $[0, 1]$, interest is in $\theta = \mathbb{E}[Y]$, and the parameter space is $\Theta = [0, 1]$. Define $\mu_1 = \mathbb{E}[Y|W = 1]$, $\lambda = \mathbb{E}[Y|W = 0]$, $\sigma^2 = \mathbb{V}(Y|W = 1)$, and $p = \mathbb{E}[W]$. For ease of exposition we assume p is known. The identified set is

$$\Theta_I = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)].$$

Imbens and Manski (2004) discuss confidence intervals for this case. The key feature of this problem, and similar ones, is that the lower and upper bounds are well-behaved functionals of the joint distribution of the data that can be estimated at the standard parametric $\sqrt{N}$ rate with an asymptotic normal distribution. In this specific example the lower and upper bound are both functions of a single unknown parameter, the conditional mean μ_1. The first step is a 95% confidence interval for μ_1. Let $N_1 = \sum_i W_i$ and $\overline{Y}_1 = \sum_i W_i \cdot Y_i / N_1$. The standard confidence interval is

$$\text{CI}_\alpha^{\mu_1} = \left[\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1}, \overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right].$$

Consider the confidence interval for the lower and upper bound:

$$\text{CI}_\alpha^{p\cdot\mu_1} = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right) \right],$$

and

$$\text{CI}_\alpha^{p\cdot\mu_1+(1-p)} = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1} \right) + (1-p), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right].$$

A simple and valid confidence interval can be based on the lower confidence bound on the lower bound and the upper confidence bound on the upper bound:

$$\text{CI}_\alpha^{\theta} = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right].$$

This is generally conservative. For each θ in the interior of Θ_I, the asymptotic coverage rate is 1. For $\theta \in \{\theta_{\text{LB}}, \theta_{\text{UB}}\}$, the coverage rate is $\alpha + (1-\alpha)/2$.

The interval can be modified to give asymptotic coverage equal to α by changing the quantiles used in the confidence interval construction, essentially using one-sided critical values,

$$\text{CI}_\alpha^{\theta} = \left[p \cdot \left(\overline{Y} - 1.645 \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + 1.645 \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right].$$

This has the problem that if $p = 0$ (when θ is point-identified), the coverage is only $\alpha-(1-\alpha)$. In fact, for values of p close to zero, the confidence interval would be shorter than the confidence interval in the point-identified case. Imbens and Manski (2004) suggest modifying the confidence interval to

$$\text{CI}^{\theta}_{\alpha} = \left[p \cdot \left(\overline{Y} - C_N \cdot \sigma/\sqrt{N}_1 \right), p \cdot \left(\overline{Y} + C_N \cdot \sigma/\sqrt{N}_1 \right) + 1 - p \right],$$

where the critical value C_N satisfies

$$\Phi\left(C_N + \sqrt{N} \cdot \frac{1-p}{\sigma/\sqrt{p}} \right) - \Phi\left(-C_N\right) = \alpha.$$

and $C_N = 1.96$ if $p = 0$. This confidence interval has asymptotic coverage 0.95, uniformly over p.

3.2. Inference: Multiple Inequalities

Here we look at inference in the Genereralized Inequality (GIR) setting. The example is a simplified version of the moment inequality type of problems discussed in CHT, Romano and Shaikh (2007ab), Pakes, Porter, Ho, and Ishii (2006), and Andrews and Guggenberger (2007). Suppose we have two moment inequalities,

$$\mathbb{E}[X] \geq \theta, \qquad \text{and} \quad \mathbb{E}[Y] \geq \theta.$$

The parameter space is $\Theta = [0, \infty)$. Let $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$. We have a random sample of size N of the pairs (X, Y). The identified set is

$$\Theta_I = [0, \min(\mu_X, \mu_Y)].$$

The key difference with the previous example is that the upper bound is no longer a smooth, well-behaved functional of the joint distribution. In the simple two-inequality example, if μ_X is close to μ_Y, the distribution of the estimator for the upper bound is not well approximated by a normal distribution. Suppose we estimate the means of X and Y by

$\overline{X}$, and $\overline{Y}$, and that the variances of X and Y are known to be equal to σ^2. A naive 95% confidence interval would be

$$C_\alpha^\theta = [0, \min(\overline{X}, \overline{Y}) + 1.645 \cdot \sigma/N].$$

This confidence interval essentially ignores the moment inequality that is not binding in the sample. It has asymptotic 95% coverage for all values of μ_X, μ_Y, as long as $\min(\mu_X, \mu_Y) > 0$, and $\mu_X \neq \mu_Y$. The first condition ($\min(\mu_X, \mu_Y) > 0$) is the same as the condition in the Imbens-Manski example. It can be dealt with in the same way by adjusting the critical value slightly based on an initial estimate of the width of the identified set.

The second condition raises a different uniformity concern. The naive confidence interval essentially assumes that the researcher knows which moment conditions are binding. This is true in large samples, unless there is a tie. However, in finite samples ignoring uncertainty regarding the set of binding moment inequalities may lead to a poor approximation, especially if there are many inequalities. One possibility is to construct conservative confidence intervals (e.g., Pakes, Porter, Ho, and Ishii, 2007). However, such intervals can be unnecessarily conservative if there are moment inequalities that are far from binding.

One would like construct confidence intervals that asymptotically ignore irrelevant inequalities, and at the same time are valid uniformly over the parameter space. Bootstrapping is unlikely to work in this setting. One way of obtaining confidence intervals that are uniformly valid is based on subsampling. See Romano and Shaikh (2007a), and Andrews and Guggenberger (2007). Little is known about finite sample properties in realistic settings.

References

Andrews, D., S. Berry, and P. Jia (2004), "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location,"unpublished manuscript, Department of Economics, Yale University.

Andrews, D., and P. Guggenberger (2004), "The Limit of Finite Sample Size and a Problem with Subsampling,"unpublished manuscript, Department of Economics, Yale University.

Aradillas-Lopez, A., and E. Tamer (2007), "The Identification Power of Equilibrium in Games,"unpublished manuscript, Department of Economics, Princeton University.

Balke, A., and J. Pearl, (1997), "Bounds on Treatment Effects from Studies with Imperfect Compliance,"*Journal of the American Statistical Association*, 92: 1172-1176.

Beresteanu, A., and F. Molinari, (2006), "Asymptotic Properties for a Class of Partially Identified Models,"Unpublished Manuscript, Department of Economics, Cornell University.

Blundell, R., M. Browning, and I. Crawford, (2007), "Best Nonparametric Bounds on Demand Responses,"Cemmap working paper CWP12/05, Department of Economics, University College London.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir, (2007), "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds,"*Econometrica*, 75(2): 323-363.

Chernozhukov, V., H. Hong, and E. Tamer (2007), "Estimation and Confidence Regions for Parameter Sets in Econometric Models,"forthcoming, *Econometrica*.

Ciliberto, F., and E. Tamer (2004), "Market Structure and Multiple Equilibria in Airline Markets,"Unpublished Manuscript.

Haile, P., and E. Tamer (2003), "Inference with an Incomplete Model of English Auctions,"*Journal of Political Economy*, Vol 111(1), 1-51.

Heckman, J., (1978), "Dummy Endogenous Variables in a Simultaneous Equations

System", *Econometrica*, Vol. 46, 931–61.

Heckman, J. J. (1990), "Varieties of Selection Bias," *American Economic Review* 80, 313-318.

Honoré, B., and E. Tamer (2006), "Bounds on Parameters in Dynamic Discrete Choice Models," *Econometrica*, 74(3): 611-629.

Hotz, J., C. Mullin, and S. Sanders, (1997), "Boudning Causal Effects Using Data from a Contaminated Natural Experiment: Analysing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64(4), 575-603.

Imbens, G., and C. Manski (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 74(6): 1845-1857.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (1995), *Identification Problems in the Social Sciences*, Cambridge, Harvard University Press.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., and J. Pepper, (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(): 997-1010.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Milgrom, P, and R. Weber (1982), "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(3): 1089-1122.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006), "Moment Inequalities and Their Application,"Unpublished Manuscript.

Robins, J., (1989), "The Analysis of Randomized and Non-randomized AIDS Trials Using a New Approach to Causal Inference in Longitudinal Studies,"in *Health Service Research*

Methodology: A Focus on AIDS, (Sechrest, Freeman, and Mulley eds), US Public Health Service, 113-159.

ROMANO, J., AND A. SHAIKH (2006a), "Inference for Partially Identified Parameters,"Unpublished Manuscript, Stanford University.

ROMANO, J., AND A. SHAIKH (2006b), "Inference for Partially Identified Sets,"Unpublished Manuscript, Stanford University.

ROSEN, A., (2005), "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," Unpublished Manuscript, Department of Economics, University College London.

WOOLDRIDGE, J (2002), "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity,"*Journal of Applied Econometrics*, 20, 39-54.

What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 10, **Tuesday**, **July 31st**, **4.30-5.30 pm**

Difference-in-Differences Estimation

These notes provide an overview of standard difference-in-differences methods that have been used to study numerous policy questions. We consider some recent advances in Hansen (2007a,b) on issues of inference, focusing on what can be learned with various group/time period dimensions and serial independence in group-level shocks. Both the repeated cross sections and panel data cases are considered. We discuss recent work by Athey and Imbens (2006) on nonparametric approaches to difference-in-differences, and Abadie, Diamond, and Hainmueller (2007) on constructing synthetic control groups.

1. Review of the Basic Methodology

Since the work by Ashenfelter and Card (1985), the use of difference-in-differences methods has become very widespread. The simplest set up is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. In the case where the same units within a group are observed in each time period, the average gain in the second (control) group is substracted from the average gain in the first (treatment) group. This removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between those groups, as well as biases from comparisons over time in the treatment group that could be the result of trends. We will treat the panel data case in Section 4.

With repeated cross sections, we can write the model for a generic member of any of groups as

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u \tag{1.1}$$

where y is the outcome of interest, $d2$ is a dummy variable for the second time period. The dummy variable dB captures possible differences between the treatment and control groups prior to the policy change. The time period dummy, $d2$, captures aggregate factors that would cause changes in y even in the absense of a policy change. The coefficient of interest, δ_1, multiplies the interaction term, $d2 \cdot dB$, which is the same as a dummy variable equal to one for those observations in the treatment group in the second period. The difference-in-differences estimate is

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}). \tag{1.2}$$

Inference based on even moderate sample sizes in each of the four groups is straightforward, and is easily made robust to different group/time period variances in the regression framework.

In some cases a more convincing analysis of a policy change is available by further refining the definition of treatment and control groups. For example, suppose a state implements a change in health care policy aimed at the elderly, say people 65 and older, and the response variable, *y*, is a health outcome. One possibility is to use data only on people in the state with the policy change, both before and after the change, with the control group being people under 65 and the treatment group being people 65 and older. The potential problem with this DD analysis is that other factors unrelated to the state's new policy might affect the health of the elderly relative to the younger population, for example, changes in health care emphasis at the federal level. A different DD analysis would be to use another state as the control group and use the elderly from the non-policy state as the control group. Here, the problem is that *changes* in the health of the elderly might be systematically different across states due to, say, income and wealth differences, rather than the policy change.

A more robust analysis than either of the DD analyses described above can be obtained by using both a different state and a control group within the treatment state. If we again label the two time periods as one and two, let *B* represent the state implementing the policy, and let *E* denote the group of elderly, then an expanded verson of (1.1) is

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB \cdot dE + \delta_0 d2 + \delta_1 d2 \cdot dB + \delta_2 d2 \cdot dE + \delta_3 d2 \cdot dB \cdot dE + u \quad (1.3)$$

The coefficient of interest is now δ_3, the coefficient on the triple interaction term, $d2 \cdot dB \cdot dE$. The OLS estimate $\hat{\delta}_3$ can be expressed as follows:

$$\hat{\delta}_3 = (\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1}) \quad (1.4)$$

where the *A* subscript means the state not implementing the policy and the *N* subscript represents the non-elderly. For obvious reasons, the estimator in (1.4) is called the *difference-in-difference-in-differences (DDD)* estimate. [The population analog of (1.4) is easily established from (1.3) by finding the expected values of the six groups appearing in (1.4).] If we drop either the middle term or the last term, we obtain one of the DD estimates described in the previous paragraph. The DDD estimate starts with the time change in averages for the elderly in the treatment state and then nets out the change in means for elderly in the control state and the change in means for the non-elderly in the treatment state. The hope is that this controls for two kinds of potentially confounding trends: changes in health status of

elderly across states (that would have nothing to do with the policy) and changes in health status of all people living in the policy-change state (possibly due to other state policies that affect everyone's health, or state-specific changes in the economy that affect everyone's health). When implemented as a regression, a standard error for $\hat{\delta}_3$ is easily obtained, including a heteroskedasticity-robust standard error. As in the DD case, it is straightforward to add additional covariates to (1.3) and inference robust to heteroskedasticity.

2. How Should We View Uncertainty in DD Settings?

The standard approach just described assumes that all uncertainty in inference enters through sampling error in estimating the means of each group/time period combination. This approach has a long history in statistics, as it is equivalent to analysis of variance. Recently, different approaches have been suggest that focus on different kinds of uncertainty – perhaps in addition to sampling error in estimating means. Recent work by Bertrand, Duflo, and Mullainathan (2004), Donald and Lang (2007), Hansen (2007a,b), and Abadie, Diamond, and Hainmueller (2007) argues for additional sources of uncertainty. In fact, in most cases the additional uncertainty is assumed to swamp the sampling error in estimating group/time period means. We already discussed the DL approach in the cluster sample notes, although we did not explicitly introduce a time dimension. One way to view the uncertainty introduced in the DL framework – and a perspective explicitly taken by ADH – is that our analysis should better reflect the uncertainty in the quality of the control groups.

Before we turn to a general setting, it is useful to ask whether introducing more than sampling error into DD analyses is necessary, or desirable. As we discussed in the cluster sample notes, the DL approach does not allow inference in the basic comparison-of-mean case for two groups. While the DL estimate is the usual difference in means, the error variance of the cluster effect cannot be estimated, and the *t* distribution is degenerate. It is also the case that the DL approach cannot be applied to the standard DD or DDD cases covered in Section 1. We either have four different means to estimate or six, and the DL regression in these cases produces a perfect fit with no residual variance. Should we conclude nothing can be learned in such settings?

Consider the example from Meyer, Viscusi, and Durbin (1995) on estimating the effects of benefit generosity on length of time a worker spends on workers' compensation. MVD have a before and after period, where the policy change was to raise the cap on covered earnings. The treatment group is high earners, and the control group is low earners – who should not have

been affected by the change in the cap. Using the state of Kentucky and a total sample size of 5,626, MVD find the DD estimate of the policy change is about 19.2% (longer time on workers' compensation). The *t* statistic is about 2.76, and the estimate changes little when some controls are added. MVD also use a data set for Michigan. Using the same DD approach, they estimate an almost identical effect: 19.1%. But, with "only" 1,524 observations, the *t* statistic is 1.22. It seems that, in this example, there is plenty of uncertainty in estimation, and one cannot obtain a tight estimate without a fairly large sample size. It is unclear what we gain by concluding that, because we are just identifying the parameters, we cannot perform inference in such cases. In this example, it is hard to argue that the uncertainty associated with choosing low earners within the same state and time period as the control group somehow swamps the sampling error in the sample means.

3. General Settings for DD Analysis: Multiple Groups and Time Periods

The DD and DDD methodologies can be applied to more than two time periods. In the first case, a full set of time-period dummies is added to (1.1), and a policy dummy replaces $d2 \cdot dB$; the policy dummy is simply defined to be unity for groups and time periods subject to the policy. This imposes the restriction that the policy has the same effect in every year, and assumption that is easily relaxed. In a DDD analysis, a full set of dummies is included for each of the two kinds of groups and all time periods, as well as all pairwise interactions. Then, a policy dummy (or sometimes a continuous policy variable) measures the effect of the policy. See Gruber (1994) for an application to mandated maternity benefits.

With many time periods and groups, a general framework considered by BDM (2004) and Hansen (2007b) is useful. The equation at the individual level is

$$y_{igt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + v_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{3.1}$$

where i indexes individual, g indexes group, and t indexes time. This model has a full set of time effects, λ_t, a full set of group effects, α_g, group/time period covariates, x_{gt} (these are the policy variables), individual-specific covariates, $\mathbf{z}_{igt}$, unobserved group/time effects, v_{gt}, and individual-specific errors, u_{igt}. We are interested in estimating $\boldsymbol{\beta}$. Equation (3.1) is an example of a *multilevel model*.

One way to write (3.1) that is useful is

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\boldsymbol{\gamma}_{gt} + u_{igt}, \; i = 1, \ldots, M_{gt}, \tag{3.2}$$

which shows a model at the individual level where both the intercepts and slopes are allowed to differ across all (g,t) pairs. Then, we think of δ_{gt} as

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}. \tag{3.3}$$

Equation (3.3) is very useful, as we can think of it as a regression model at the group/time period level.

As discussed by BDM, a common way to estimate and perform inference in (3.1) is to ignore v_{gt}, in which case the observations at the individual level are treated as independent. When v_{gt} is present, the resulting inference can be very misleading. BDM and Hansen (2007b) allow serial correlation in $\{v_{gt} : t = 1,2,\ldots,T\}$ and assume independence across groups, g.

A simple way to proceed is to view (3.3) as ultimately of interest. We observe $\mathbf{x}_{gt}$, λ_t is handled with year dummies,and α_g just represents group dummies. The problem, then, is that we do not observe δ_{gt}. But we can use the individual-level data to estimate the δ_{gt}, provided the group/time period sizes, M_{gt}, are reasonably large. With random sampling within each (g,t), the natural estimate of δ_{gt} is obtained from OLS on (3.2) for each (g,t) pair, assuming that $E(\mathbf{z}_{igt}'u_{igt}) = \mathbf{0}$. (In most DD applications, this assumption almost holds by definition, as the individual-specific controls are included to improve estimation of δ_{gt}.) If a particular model of heteroskedasticity suggests itself, and $E(u_{it}|\mathbf{z}_{igt}) = 0$ is assumed, then a weighted least squares procedure can be used. Sometimes one wishes to impose some homogeneity in the slopes – say, $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}_g$ or even $\boldsymbol{\gamma}_{gt} = \boldsymbol{\gamma}$ – in which case pooling can be used to impose such restrictions. In any case, we proceed as if the M_{gt} are large enough to ignore the estimation error in the $\hat{\delta}_{gt}$; instead, the uncertainty comes through v_{gt} in (3.3). Hansen (2007b) considers adjustments to inference that accounts for sampling error in the $\hat{\delta}_{gt}$, but the methods are more complicated. The minimum distance approach we discussed in the cluster sampling notes, applied in the current context, effectively drops v_{gt} from (3.3) and views $\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta}$ as a set of deterministic restrictions to be imposed on δ_{gt}. Inference using the efficient minimum distance estimator uses only sampling variation in the $\hat{\delta}_{gt}$, which will be independent across all (g,t) if they are separately estimated, or which will be correlated if pooled methods are used.

Because we are ignoring the estimation error in $\hat{\delta}_{gt}$, we proceed simply by analyzing the panel data equation

$$\hat{\delta}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + v_{gt}, \quad t = 1,\ldots,T, g = 1,\ldots,G, \tag{3.4}$$

where we keep the error as v_{gt} because we are treating $\hat{\delta}_{gt}$ and δ_{gt} interchangeably. If we assume that We can apply the BDM findings and Hansen (2007a) results directly to this equation. Namely, if we estimate (3.4) by OLS – which means full year and group effects, along with x_{gt} – then the OLS estimator has satisfying properties as G and T both increase, provided $\{v_{gt} : t = 1, 2, \ldots, T\}$ is a weakly dependent (mixing) time series for all g. The simulations in BDM and Hansen (2007a) indicate that cluster-robust inference, where each cluster is a set of time periods, work reasonably well when $\{v_{gt}\}$ follows a stable AR(1) model and G is moderately large.

Hansen (2007b), noting that the OLS estimator (the fixed effects estimator) applied to (3.4) is inefficient when v_{gt} is serially uncorrelated (and possibly heteroskedastic), proposes feasible GLS. As is well known, if T is not large, estimating parameters for the variance matrix $\Omega_g = Var(\mathbf{v}_g)$, where $\mathbf{v}_g$ is the $T \times 1$ error vector for each g, is difficult when group effects have been removed. In other words, using the FE residuals, $\hat{v}_{gt}$, to estimate Ω_g can result in severe bias for small T. Solon (1984) highlighted this problem for the homoskedastic AR(1) model. Of course, the bias disappears as $T \to \infty$, and regression packages such as Stata, that have a built-in command to do fixed effects with AR(1) errors, use the usual AR(1) coefficient $\hat{\rho}$, obtained from

$$\hat{v}_{gt} \text{ on } \hat{v}_{g,t-1}, \; t = 2, \ldots, T, g = 1, \ldots, G. \tag{3.5}$$

As discussed in Wooldridge (2003) and Hansen (2007b), one way to account for the bias in $\hat{\rho}$ is to still use a fully robust variance matrix estimator. But Hansen's simulations show that this approach is quite inefficient relative to his suggestion, which is to bias-adjust the estimator $\hat{\rho}$ and then use the bias-adjusted estimator in feasible GLS. (In fact, Hansen covers the general $AR(p)$ model.) Hansen derives many attractive theoretical properties of his the estimator.An iterative bias-adjusted procedure has the same asymptotic distribution as $\hat{\rho}$ in the case $\hat{\rho}$ should work well: G and T both tending to infinity. Most importantly for the application to DD problems, the feasible GLS estimator based on the iterative procedure has the same asymptotic distribution as the GLS etsimator when $G \to \infty$ and T is fixed. When G and T are both large, there is no need to iterated to achieve efficiency.

Hansen further shows that, even when G and T are both large, so that the unadjusted AR coefficients also deliver asymptotic efficiency, the bias-adusted estimates deliver higher-order improvements in the asymptotic distribution. One limitation of Hansen's results is that they assume $\{\mathbf{x}_{gt} : t = 1, \ldots, T\}$ are strictly exogenous. We know that if we just use OLS – that is,

the usual fixed effects estimate – strict exogeneity is not required for consistency as $T \to \infty$. GLS, in exploiting correlations across different time periods, tends to exacerbate bias that results from a lack of strict exogeneity. In policy analysis cases, this is a concern if the policies can switch on and off over time, because one must decide whether the decision to implement or remove a program is related to past outcomes on the response.

With large G and small T, one can estimate an unstricted variance matrix Ω_g and proceed with GLS – this is the approach suggested by Kiefer (1980) and studied more recently by Hausman and Kuersteiner (2003). It is equivalent to dropping a time period in the time-demeaned equation and proceeding with full GLS (and this avoids the degeneracy in the variance matrix of the time-demeaned errors). Hausman and Kuersteiner show that the Kiefer approach works pretty well when $G = 50$ and $T = 10$, although substantial size distortions exist for $G = 50$ and $T = 20$.

Especially if the M_{gt} are not especially large, we might worry about ignoring the estimation error in the $\hat{\delta}_{gt}$. One simple way to avoid ignoring the estimation error in $\hat{\delta}_{gt}$ is to aggregate equation (3.1) over individuals, giving

$$\bar{y}_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \bar{\mathbf{z}}_{gt}\boldsymbol{\gamma} + v_{gt} + \bar{u}_{gt}, \; t = 1, .., T, g = 1, \ldots, G. \tag{3.6}$$

Of course, this equation can be estimated by fixed effects, too, and fully robust inference is available using Hansen (2007a) because the composite error, $\{r_{gt} \equiv v_{gt} + \bar{u}_{gt}\}$, is weakly dependent. Fixed Effects GLS using an unrestricted variance matrix can be used with large G and small T. The complication with using specific time series model for the error is the presence of $\bar{u}_{gt}$. With different M_{gt}, $Var(\bar{u}_{gt})$ is almost certainly heteroskedastic (and might be with the same M_{gt}, of course). So, even if we specify, say, an AR(1) model $v_{gt} = \rho v_{g,t-1} + e_{gt}$, the variance matrix of $\mathbf{r}_g$ is more complicated. One possibility is to just assume the composite error, r_{gt}, follows a simple model, implement Hansen's methods, but then use fully robust inference.

The Donald and Land (2007) approach applies in the current setting by using finite sample analysis applied to the pooled regression (3.4). However, DL assume that the errors $\{v_{gt}\}$ are uncorrelated across time, and so, even though for small G and T it uses small degrees-of-freedom in a t distribution, it does not account for uncertainty due to serial correlation in $\{v_{gt} : t = 1, \ldots, T\}$.

4. Individual-Level Panel Data

Individual-level panel data is a powerful tool for estimating policy effects. In the simplest

case we have two time periods and a binary program indicator, w_{it}, which is unity if unit i participates in the program at time t. A simple, effective model is

$$y_{it} = \alpha + \eta d2_t + \tau w_{it} + c_i + u_{it}, t = 1, 2, \tag{4.1}$$

where $d2_t = 1$ if $t = 2$ and zero otherwise, c_i is an observed effect, and u_{it} are the idiosyncratic errors. The coefficient τ is the treatment effect. A simple estimation procedure is to first difference to remove c_i :

$$(y_{i2} - y_{i1}) = \eta + \tau(w_{i2} - w_{i1}) + (u_{i2} - u_{i1}) \tag{4.2}$$

or

$$\Delta y_i = \eta + \tau \Delta w_i + \Delta u_i. \tag{4.3}$$

If $E(\Delta w_i \Delta u_i) = 0$, that is, the change in treatment status is uncorrelated with changes in the idiosyncratic errors, then OLS applied to (4.3) is consistent. The leading case is when $w_{i1} = 0$ for all i, so that no units we exposed to the program in the initial time period. Then the OLS estimator is

$$\hat{\tau} = \Delta\bar{y}_{treat} - \Delta\bar{y}_{control}, \tag{4.4}$$

which is a difference-in-differences estimate except that we different the means of the same units over time.This same estimate can be derived without introducing heterogeneity by simply writing the equation for y_{it} with a full set of group-time effects. Also, (4.4) is not the same estimate obtained from the regression y_{i2} on $1, y_{i1}, w_{i2}$ – that is, using y_{i1} as a control in a cross section regression. The estimates can be similar, but their consistency is based on different assumptions.

More generally, with many time periods and arbitrary treatment patterns, we can use

$$y_{it} = \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + c_i + u_{it}, \; t = 1, \ldots, T, \tag{4.5}$$

which accounts for aggregate time effects and allows for controls, $\mathbf{x}_{it}$. Estimation by FE or FD to remove c_i is standard, provided the policy indicator, w_{it}, is strictly exogenous: correlation beween w_{it} and u_{ir} for any t and r causes inconsistency in both estimators, although the FE estimator typically has smaller bias when we can assume conteporaneous exogeneity, $Cov(w_{it}, u_{it}) = 0$. Strict exogeneity can be violated if policy assignment changes in reaction to past outcomes on y_{it}. In cases where $w_{it} = 1$ whenever $w_{ir} = 1$ for $r < t$, strict exogeneity is usually a reasonable assumption.

Equation (4.5) allows policy designation to depend on a level effect, c_i, but w_{it} might be

correlated with unit-specific trends in the response, too. This suggests the "correlated random trend" model

$$y_{it} = c_i + g_i t + \lambda_t + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma} + u_{it}, \; t = 1, \ldots, T, \tag{4.6}$$

where g_i is the trend for unit i. A general analysis allows arbitrary corrrelation between (c_i, g_i) and w_{it}, which requires at least $T \geq 3$. If we first difference, we get

$$\Delta y_{it} = g_i + \eta_t + \tau \Delta w_{it} + \Delta \mathbf{x}_{it}\boldsymbol{\gamma} + \Delta u_{it}, \; t = 2, \ldots, T, \tag{4.7}$$

where $\eta_t = \lambda_t - \lambda_{t-1}$ is a new set of time effects. We can estimate (4.7) by differencing again, or by using FE. The choice depends on the serial correlation properties in $\{\Delta u_{it}\}$ (assume strict exogeneity of treatment and covariates). If Δu_{it} is roughly uncorrelated, FE is preferred. If the original errors $\{u_{it}\}$ are essentially uncorrelated, applying FE to (4.6), in the general sense of sweeping out the linear trends from the response, treatment, and covariates, is preferred. Fully robust inference using cluster-robust variance estimators is straightforward. Of course, one might want to allow the effect of the policy to change over time, which is easy by interacting time dummies with the policy indicator.

We can derive standard panel data approaches using the counterfactural framework from the treatment effects literature.For each (i, t), let $y_{it}(1)$ and $y_{it}(0)$ denote the counterfactual outcomes, and assume there are no covariates. One way to state the assumption of unconfoundedness of treatment is that, for time-constant heterogeneity c_{i0} and c_{i1},

$$E(y_{it0}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it0}|c_{i0}) \tag{4.8}$$

$$E(y_{it1}|\mathbf{w}_i, c_{i0}, c_{i1}) = E(y_{it1}|c_{i1}), \tag{4.9}$$

where $\mathbf{w}_i = (w_{i1}, \ldots, w_{iT})$ is the time sequence of all treatments. We saw this kind of strict exogeneity assumption conditional on latent variables several times before. It allows treatment to be correlated with time-constant heterogeneity, but does not allow treatment in any time period to be correlated with idiosyncratic changes in the counterfactuals. Next, assume that the expected gain from treatment depends at most on time:

$$E(y_{it1}|c_{i1}) = E(y_{it0}|c_{i0}) + \tau_t, \; t = 1, \ldots, T. \tag{4.10}$$

Writing $y_{it} = (1 - w_{it})y_{it0} + w_{it}y_{i1}$, and using (4.8), (4.9), and (4.10) gives

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) &= E(y_{it0}|c_{i0}) + w_{it}[E(y_{it1}|c_{i1}) - E(y_{it0}|c_{i0})] \\ &= E(y_{it0}|c_{i0}) + \tau_t w_{it}. \end{aligned} \tag{4.11}$$

If we now impose an additive structure on $E(y_{it0}|c_{i0})$, namely,

$$E(y_{it0}|c_{i0}) = \alpha_{t0} + c_{i0}, \tag{4.12}$$

then we arrive at

$$E(y_{it}|\mathbf{w}_i, c_{i0}, c_{i1}) = \alpha_{t0} + c_{i0} + \tau_t w_{it}, \tag{4.13}$$

an estimating equation that leads to well-known procedures. Because $\{w_{it} : t = 1, \ldots, T\}$ is strictly exogenous conditional on c_{i0}, we can use fixed effects or first differencing, with a full set of time period dummies. A standard analysis would use $\tau_t = \tau$, but, of course, we can easily allow the effects of the policy to change over time.

Of course, we can add covariates $\mathbf{x}_{it}$ to the conditioning sets and assume linearity, say $E(y_{it0}|\mathbf{x}_{it}, c_{i0}) = \alpha_{t0} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 + c_{i0}$. If (4.8) becomes

$$E(y_{it0}|\mathbf{w}_i, \mathbf{x}_i, c_{i0}, c_{i1}) = E(y_{it0}|\mathbf{x}_{it}, c_{i0}), \tag{4.14}$$

and similarly for (4.9), then the estimating equation simply adds $\mathbf{x}_{it}\boldsymbol{\gamma}_0$ to (4.13). More interesting models are obtained by allowing the gain from treatment to depend on heterogeneity. Suppose we assume, in addition to the ignorability assumption in (4.14) (and the equivalent condition for y_{it1})

$$E(y_{it1} - y_{it0}|\mathbf{x}_{it}, c_{i0}, c_{i1}) = \tau_t + a_i + (\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} \tag{4.15}$$

where a_i is a function of (c_{i0}, c_{i1}) normalized so that $E(a_i) = 0$ and $\boldsymbol{\xi}_t = E(\mathbf{x}_{it})$. Equation (4.15) allows the gain from treatment to depend on time, unobserved heterogeneity, and observed covariates. Then

$$\begin{aligned} E(y_{it}|\mathbf{w}_i, \mathbf{x}_i, c_{i0}, a_i) = \alpha_{t0} + \tau_t w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 \\ + w_{it}(\mathbf{x}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} + a_i w_{it}. \end{aligned} \tag{4.16}$$

This is a correlated random coefficient model because the coefficient on w_{it} is $(\tau_t + a_i)$, which has expected value τ_t. Generally, we want to allow w_{it} to be correlated with a_i and c_{i0}. With small T and large N, we do not try to estimate the a_i (nor the c_{i0}). But an extension of the within transformation effectively eliminates $a_i w_{it}$. Suppose we simplify a bit and assume $\tau_t = \tau$ and drop all other covariates. Then, a regression that appears to suffer from an incidental parameters problem turns out to consistently estimate τ: Regress y_{it} on year dummies, dummies for each cross-sectional observation, and latter dummies interacted with w_{it}. In other words, we estimate

$$\hat{y}_{it} = \hat{\alpha}_{t0} + \hat{c}_{i0} + \hat{\tau}_i w_{it}. \tag{4.17}$$

While $\hat{\tau}_i$ is usually a poor estimate of $\tau_i = \tau + a_i$, their average is a good estimator of τ :

$$\hat{\tau} = N^{-1} \sum_{i=1}^{N} \hat{\tau}_i. \tag{4.18}$$

A standard error can be calculated using Wooldridge (2002, Section 11.2) or bootstrapping.

We can apply the results from the linear panel data notes to determine when the usual FE estimator – that is, the one that ignores $a_i w_{it}$ – is consistent for τ. In addition to the unconfoundedness assumption, sufficient is

$$E(\tau_i | \ddot{w}_{it}) = E(\tau_i) = \tau, t = 1, \ldots, T, \tag{4.19}$$

where $\ddot{w}_{it} = w_{it} - \bar{w}_i$. Essentially, the individual-specific treatment effect can be correlated with the average propensity to recieve treatment, $\bar{w}_i$, but not the deviations for any particular time period.

Assumption (4.19) is not completely general, and we might want a simple way to tell whether the treatment effect is heterogeneous across individuals. Here, we an exploit correlation between the τ_i and treatment. Recalling that $\tau_i = \tau + a_i$, a useful assumption (that need not hold for obtaining a test) is

$$E(a_i | w_{i1}, \ldots w_{iT}) = E(a_i | \bar{w}_i) = \rho(\bar{w}_i - \mu_{\bar{w}_i}), \tag{4.20}$$

where other covariates have been suppressed. Then we can estimate the equation (with covariates)

$$\begin{aligned} y_{it} = \alpha_{t0} + \tau w_{it} + \mathbf{x}_{it}\boldsymbol{\gamma}_0 + w_{it}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)\boldsymbol{\delta} \\ + \rho w_{it}(\bar{w}_i - \bar{w}) + c_{i0} + e_{it} \end{aligned} \tag{4.21}$$

by standard fixed effects. Then, we use a simple t test on $\hat{\rho}$, robust to heteroskedasticity and serial correlation. If we reject, it does not mean the mean usual FE estimator is inconsistent, but it could be.

5. Semiparametric and Nonparametric Approaches

Return to the setting with two groups and two time periods. Athey and Imbens (2006) generalize the standard DD model in several ways. Let the two time periods be $t = 0$ and 1 and label the two groups $g = 0$ and 1. Let $Y_i(0)$ be the counterfactual outcome in the absense of intervention and $Y_i(1)$ the counterfactual outcome with intervention. AI assume that

$$Y_i(0) = h_0(U_i, T_i), \tag{5.1}$$

where T_i is the time period and

$$h_0(u, t) \text{ strictly increasing in } u \text{ for } t = 0, 1 \tag{5.2}$$

The random variable U_i represents all unobservable characteristics of individual i. Equation (5.1) incorporates the idea that the outcome of an individual with $U_i = u$ will be the same in a given time period, irrespective of group membership. The strict monotonicity assumption in (5.2) rules out discrete responses, but Athey and Imbens (2006) provide bounds under weak monotonicity, and show how, with additional assumptions, point identification be be recovered.

The distribution of U_i is allowed to vary across groups, but not over time within groups, so that

$$D(U_i|T_i, G_i) = D(U_i|G_i). \tag{5.3}$$

This assumption implies that, within group, the population distribution is stable over time.

The standard DD model can be expressed in this way, with

$$h_0(u,t) = u + \delta \cdot t \tag{5.4}$$

and

$$U_i = \alpha + \gamma G_i + V_i,\ V_i \perp (G_i, T_i) \tag{5.5}$$

although, because of the linearity, we can get by with the mean independence assumption $E(V_i|G_i, T_i) = 0$. If the treatment effect is constant across individuals, $\tau = Y_i(1) - Y_i(0)$, then we can write

$$Y_i = \alpha + \beta T_i + \gamma G_i + \tau G_i T_i + V_i, \tag{5.6}$$

where $Y_i = (1 - G_i T_i)Y_i(0) + G_i T_i Y_i(1)$ is the realized outcome. Because $E(V_i|G_i, T_i) = 0$, the parameters in (5.6) can be estimated by OLS.

Athey and Imbens call the extension of the usual DD model the *changes-in-changes* (CIC) model. They show not only how to recover the average treatment effect, but also that the distribution of the counterfactual outcome conditional on intervention, that is

$$D(Y_i(0)|G_i = 1, T_i = 1), \tag{5.7}$$

is identified. The distribution of $D(Y_i(1)|G_i = 1, T_i = 1)$ is identified by the data because $Y_i = Y_i(1)$ when $G_i = T_i = 1$. The extra condition AI use is that the support of the distribution of $D(U_i|G_i = 1)$ is contained in the support of $D(U_i|G_i = 0)$, written as

$$\mathbb{U}_1 \subseteq \mathbb{U}_0. \tag{5.8}$$

Let $F^0_{gt}(y)$ the be cumulative distribution function of $D(Y_i(0)|G_i = g, T_i = t)$ for $g = 1, 2$ and $t = 1, 2$, and let $F_{gt}(y)$ be the cdf for the observed outcome Y_i conditional on $G_i = g$ and

$T_i = t$. By definition, $F_{gt}(y)$ is generally identified from the data, assuming random sampling for each (g,t) pair. AI show that, under (5.1), (5.2), (5.3), and (5.8),

$$F_{11}^{(0)}(y) = F_{10}(F_{00}^{-1}(F_{01}(y))), \tag{5.9}$$

where $F_{00}^{-1}(\cdot)$ is the inverse function of F_{00}^{-1}, which exists under the strict monotonicity assumption. Notice that all of the cdfs appearing on the right hand size of (5.9) are estimable from the data; they are simply the cdfs for the observed outcomes conditional on different (g,t) pairs. Because $F_{11}^{(1)}(y) = F_{11}(y)$, we can estimate the entire distributions of both counterfactuals conditional on intervention, $G_i = T_i = 1$.

The average treatment effect in the CIC framework as

$$\begin{aligned} \tau_{CIC} &= E[Y(1)|G = 1, T = 1] - E[Y(0)|G = 1, T = 1]. \\ &= E(Y_{11}(1)) - E(Y_{11}(0)), \end{aligned} \tag{5.10}$$

where we drop the *i* subscript, $Y_{gt}(1)$ is a random variable having distribution $D(Y(1)|G = g,t)$, and $Y_{gt}(0)$ is a random variable having distribution $D(Y(0)|G = g,t)$. Under the same assumptions listed above,

$$\tau_{CIC} = E(Y_{11}) - E[F_{01}^{-1}(F_{00}(Y_{10}))] \tag{5.11}$$

where Y_{gt} is a random variable with distribution $D(Y|G = g,t)$. Given random samples from each subgroup, a generally consistent estimator of τ_{CIC} is

$$\hat{\tau}_{CIC} = N_{11}^{-1} \sum_{i=1}^{N_{11}} Y_{11,i} - N_{10}^{-1} \sum_{i=1}^{N_{10}} \hat{F}_{01}^{-1}(\hat{F}_{00}(Y_{10,i})), \tag{5.12}$$

for consistent estimators $\hat{F}_{00}$ and $\hat{F}_{01}$ of the cdfs for the control groups in the initial and later time periods, respectively. Now, $Y_{11,i}$ denotes a random draw on the observed outcome for the $g = 1, t = 1$ group and similarly for $Y_{10,i}$. Athey and Imbens establish weak conditions under which $\hat{\tau}_{CIC}$ is $\sqrt{N}$-asymptotically normal (where, naturally, observations must accumulate within each of the four groups). In the case where the distributions of Y_{10} and Y_{00} are the same, a simple difference in means for the treatment group over time.

The previous approach can be applied either with repeated cross sections or panel data. Athey and Imbens discuss how the assumptions can be relaxed with panel data, and how alternative estimation strategies are available. In particular, if U_{i0} and U_{i1} represent unobservables for unit *i* in the initial and later time periods, respectively, then (5.3) can be modified to

$$D(U_{i0}|G_i) = D(U_{i1}|G_i), \tag{5.13}$$

which allows for unobservd components structures $U_{it} = C_i + V_{it}$ where V_{it} has the same distribution in each time period.

As discussed by AI, with panel data there are other estimation approaches. As discussed earlier, Altonji and Matzkin (2005) use exchangeability assumptions to identify average partial effects. To illustrate how their approach might applie, suppose the counterfactuals satisfy the ignorability assumption

$$E(Y_{it}(g)|W_{i1}, \ldots, W_{iT}, U_i) = h_{tg}(U_i), t = 1, \ldots, T, g = 0, 1. \tag{5.14}$$

The treatment effect for unit i in period t is $h_{t1}(U_i) - h_{t0}(U_i)$, and the average treatment effect is

$$\tau_t = E[h_{t1}(U_i) - h_{t0}(U_i)], t = 1, \ldots, T. \tag{5.15}$$

Suppose we make the assumption

$$D(U_i|W_{i1}, \ldots, W_{iT}) = D(U_i|\bar{W}_i), \tag{5.16}$$

which means that only the intensity of treatment is correlated with heterogeneity. Under (5.14) and (5.16), it can be shown that

$$E(Y_{it}|W_i) = E[E(Y_{it}|W_i, U_i)|W_i] = E(Y_{it}|W_{it}, \bar{W}_i). \tag{5.17}$$

The key is that $E(Y_{it}|W_i)$ does not depend on $\{W_{i1}, \ldots, W_{iT}\}$ in an unrestricted fashion; it is a function only of $(W_{it}, \bar{W}_i)$. If W_{it} are continuous, or take on numerous values, we can use local smoothing methods to estimate $E(y_{it}|W_{it}, \bar{W}_i)$. In the treatment effect case, estimation is very simple because $(W_{it}, \bar{W}_i)$ can take on only $2T$. The average treatment effect can be estimated as

$$\hat{\tau}_t = N^{-1} \sum_{i=1}^{n} [\hat{\mu}_t^Y(1, \bar{W}_i) - \hat{\mu}_t^Y(0, \bar{W}_i)]. \tag{5.18}$$

If we pool across t (as well as i) and use a linear regression, Y_{it} on $1, d2_t, \ldots, dT_t, W_{it}, \bar{W}_i, t = 1, \ldots, T; i = 1, \ldots, N$, we obtain the usual fixed effects estimate $\hat{\tau}_{FE}$ as the coefficient on W_{it}. Wooldridge (2005) describes other scenarios and compares this strategy to other approaches. As we discussed earlier, a condtional MLE logit can estimate parameters by not generally ATEs, and require conditional independence. Chamberlain's correlated random effects probit models the heterogeneity as $U_i|W_i \sim \text{Normal}(\xi_0 + \xi_1 W_{i1} + \ldots + \xi_T W_{iT}, \eta^2)$, which identifies the ATEs without assuming exchangeability but maintaining a distributional assumption (and functional form for the

response probability).

For the leading case of two time periods, where treatment does not occur in the initial time period for any unit, but does for some units in the second time period, Abadie (2005) provides methods for both repeated cross sections and panel data that use unconfoundedness assumptions on changes over time. Here we describe the panel data approach. Omitting the *i* subscript, for any unit from the population there are counterfactual outcomes, which we write as $Y_t(w)$, where $t = 0, 1$ are the two time periods and $w = 0, 1$ represent control and treatment. In this setup, interest lies in two parameters, the average treatment effect in the second time period,

$$\tau_{ATE} = E[Y_1(1) - Y_1(0)], \tag{5.19}$$

or the average treatment effect on the treated,

$$\tau_{ATT} = E[Y_1(1) - Y_1(0)|W = 1]. \tag{5.20}$$

Remember, in the current setup, no units are treated in the initial time period, so $W = 1$ means treatment in the second time period.

As in Heckman, Ichimura, Smith, and Todd (1997), Abadie uses unconfoundedness assumptions on changes over time to identify τ_{ATT}, and straightforward extensions serve to identify τ_{ATE}. Given covariates X (that, if observed in the second time period, should not be influenced by the treatment), Abadie assumes

$$E[Y_1(0) - Y_0(0)|X, W] = E[Y_1(0) - Y_0(0)|X], \tag{5.21}$$

so that, conditional on X, treatment status is not related to the gain over time in the absense of treatment. In addition, the overlap assumption,

$$0 < P(W = 1|X) < 1 \tag{5.22}$$

is critical. (Actually, for estimating τ_{ATT}, we only need $P(W = 1|X) < 1$.) Under (5.21) and (5.22), it can be shown that

$$\tau_{ATT} = [P(W = 1)]^{-1} E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{[1 - p(X)]} \right\},$$

where Y_1 is the observed outcome in period 1, Y_0, is the outcome in period 0, and $p(X) = P(W = 1|X)$ is the propensity score. Dehejia and Wahba (1999) derived (5.23) for the cross-sectional case; see also Wooldridge (2002, Chapter 18). All quantities in (5.23) are observed or, in the case of the $p(X)$ and $\rho = P(W = 1)$, can be estimated. As in Hirano, Imbens, and Ridder (2003), a flexible logit model can be used for $p(X)$; the fraction of units

treated would be used for $\hat{\rho}$. Then

$$\hat{\tau}_{ATT} = \hat{\rho}^{-1} N^{-1} \sum_{i=1}^{N} \left\{ \frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{[1 - \hat{p}(X_i)]} \right\} \tag{5.23}$$

is consistent and $\sqrt{N}$-asymptotically normal. HIR discuss variance estimation. Imbens and Wooldridge (2007) provide a simple adjustment available in the case that $\hat{p}(\cdot)$ is treated as a parametric model.

If we also add

$$E[Y_1(1) - Y_0(1)|X, W] = E[Y_1(0) - Y_0(0)|X], \tag{5.24}$$

so that treatment is mean independent of the gain in the treated state, then

$$\tau_{ATE} = E\left\{ \frac{[W - p(X)](Y_1 - Y_0)}{p(X)[1 - p(X)]} \right\}, \tag{5.25}$$

which dates back to Horvitz and Thompson (1952); see HIR. Now, to estimate the ATE over the specified population, the full overlap assumption in (5.22) is needed, and

$$\hat{\tau}_{ATE} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{[W_i - \hat{p}(X_i)]\Delta Y_i}{\hat{p}(X_i)[1 - \hat{p}(X_i)]} \right\}. \tag{5.26}$$

Hirano, Imbens, and Ridder (2003) study this estimator in detail where $\hat{p}(x)$ is a series logit estimator. If we treat this estimator parametrically, a simple adjustment makes valid inference on $\hat{\tau}_{ATE}$ simple. Let $\hat{K}_i$ be the summand in (5.26) less $\hat{\tau}_{ATE}$, and let $\hat{D}_i = h(X_i)[W_i - \Lambda(h(X_i)\hat{\gamma})]$ be the gradient (a row vector) from the logit estimation. Compute the residuals, $\hat{R}_i$ from the OLS regression $\hat{K}_i$ on $\hat{D}_i$, $i = 1, \ldots, N$. Then, a consistent estimator of $Avar\sqrt{N}(\hat{\tau}_{ATE} - \tau_{ATE})$ is just the sample variance of the $\hat{R}_i$. This is never greater than if we ignore the estimation of $p(x)$ and just use the sample variance of the $\hat{K}_i$ themselves.

Under the unconfoundedness assumption, other strategies are available for estimating the ATE and ATT. One possibility is to run the regression

$$\Delta Y_i \text{ on } 1, W_i, \hat{p}(X_i), \; i = 1, \ldots, N,$$

which was studied by Rosenbaum and Rubin (1983) in the cross section case. The coefficient on W_i is the estimated ATE, although it requires some functional form restrictions for consistency. This is much preferred to pooling across t and running the regression Y_{it} on 1, $d1_t$, $d1_t \cdot W_i$, $\hat{p}(X_i)$. This latter regression requires unconfoundedness in the levels, and as dominated by the basic DD estimate from ΔY_i on $1, W_i$: putting in any time-constant function

as a control in a pooled regression is always less general than allowing an unobserved effect and differencing it away.

Regression adjustment is also possible under the previous assumptions. As derived by HIST,

$$E[Y_1(1) - Y_0(1)|X, W = 1] = \{[E(Y_1|X, W = 1) - E(Y_1|X, W = 0)] - [E(Y_0|X, W = 1) - E(Y_0|X, W = 0)]\} \tag{5.27}$$

where, remember, Y_t denotes the observed outcome for $t = 0, 1$. Each of the four conditional expectations on the right hand side is estimable using a random sample on the appropriate subgroup. Call each of these $\hat{\mu}_{wt}(x)$ for $w = 0, 1$ and $t = 0, 1$. Then a consistent estimator of τ_{ATT} is

$$N_1^{-1} \sum_{i=1}^{N} W_i \{[\hat{\mu}_{11}(X_i) - \hat{\mu}_{01}(X_i)] - [\hat{\mu}_{10}(X_i) - \hat{\mu}_{00}(X_i)]\}. \tag{5.28}$$

Computationally, this requires more effort than the weighted estimator proposed by Abdadie. Nevertheless, with flexible parametric functional forms that reflect that nature of *Y*, implementing (5.28) is not difficult. If *Y* is binary, then the $\hat{\mu}_{wt}$ should be obtained from binary response models; if *Y* is nonnegative, perhaps a count variable, then $\mu_{wt}(x) = \exp(x\beta_{wt})$ is attractive, with estimates obtained via Poisson regression (quasi-MLE).

6. Synthetic Control Methods for Comparative Case Studies

In Section 3 we discussed difference-in-differences methods that ignore sampling uncertain in the group/time period means (more generally, regression coefficients). Abadie, Diamond, and Haimmueller (2007), building on the work of Abadie and Gardeazabal (2003), argue that in policy analysis at the aggregate leve, there is no estimation uncertainty: the goal is to determine the effect of a policy on an entire population – say, a state – and the aggregate is measured without error (or very little error). The application in ADH is the effects of California's tobacco control program on state-wide smoking rates.

Of course, one source of uncertainty in any study using data with a time series dimension is the change in outcomes over time, even if the outcomes are aggregates measured without error. Event study methodology is one such example: often, time series regressions for a single entity, such as a state, are used to determine the effect of a policy (speed limit change, tobacco control program, and so on) on an aggregate outcome. But such event studies can suffer because they do not use a control group to account for aggregate effects that have nothing to

do with the specific state policy.

In the context of case control studies, where a time series is available for a particular unit – the treatment group – there are often many potential control groups. For example, in the tobacco control example, each state in the U.S. is a potential control for California (provided a state did not undergo a similar policy). ADH study this setup and emphasize the uncertainty associated with choosing suitable control groups. They point out that, even in the absense of sampling error, surely someone analyzing a state-level policy must nevertheless seal with uncertainty.

The approach of ADH is to allow one to select a synthetic control group out of a collection of possible controls. For example, in the California tobacco control case, ADH identify 38 states that did not implement such programs during the time period in question. Rather than just use a standard fixed effects analysis – which effectively treats each state as being of equal quality as a control group – ADH propose choosing a weighted average of the potential controls. Of course, choosing a suitable control group or groups is often done informally, including matching on pre-treatment predictors. ADH formalize the procedure by optimally choosing weights, and they propose methods of inference.

Consider a simple example, with only two time periods: one before the policy and one after. Let y_{it} be the outcome for unit i in time t, with $i = 1$ the (eventually) treated unit. Suppose there are J possible controls, and index these as $\{2, \ldots, J+1\}$. Let $\mathbf{x}_i$ be observed covariates for unit i that are not (or would not be) affected by the policy; $\mathbf{x}_i$ may contain period $t = 2$ covariates provided they are not affected by the policy. Generally, we can estimate the effect of the policy as

$$y_{12} - \sum_{j=2}^{J+1} w_j y_{j2},$$

where w_j are nonnegative weights that add up to one. The question is: how can we choose the weights – that is, the synthetic control – to obtain the best estimate of the intervention effect? ADH propose choosing the weights so as to minimize the distance between, in this simple case, $(y_{11}, \mathbf{x}_1)$ and $\sum_{j=2}^{J+1} w_j \cdot (y_{j1}, \mathbf{x}_j)$, or some linear combinations of elements of $(y_{11}, \mathbf{x}_1)$ and $(y_{j1}, \mathbf{x}_j)$. The optimal weights – which differ depending on how we define distance – produce the synthetic control whose pre-intervention outcome and predictors of post-intervention outcome are "closest." With more than two time periods, one can use averages of

pre-intervention outcomes, say, or weighted averages that give more weight to more recent pre-intervention outcomes.

ADH propose permutation methods for inference, which require estimating a placebo treatment effect for each region (potential control), using the same synthetic control method as for the region that underwent the intervention. In this way, one can compare the estimated intervention effect using the synthetic control method is substantially larger than the effect estimated from a region chosen at random. The inference is exact even in the case the aggregate outcomes are estimated with error using individual-level data.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 11, Wednesday, Aug 1st, 9.00-10.30am

Discrete Choice Models

1. Introduction

In this lecture we discuss multinomial discrete choice models. The modern literature on these models goes back to the work by Daniel McFadden in the seventies and eighties, (McFadden, 1973, 1981, 1982, 1984). In the nineties these models received much attention in the Industrial Organization literature, starting with Berry (1994), Berry, Levinsohn, Pakes (1995, BLP), and Goldberg (1995). In the IO literature the applications focused on demand for differentiated products, in settings with relatively large numbers of products, some of them close substitutes. In these settings a key feature of the conditional logit model, namely the Independence of Irrelevant Alternatives (IIA), was viewed as particularly unattractive. Three approaches have been used to deal with this. Goldberg (1995) used nested logit models to avoid the IIA property. McCulloch and Rossi (1994), and McCulloch, Polson and Rossi (2000) studied multinomial probit models with relatively unrestricted covariance matrices for the unobserved components. BLP, McFadden and Train (2000) and Berry, Levinsohn and Pakes (2004) uses random effects or mixed logit models, in BLP in combination with unobserved choice characteristics and using methods that allow for estimation using only aggregate choice data. The BLP approach has been very influential in the subsequent empirical IO literature.

Here we discuss these models. We argue that the random effects approach to avoid IIA is indeed very attractive, both substantively and computationally, compared to the nested logit or unrestricted multinomial probit models. In addition to the use of random effects to avoid the IIA property, the inclusion in the BLP methodology of unobserved choice characteristics, and the ability to estimate the models with market share rather than individual level data makes their methods very flexible and widely applicable. We discuss extensions to the BLP set up allowing multiple unobserved choice characteristics, and the richness required for these

models to rationalize general choice data based on utility maximization. We also discuss the potential benefits of using Bayesian methods.

2. Multinomial and Conditional Logit Models

First we briefly review the multinomial and conditional logit models.

2.1 Multinomial Logit Models

We focus on models for discrete choice with more than two choices. We assume that the outcome of interest, the choice Y_i takes on non-negative, un-ordered integer values between zero and J; $Y_i \in \{0, 1, \ldots, J\}$. Unlike the ordered case there is no particular meaning to the ordering. Examples are travel modes (bus/train/car), employment status (employed/unemployed/out-of-the-laborforce), car choices (suv, sedan, pickup truck, convertible, minivan), and many others.

We wish to model the distribution of Y in terms of covariates. In some cases we will distinguish between covariates Z_i that vary by units (individuals or firms), and covariates that vary by choice (and possibly by individual), X_{ij}. Examples of the first type include individual characteristics such as age or education. An example of the second type is the cost associated with the choice, for example the cost of commuting by bus/train/car, or the price of a product, or the speed of a computer chip. This distinction is important from the substantive side of the problem. McFadden developed the interpretation of these models through utility maximizing choice behavior. In that case we may be willing to put restrictions on the way covariates affect utilities: characteristics of a particular choice should affect the utility of that choice, but not the utilities of other choices.

The strategy is to develop a model for the conditional probability of choice j given the covariates. Suppose we only have individual-specific covariates, and the model is $\Pr(Y_i = j|Z_i = z) = P_j(z;\theta)$. Then the log likelihood function is

$$L(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{J} 1\{Y_i = j\} \cdot \ln P_j(Z_i;\theta).$$

A natural extension of the binary logit model is to model the response probability as

$$\Pr(Y_i = j|Z_i = z) = \frac{\exp(z'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for choices $j = 1, \ldots, J$ and

$$\Pr(Y_i = 0|Z_i = z) = \frac{1}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for the first choice. The γ_l here are choice-specific parameters. This multinomial logit model leads to a very well-behaved likelihood function, and it is easy to estimate using standard optimization techniques. Interestingly, it can be viewed as a special case of the following conditional logit.

2.2 Conditional Logit Models

Suppose all covariates vary by choice (and possibly also by individual, but that is not essential here). Then McFadden proposed the conditional logit model:

$$\Pr(Y_i = j|X_{i0}, \ldots, X_{iJ}) = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)},$$

for $j = 0, \ldots, J$. Now the parameter vector β is common to all choices, and the covariates are choice-specific.

The multinomial logit model can be viewed as a special case of the conditional logit model. Suppose we have a vector of individual characteristics Z_i of dimension K, and J vectors of coefficients γ_j, each of dimension K. Then define for choice j, $j = 1, \ldots, J$, the vector of covariates X_{ij} as the vector of dimension $K \times J$, with all elements equal to zero other than the elements $K \times (j-1) + 1$ to $K \times j$ which are equal to Z_i:

$$X_{i1} = \begin{pmatrix} Z_i \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad X_{ij} = \begin{pmatrix} 0 \\ \vdots \\ Z_i \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad X_{iJ} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ Z_i \end{pmatrix}, \text{ and } X_{i0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and define the common parameter vector β, of dimension $K \cdot J$, as

$$\beta = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_J \end{pmatrix}.$$

Then

$$\Pr(Y_i = j|Z_i) = \frac{\exp(Z_i'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)} = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = j|X_{i0}, \ldots, X_{iJ}),$$

for $j = 1, \ldots, J$, and

$$\Pr(Y_i = 0|Z_i) = \frac{1}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)} = \frac{\exp(X_{i0}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = 0|X_{i0}, \ldots, X_{iJ}).$$

2.3 Link with Utility Maximization

McFadden motivates the conditional logit model by extending the single latent index model to multiple choices. Suppose that the utility, for individual i, associated with choice j, is

$$U_{ij} = X_{ij}'\beta + \varepsilon_{ij}. \tag{1}$$

Furthermore, let individual i choose option j (so that $Y_i = j$) if choice j provides the highest level of utility, or

$$Y_i = j \quad \text{if } U_{ij} \geq U_{il} \text{ for all } l = 0, \ldots, J,$$

(ties have probability zero because of the continuity of the distribution for ε).

Now suppose that the ε_{ij} are independent accross choices and individuals and have type I extreme value distributions. Then the choice Y_i follows the conditional logit model. The type I extreme value distribution has cumulative distribution function

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \qquad \text{and pdf} \quad f(\epsilon) = \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)).$$

This distribution has a unique mode at zero, a mean equal to 0.58, and a a second moment of 1.99 and a variance of 1.65. See Figure 1 for the probability density function and the comparison with the normal density. Note the assymmetry of the distribution.

Given the extreme value distribution the probability of choice 0 is

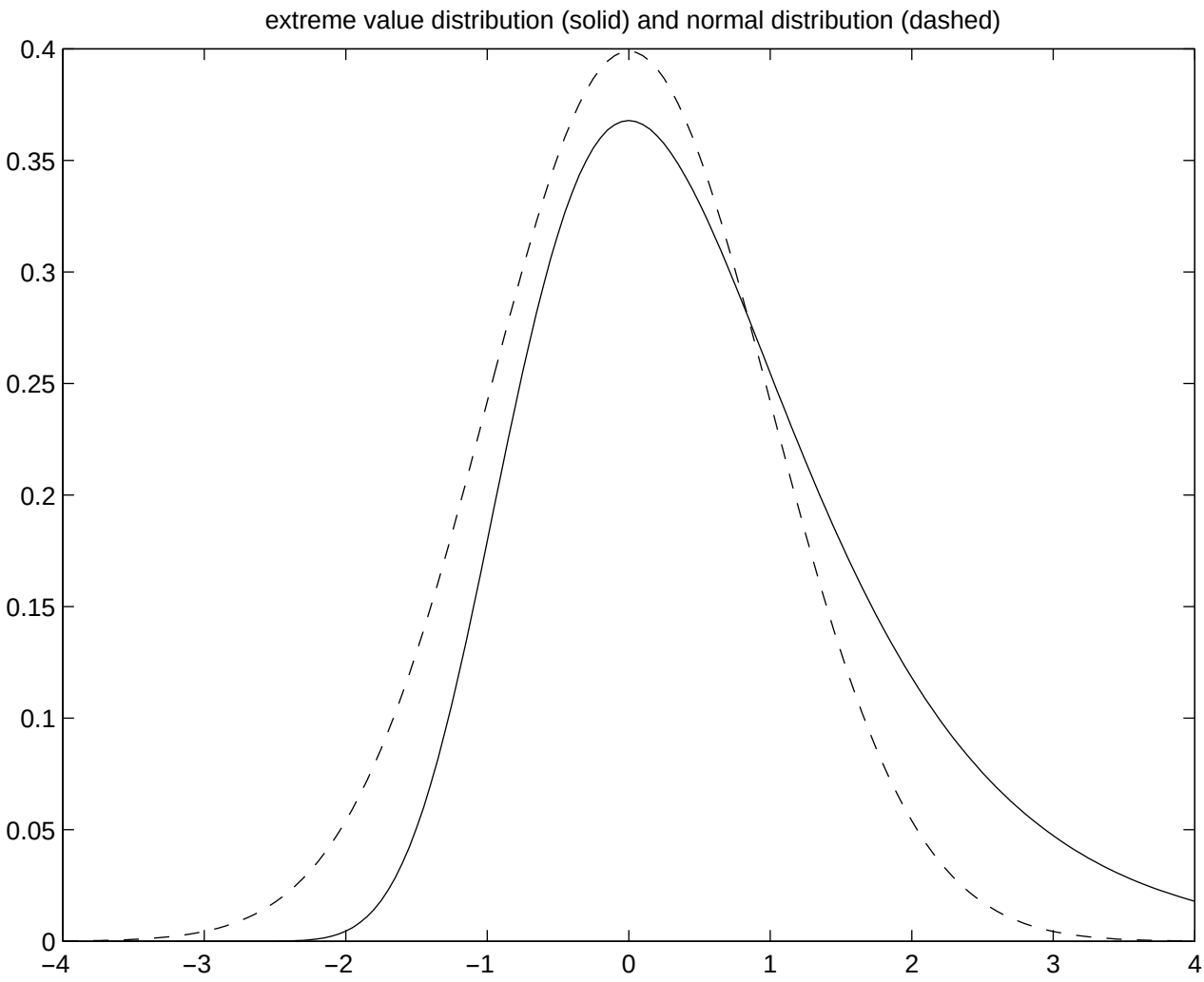
$$\begin{aligned}
\Pr(Y_i = 0|X_{i0},\ldots,X_{iJ}) &= \Pr(U_{i0} > U_{i1},\ldots,U_{i0} > U_{iJ}) \\
&= \Pr(\varepsilon_{i0} + X_{i0}'\beta - X_{i1}'\beta > \varepsilon_{i1},\ldots,\varepsilon_{i0} + X_{i0}'\beta - X_{iJ}'\beta > \varepsilon_{iJ}) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_{i0}+X_{i0}'\beta-X_{i1}'\beta}\cdots\int_{-\infty}^{\varepsilon_{i0}+X_{iJ}'\beta-X_{iJ}'\beta} f(\varepsilon_{i0})\ldots f(\varepsilon_{iJ})d\varepsilon_{iJ}\ldots,d\varepsilon_{i0} \\
&= \int_{-\infty}^{\infty}\exp(-\varepsilon_{0i})\exp(-\exp(-\varepsilon_{0i})\cdot\exp(-\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{i1}'\beta))\ldots \\
&\qquad\qquad \times\exp(-\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{iJ}'\beta))d\varepsilon_{i0} \\
&= \int_{-\infty}^{\infty}\exp(-\varepsilon_{0i})\exp\Big[-\exp(-\varepsilon_{0i})-\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{i1}'\beta))\ldots \\
&\qquad\qquad -\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{iJ}'\beta)\Big]d\varepsilon_{i0} \\
&= \frac{\exp(X_{i0}'\beta)}{\sum_{j=0}^{J}\exp(X_{j0}'\beta)}.
\end{aligned}$$

To see the different steps in this derivation note that

$$\int_{-\infty}^{c}\exp(-\epsilon)\cdot\exp(-\exp(-\epsilon))d\epsilon = F(c) = \exp(-\exp(-c)),$$

for the extreme value distribution. Also,

$$\int_{-\infty}^{\infty}\exp(-\epsilon)\cdot\exp(-\exp(-\epsilon-c))d\epsilon$$


extreme value distribution (solid) and normal distribution (dashed)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4
−3
−2
−1
0
1
2
3
4

$$= \int_{-\infty}^{\infty} \exp(-\eta + c) \cdot \exp(-\exp(-\eta))d\eta$$

$$= \exp(c) \cdot \int_{-\infty}^{\infty} \exp(-\eta) \cdot \exp(-\exp(-\eta))d\eta = \exp(c),$$

by change of variables, which we apply with

$$c = -\ln\left(1 + \exp(X_{i1}'\beta - X_{i0}'\beta) + \ldots + \exp(X_{iJ}'\beta - X_{i0}'\beta)\right).$$

3. Independence of Irrelevant Alternatives

The main problem with the conditional logit is the property of Independence of Irrelevant Alternative (IIA). Consider the conditional probability of choosing j given that you choose either j or l:

$$\Pr(Y_i = j | Y_i \in \{j, l\}) = \frac{\Pr(Y_i = j)}{\Pr(Y_i = j) + \Pr(Y_i = l)} = \frac{\exp(X_{ij}'\beta)}{\exp(X_{ij}'\beta) + \exp(X_{il}'\beta)}.$$

This probability does not depend on the characteristics X_{im} of alternatives m other than j and l. This is sometimes unattractive. The traditional example is McFadden's famous blue bus/red bus example. Suppose there are initially three choices: commuting by car, by red bus or by blue bus. It would seem reasonable be to assume that people have a preference over cars versus buses, but are indifferent between red versus blue buses. One could capture this by assuming that

$$U_{i,\text{redbus}} = U_{i,\text{bluebus}},$$

with the choice between the blue and red bus being random. So, to be explicit, suppose that $X_{i,\text{bluebus}} = X_{i,\text{redbus}} = X_{i,\text{bus}}$. Then suppose that the probability of commuting by bus is

$$\Pr(Y_i = \text{bus}) = \Pr(Y_i = \text{redbus or bluebus}) = \frac{\exp(X_{i,\text{bus}}'\beta)}{\exp(X_{i,\text{bus}}'\beta) + \exp(X_{i,\text{car}}'\beta)},$$

and the probability of choosing a red bus or blue bus, conditional on choosing a bus, is

$$\Pr(Y_i = \text{redbus} | Y_i = \text{bus}) = \frac{1}{2}.$$

That would imply that the conditional probability of commuting by car, given that one commutes by car or red bus, would differ from the same conditional probability if there is no blue bus. Presumably taking away the blue bus choice would lead all the current blue bus users to shift to the red bus, and not to cars.

The conditional logit model does not allow for this type of substitution pattern. Another way of stating the problems with the conditional logit model is to say that it generates unrealistic substitution patterns. Let us make that argument more specific. Suppose that individuals have the choice out of three Berkeley restaurants, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B). Suppose the two characteristics of the restaurants are price with $P_C = 95$, $P_L = 80$, and $P_B = 5$, and quality, with $Q_C = 10$, $Q_L = 9$, and $Q_B = 2$. Suppose that market shares for the three restaurants are $S_C = 0.10$, $S_L = 0.25$, and $S_B = 0.65$. These numbers are roughly consistent with a conditional logit model where the utility associated with individual i and restaurant j is

$$U_{ij} = -0.2 \cdot P_j + 2 \cdot Q_j + \epsilon_{ij},$$

with independent extreme value ϵ_{ij}, and individuals go to the restaurant with the highest utility. Now suppose that we raise the price at Lalime's to 1000 (or raise it to infinity, corresponding to taking it out of business). In that case the prediction of the conditional logit model is that the market shares for Chez Panisse and the Bongo Burger go to $\tilde{S}_C = 0.13$ and $\tilde{S}_B = 0.87$. That seems implausible. The people who were planning to go to Lalime's would appear to be more likely to go to Chez Panisse if Lalime's is closed than to go to the Bongo Burger, and so one would expect $\tilde{S}_C \approx 0.35$ and $\tilde{S}_B \approx 0.65$. The model on the other hand predicts that most of the individuals who would have gone to Lalime's will now dine (if that is the right term) at the Bongo Burger.

Recall the latent utility set up with the utility for individual i and choice j equal to

$$U_{ij} = X'_{ij}\beta + \epsilon_{ij}. \tag{2}$$

In the conditional logit model we assume independent ϵ_{ij} with extreme value distributions. This is essentially what creates the IIA property. (This is not completely correct, because other distributions for the unobserved, say with normal errors, we would not get IIA exactly, but something pretty close to it.) The solution is to allow in some fashion for correlation between the unobserved components in the latent utility representation. In particular, with a choice set that contains multiple versions of essentially the same choice (like the red bus or the blue bus), we should allow the latent utilities for these choices to be identical, or at least very close. In order to achieve this the unobserved components of the latent utilities would have to be highly correlated for those choices. This can be done in a number of ways.

4. Models without Independence of Irrelevant Alternatives

Here we discuss three ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the unobserved components of the latent utility. All of these originate in some form or another in McFadden's work (e.g., McFadden, 1981, 1982, 1984). The first is the nested logit model where the researcher groups together sets of choices. In the simple version with a single layer of nests this allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation between the unobserved components of choices in different nests. Second, the unrestricted multinomial probit model with no restrictions on the covariance between unobserved components, beyond normalizations. Third, the mixed or random coefficients logit where the marginal utilities associated with choice characteristics are allowed to vary between individuals. This generates positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

4.1 Nested Logit

One way to induce correlation between the choices is through nesting them. Suppose the

set of choices $\{0, 1, \ldots, J\}$ can be partitioned into S sets $B_1, \ldots, B_S$, so that the full set of choices can be written as

$$\{0, 1, \ldots, J\} = \cup_{s=1}^{S} B_s.$$

Let Z_s be set-specific characteristics. (It may be that the set of set specific variables is empty, or just a vector of indicators, with Z_s an S-vector of zeros with a one for the sth element.) Now let the conditional probability of choice j given that your choice is in the set B_s, or $Y_i \in B_s$ be equal to

$$\Pr(Y_i = j | X_i, Y_i \in B_s) = \frac{\exp(\rho_s^{-1} X_{ij}'\beta)}{\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}'\beta)},$$

for $j \in B_s$, and zero otherwise. In addition suppose the marginal probability of a chocie in the set B_s is

$$\Pr(Y_i \in B_s | X_i) = \frac{\exp(Z_s'\alpha) \left(\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}'\beta)\right)^{\rho_s}}{\sum_{t=1}^{S} \exp(Z_t'\alpha) \left(\sum_{l \in B_t} \exp(\rho_t^{-1} X_{il}'\beta)\right)^{\rho_s}}.$$

If we fix $\rho_s = 1$ for all s, then

$$\Pr(Y_i = j | X_i) = \frac{\exp(X_{ij}'\beta + Z_s'\alpha)}{\sum_{t=1}^{S} \sum_{l \in B_t} \exp(X_{il}'\beta + Z_t\alpha)},$$

and we are back in the conditional logit model.

In general this model corresponds to individuals choosing the option with the highest utility, where the utility of choice j in set B_s for individual i is

$$U_{ij} = X_{ij}'\beta + Z_s'\alpha + \epsilon_{ij},$$

where the joint distribution function of the ϵ_{ij} is

$$F(\epsilon_{i0}, \ldots, \epsilon_{iJ}) = \exp\left(-\sum_{s=1}^{S} \left(\sum_{j \in B_s} \exp\left(-\rho_s^{-1}\epsilon_{ij}\right)\right)^{\rho_s}\right).$$

Within the sets the correlation coefficient for the ϵ_{ij} is approximately equal to $1-\rho$. Between the sets the ϵ_{ij} are independent.

The nested logit model could capture the blue bus/red bus example by having two nests, the first $B_1 = \{\text{redbus}, \text{bluebus}\}$, and the second one $B_2 = \{\text{car}\}$.

How do you estimate these models? One approach is to construct the log likelihood and directly maximize it. That is complicated, especially since the log likelihood function is not concave, but it is not impossible. An easier alternative is to directly use the nesting structure. Within a nest we have a conditional logit model with coefficients β/ρ_s. Hence we can directly estimate β/ρ_s using the concavity of the conditional logit model. Denote these estimates of β/ρ_s by $\widehat{\beta/\rho_s}$. Then the probability of a particular set B_s can be used to estimate ρ_s and α through

$$\Pr(Y_i \in B_s|X_i) = \frac{\exp(Z_s'\alpha)\left(\sum_{l\in B_s}\exp(X_{il}'\widehat{\beta/\rho_s})\right)^{\rho_s}}{\sum_{t=1}^S \exp(Z_t'\alpha)\left(\sum_{l\in B_t}\exp(X_{il}'\widehat{\beta/\rho_t})\right)^{\rho_s}} = \frac{\exp(Z_s'\alpha+\rho_s\hat{W}_s)}{\sum_{t=1}^S\exp(Z_t'\alpha+\rho_t\hat{W}_t)},$$

where

$$\hat{W}_s = \ln\left(\sum_{l\in B_s}\exp(X_{il}'\widehat{\beta/\rho_s})\right),$$

known as the "inclusive values". Hence we have another conditional logit model back that is easily estimable. These two-step estimators are not efficient. The variance/covariance matrix is provided in McFadden (1981).

These models can be extended to many layers of nests. See for an impressive example of a complex model with four layers of multiple nests Goldberg (1995). Figure 2 shows the nests in the Goldberg application. The key concern with the nested logit models is that results may be sensitive to the specification of the nest structure. The researcher chooses the choices that are potentially close, with the data being used to estimate the amount of correlation. In contrast, in the random effects models, choices can only be close if they are close in terms of observed choice characteristics, with the data being used to estimate the

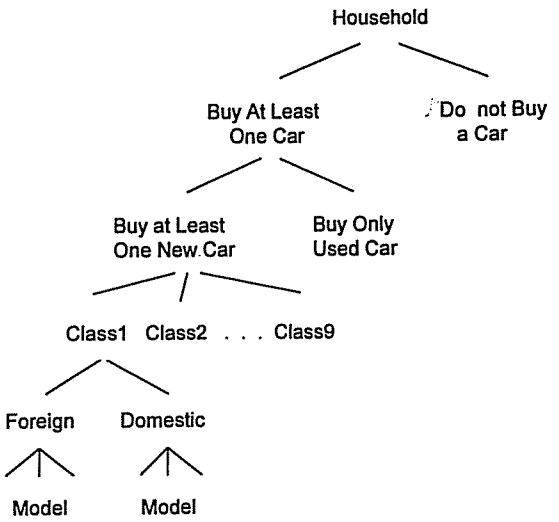


FIGURE 1.—Automobile choice model.

relative importance of the various choice characteristics. In that sense the nested logit model can be more flexible, allowing the researcher to group together choices that are far apart in terms of observed choice characteristics, but it is more demanding in requiring the researcher to make these decisions *a priori*.

4.2 Multinomial Probit

A second possibility is to directly free up the covariance matrix of the error terms. This is more natural to do in the multinomial probit case. See McCulloch and Rossi (1994) McCulloch, Polson, and Rossi (2000) for general discussion.

We specify:

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X_{i0}'\beta + \epsilon_{i0} \\ X_{i1}'\beta + \epsilon_{i1} \\ \vdots \\ X_{iJ}'\beta + \epsilon_{iJ} \end{pmatrix},$$

with

$$\epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} \Big| X_i \sim \mathcal{N}(0, \Omega),$$

for some relatively unrestricted $(J+1) \times (J+1)$ covariance matrix Ω. We do need some normalizations on Ω beyond symmetry. Recall that in the binary choice case (which corresponds to $J = 1$) there were no free parameters in the distribution of ϵ, which implies three restrictions on the symmetric matrix Ω.

In principle we can derive the probability for each choice given the covariates, construct the likelihood function based on that, and maximize it using an optimization algorithm like Davidon-Fletcher-Powell (Gill, Murray, and Wright, 1981) or something similar. In practice this is very difficult with $J \geq 3$. Evaluating the probabilities involves calculating a third order integral involving normal densities. This is difficult to to using standard integration methods. There are two alternatives.

There is a substantial literature on simulation methods for computing estimates in these models. See for an early example Manski and Lerman (1981), general studies McFadden (1989), and Pakes and Pollard (1989), and Hajivassiliou and Ruud (1994) for a review. Geweke, Keane, and Runkle (1994) and Hajivasilliou and McFadden (1990) proposed a way of calculating the probabilities in the multinomial probit models that allowed researchers to deal with substantially larger choice sets. A simple attempt to estimate the probabilities would be to draw the ϵ_i from a multivariate normal distribution and calculate the probability of choice j as the number of times choice j corresponded to the highest utility. This does not work well in cases with many (more than four) choices. The Geweke-Hajivasilliou-Keane (GHK) simulator uses a more complicated procedure that draws sequentially and combines the draws with the calculation of univariate normal integrals so that the resulting probabilities are smooth in the parameters.

From a Bayesian perspective drawing from the posterior distribution of β and Ω is straightforward. The key is setting up the vector of unobserved random variables as

$$\theta = (\beta, \Omega, U_{i0}, \ldots, U_{iJ}) ,$$

and defining the most convenient partition of this vector. Suppose we know the latent utilities U_i for all individuals. Then the normality makes this a standard linear model problem, and we can sample sequentially from $\beta|\Omega$ and $\Omega|\beta$ given the appropriate conjugate prior distributions (normal for β and inverse Wishart for Ω). Given the parameters drawing from the unobserved utilities can be done sequentially: for each unobserved utility given the others we would have to draw from a truncated normal distribution, which is straightforward. See McCulloch, Polson, and Rossi (2000) for details.

The attraction of this approach is that there are no restrictions on which choices are close. In contrast, in the nested logit approach the researcher specifies which choices are potentially close, and in the random effects approach only choices that are close in terms of observed choice characteristics can be close. The difficulty, however, with the unrestricted multinomial probit approach is that with a reasonable number of chocies this frees up a

large number of parameters (all elements in the $(J+1) \times (J+1)$ dimensional covariance matrix of latent utilities, minus some that are fixed by normalizations.) Estimating all these covariance parameters precisely, based on only first choice data (as opposed to data where we know for each individual additional orderings, e.g., first and second choices), is difficult with the sample sizes typically available.

4.3 Random Coefficient (Mixed) Logit (or Probit)

A third possibility to get around the IIA property is to allow for unobserved heterogeneity in the slope coefficients. This is a very natural idea. Why do we fundamentally think that if Lalime's price goes up, the individuals who were planning to go Lalime's go to Chez Panisse instead, rather than to the Bongo Burger? The reason is that we think individuals who have a taste for Lalime's are likely to have a taste for close substitute in terms of observable characteristics, Chez Panisse as well, rather than for the Bongo Burger.

We can model this by allowing the marginal utilities to vary at the individual level:

$$U_{ij} = X_{ij}'\beta_i + \epsilon_{ij},$$

where the ϵ_{ij} are again independent of everything else, and of each other, either extreme value, or normal. We can also write this as

$$U_{ij} = X_{ij}'\overline{\beta} + \nu_{ij},$$

where

$$\nu_{ij} = \epsilon + _{ij} + X_{ij} \cdot (\beta_i - \overline{\beta}),$$

which is no longer independent across choices. The key ingredient is the vector of individual specific taste parameters β_i. See for a general discussion of such models and their properties in approximating general choice patterns McFadden and Train (2000). One possibility is to

assume the existence of a finite number of types of individuals, similar to the mixture models used by Heckman and Singer (1984) in duration settings:

$$\beta_i \in \{b_0, b_1, \ldots, b_K\},$$

with

$$\Pr(\beta_i = b_k|Z_i) = p_k, \qquad \text{or} \quad \Pr(\beta_i = b_k|Z_i) = \frac{\exp(Z_i'\gamma_k)}{1 + \sum_{l=1}^{K} \exp(Z_i'\gamma_l)}.$$

Here the taste parameters take on a finite number of values, and we have a finite mixture. We can use either Gibbs sampling with the indicator of which mixture an observations belongs to as an unobserved random variable, or use the EM algorithm (Dempster, Laird, and Rubin, 1977).

Alternatively we could specify

$$\beta_i|Z_i \sim \mathcal{N}(Z_i'\gamma, \Sigma),$$

where we use a normal (continuous) mixture of taste parameters. Just evaluating the likelihood function would be very difficult in this setting if there is a large number of choices. This would involve integrating out the random coefficients which could be very computationally intensive. See McFadden and Train (2000). Using Gibbs sampling with the unobserved β_i as additional unobserved random variables may be an effective way of doing inference.

5. Berry-Levinsohn-Pakes

Here we consider again random effects logit models. BLP extended these models to allow for unobserved product characteristics, endogeneity of choice characteristics, and developed methods that allowed for consistent estimation without individual level choice data. Their approach has been widely used in Industrial Organization, where it is used to model demand for differentiated products, often in settings with a large number of products. See Nevo (2000) and Ackerberg, Benkard, Berry, and Pakes (2005) for reviews and references.

Compared to the earlier examples we have looked at there is an emphasis in this study, and those that followed it, on the large number of goods and the potential endogeneity of some of the product characteristics. (Typically one of the regressors is the price of the good.) In addition the procedure only requires market level data. We do not need individual level purchase data, just market shares and estimates of the distribution of individual characteristics by market. In practice we need a fair amount of variation in these things to estimate the parameters well, but in principle this is less demanding in terms of data required. On the other hand, we do need data by market, where before we just needed individual purchases in a single market (although to identify price effects we would need variation in prices by individuals in that case).

The data have three dimensions: products, indexed by $j = 0, \ldots, J$, markets, $t = 1, \ldots, T$, and individuals, $i = 1, \ldots, N_t$. We only observe one purchase per individual. The large sample approximations are based on large N and T, and fixed J.

Let us go back to the random coefficients model, now with each utility indexed by individual, product and market:

$$U_{ijt} = \beta_i' X_{jt} + \zeta_{jt} + \epsilon_{ijt}.$$

The ζ_{jt} is a unobserved product characteristic. This component is allowed to vary by market and product. It can include product and market dummies (for example, we can have $\zeta_{jt} = \zeta_j + \zeta_t$). Unlike the observed product characteristics this unobserved characteristic does not have a individual-specific coefficient. The inclusion of this component allows the model to rationalize any pattern of market shares. The observed product characteristics may include endogenous characteristics like the price.

The ϵ_{ijt} unobserved components have extreme value distributions, independent across all individuals i, products j, and markets t.

The random coefficients β_i, with dimension equal to that of the observable characteristics X_{jt}, say K, are assumed to be related to individual observable characteristics. We postulate

the following linear form:

$$\beta_i = \beta + Z_i'\Gamma + \eta_i,$$

with

$$\eta_i | Z_i \sim \mathcal{N}(0, \Sigma).$$

So if the dimension of Z_i is $L \times 1$, then Γ is a $L \times K$ matrix. The Z_i are normalized to have mean zero, so that the β's are the average marginal utilities. The normality assumption is not necessary, and unlikely to be important. Other distributional assumptions can be substituted.

BLP developed an approach to estimate models of this type that does not require individual level data. Instead it exploits aggregate (market level) data in combination with estimates of the distribution of Z_i. Specifically the data consist of estimated shares $\hat{s}_{tj}$ for each choice j in each market t, combined with observations from the marginal distribution of individual characteristics (the Z_i's) for each market, often from representative data sets such as the CPS.

First write the latent utilities as

$$U_{ijt} = \delta_{jt} + \nu_{ijt} + \epsilon_{ijt},$$

where

$$\delta_{jt} = \beta' X_{jt} + \zeta_{jt}, \qquad \text{and} \quad \nu_{ijt} = (Z_i'\Gamma + \eta_i)' X_{jt}.$$

Now consider for fixed Γ and Σ and δ_{jt} the implied market share for product j in market t, s_{jt}. This can be calculated analytically in simple cases. For example with $\Gamma_{jt} = 0$ and $\Sigma = 0$, the market share is a very simple function of the δ_{jt}:

$$s_{jt}(\delta_{jt}, \Gamma = 0, \Sigma = 0) = \frac{\exp(\delta_{jt})}{\sum_{l=0}^{J} \exp(\delta_{lt})}.$$

More generally, this is a more complex relationship. We can always calculate the implied market share by simulation: draw from the distribution of Z_i in market t, draw from the distribution of η_i, and calculate the implied purchase probability (or even simulate the implied purchase by also drawing from the distribution of ϵ_{ijt}). Do that repeatedly and you will be able to calculate the market share for this product/market. Call the vector function obtained by stacking these functions for all products and markets $s(\delta, \Gamma, \Sigma)$.

Next, fix only Γ and Σ. For each value of δ_{jt} we can find the implied market share. Now find the vector of δ_{jt} such that the implied market shares are equal to the observed market shares $\hat{s}_{jt}$ for all j, t. BLP suggest using the following algorithm. Given a starting value for δ^0_{jt}, use the updating formula:

$$\delta^{k+1}_{jt} = \delta^k_{jt} + \ln s_{jt} - \ln s_{jt}(\delta^k, \Gamma, \Sigma).$$

BLP show this is a contraction mapping, and so it defines a function $\delta(s, \Gamma, \Sigma)$ expressing the δ as a function of observed market shares, and parameters Γ and Σ. In order to implement this, one needs to approximate the implied market shares accurately for each iteration in the contraction mapping, and then you will need to do this repeatedly to get the contraction mapping to converge.

Note that does require that each market share is accurately estimated. If all we have is an estimated market share, then even if this is unbiased, the procedures will not necessarily work. In that case the log of the estimated share is not unbiased for the log of the true share. In practice the precision of the estimated market share is so much higher than that of the other parameters that this is unlikely to matter.

Given this function $\delta(s, \Gamma, \Sigma)$ define the residuals

$$\omega_{jt} = \delta_{jt}(s, \Gamma, \Sigma) - \beta' X_{jt}.$$

At the true values of the parameters and the true market shares this is equal to the unobserved product characteristic ζ_{jt}.

Now we can use GMM or instrumental variable methods. We assume that the unobserved product characteristics are uncorrelated with observed product characteristics (other than typically price). This is not sufficient since the observed product characteristics enter directly into the model. We need more instruments, and typically use things like characteristics of other products by the same firm, or average characteristics by competing products. The general GMM machinery will also give us the standard errors for this procedure. This is where the method is most challenging. Finding values of the parameters that set the average moments closest to zero can be difficult.

It is instructive to see what this approach does if we in fact have, and know we have, a conditional logit model with fixed coefficients. In that case $\Gamma = 0$, and $\Sigma = 0$. Then we can invert the market share equation to get the market specific unobserved choice-characteristics

$$\delta_{jt} = \ln s_{jt} - \ln s_{0t},$$

where we set $\delta_{0t} = 0$. (this is typically the outside good, whose average utility is normalized to zero). The residual is

$$\zeta_{jt} = \delta_{jt} - \beta' X_{jt} = \ln s_{jt} - \ln s_{0t} - \beta' X_{jt}.$$

With a set of instruments W_{jt}, we run the regression

$$\ln s_{jt} - \ln s_{0t} = \beta' X_{jt} + \epsilon_{jt},$$

using W_{jt} as instrument for X_{jt}, using as the observational unit the market share for product j in market t.

So here the technique is very transparent. It amounts to transforming the market shares to something linear in the coefficients so we can use two-stage-least-squares. More generally the transformation is going to be much more difficult with the random coefficients implying that there is no analytic solution. Computationally these things can get very complicated. Note however that we can estimate these models now without having individual level data,

and that at the same time we can get a fairly flexible model for the substitution patterns. At the same time you would expect to need a lot of structure to get the parameters precisely estimated just as in the other models. Of course if you compare the current model to the nested logit model you can impose such structure by imposing restrictions on the covariance matrix.

Comparisons of the models are difficult. Obviously if the structure imposed is correct it helps, but we typically do not know what the truth is, so we cannot conclude which one is better on the basis of the data typically available.

6. Models with Multiple Unobserved Choice Characteristics

The BLP approach allows for a single unobserved choice characteristic. This is essential for their estimation strategy that requires only market share data, and exploits a one-to-one relationship between market-specific unobserved product characteristics and market shares given other parameters and covariates. With individual level data one may be able to, and wish to allow for, multiple unobserved product characteristics. Elrod and Keane (1995), Goettler and Shachar (2001), and Athey and Imbens (2007), among others, study such models, in all cases with the unobserved choice characteristics constant across markets. Athey and Imbens model the latent utility for individual i in market t for choice j as

$$U_{ijt} = X_{it}'\beta_i + \zeta_j'\gamma_i + \epsilon_{ijt},$$

with the individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Delta Z_i, \Omega).$$

Even in the case with all choice characteristics exogenous, maximum likelihood estimation would be difficult. Athey and Imbens show that Bayesian methods, and in particular markov-chain-monte-carlo methods are effective tools for conducting inference in these settings.

7. Hedonic Models and the Motivation for a Choice and Individual Specific Error Term

Recently researchers have reconsidered using pure characteristics models for discrete choices, that is models with no idiosyncratic error ϵ_{ij}, instead relying solely on the presence of a few unobserved product characteristics and unobserved variation in taste parameters to generate stochastic choices. Such an error term is the only source of stochastic variation in the original multinomial choice models with only observed choice and individual characteristics, but in models with unobserved choice and individual characteristics their presence needs more motivation. Athey and Imbens (2007) discuss two arguments for including the additive error term.

First, the pure characteristics model can be extremely sensitive to measurement error, because it can predict zero market shares for some products. Consider a case where choices are generated by a pure characteristics model that implies that a particular choice j has zero market share. Now suppose that there is a single unit i for whom we observe, due to measurement error, the choice $Y_i = j$. Irrespective of the number of correctly measured observations available that were generated by the pure characteristics model, the estimates of the latent utility function will not be close to the true values corresponding to the pure characteristics model due to the single mismeasured observation. Such extreme sensitivity puts a lot of emphasis on the correct specification of the model and the absence of measurement error, and is undesirable in most settings.

Thus, one might wish to generalize the model to be robust against small amounts of measurement error of this type. One possibility is to define the optimal choice Y_i^* as the choice that maximizes the utility and assume that the observed choice Y_i is equal to the optimal choice Y_i^* with probability $1-\delta$, and with probability $\delta/(J-1)$ any of the other choices is observed:

$$\Pr(Y_i = y | Y_i^*, X_i, \nu_i, Z_1, \ldots, Z_J, \zeta_1, \ldots, \zeta_J) = \begin{cases} 1-\delta & \text{if } Y = Y_i^*, \\ \delta/(J-1) & \text{if } Y \neq Y_i^*. \end{cases}$$

This nests the pure characteristics model (by setting $\delta = 0$), without having the disad-

vantages of extreme sensitivity to mismeasured choices that the pure characteristics model has. If the true choices are generated by the pure characteristics model the presence of a single mismeasured observation will not prevent the researcher from estimating the true utility function. However, this specific generalization of the pure characteristics model has an unattractive feature: if the optimal choice Y_i^* is not observed, all of the remaining choices are equally likely. One might expect that choices with utilities closer to the optimal one are more likely to be observed conditional on the optimal choice not being observed.

An alternative modification of the pure characteristics model is based on adding an idiosyncratic error term to the utility function. This model will have the feature that, conditional on the optimal choice not being observed, a close-to-optimal choice is more likely than a far-from-optimal choice. Suppose the true utility is U_{ij}^* but individuals base their choice on the maximum of mismeasured version of this utility:

$$U_{ij} = U_{ij}^* + \epsilon_{ij},$$

with an extreme value ϵ_{ij}, independent across choices and individuals. The ϵ_{ij} here can be interpreted as an error in the calculation of the utility associated with a particular choice. This model does not directly nest the pure characteristics model, since the idiosyncratic error term has a fixed variance. However, it approximately nests it in the following sense. If the data are generated by the pure characteristics model with the utility function $g(x, \nu, z, \zeta)$, then the model with the utility function $\lambda \cdot g(x, \nu, z, \zeta) + \epsilon_{ij}$ leads, for sufficiently large λ, to choice probabilities that are arbitrarily close to the true choice probabilities (e.g., Berry and Pakes, 2007).

Hence, even if the data were generated by a pure characteristics model, one does not lose much by using a model with an additive idiosyncratic error term, and one gains a substantial amount of robustness to measurement or optimization error.

References

Ackerberg, D., L. Benkard, S. Berry, and A. Pakes, (2005), "Econometric Tools for Analyzing Market Outcomes," forthcoming, *Handbook of Econometrics*, Vol 5, Heckman and Leamer (eds.)

Amemiya, T., and F. Nold, (19750, "A Modified Logit Model," *Review of Economics and Statistics*, Vol 57(2), 255-257.

Athey, S., and G. Imbens, (2007), "Discrete Choice Models with Multiple Unobserved Product Characteristics," *International Economic Review*, forthcoming.

Bajari, P., and L. Benkard, (2004), "Demand Estimation with Heterogenous Consumers and Unobserved Product Characteristics: A Hedonic Approach," Stanford Business School.

Berry, S., (1994), "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, Vol. 25, 242-262.

Berry, S., J. Levinsohn, and A. Pakes, (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, Vol. 63, 841-890.

Berry, S., J. Levinsohn, and A. Pakes (2004), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, Vol 112(1), 68-105.

Berry, S., O. Linton, and A. Pakes, (2004), "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems ", *Review of Economic Studies*, Vol. 71, 613-654.

Berry, S., and A. Pakes, (2007), "The Pure Characteristics Discrete Choice Model of Differentiated Products Demand," *International Economic Review*, forthcoming.

Dempster, A., N. Laird, and D. Rubin, (1974), "Maximum Likelihood from Incomplete Data via the EM Algorithm", (with discussion), *Journal of the Royal Statistical*

Society, Series B, Vol. 39, 1-38.

Elrod, T., and M. Keane, (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data, "*Journal of Marketing Research*, Vol. XXXII, 1-16.

Geweke, J., M. Keane, and D. Runkle, (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 76, No 4, 609-632.

Gill, P., W. Murray, and M. Wright, (1981), *Practical Optimization*, Harcourt Brace and Company, London

Goettler, J., and R. Shachar (2001), "Spatial Competition in the Network Television Industry," *RAND Journal of Economics*, Vol. 32(4), 624-656.

Goldberg, P., (1995), "Product Differentiation and Oligopoly in International Markets: The Case of the Automobile Industry," *Econometrica*, 63, 891-951.

Hajivassiliou, V., and P. Ruud, (1994), "Classical Estimation Methods for LDV Models Using Simulation," in Engle and McFadden 9eds.), *Handbook of Econometrics*, Vol 4, Chapter 40, Elseviers.

Hajivassiliou, V., and D. McFadden, (1990, "The method of simulated scores,", with application to models of external debt crises," unpublished manuscript, Department of Economics, Yale University.

Heckman, J., and B. Singer, (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2).

Manski, C., and S. Lerman,, (1981) "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds.), 305-319, MIT Press, Cambridge, MA.

McCulloch, R., and P. Rossi, (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model, "*Journal of Econometrics* 64 207-240.

McCulloch, R., N. Polson, and P. Rossi, (2000) "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters, "*Journal of Econometrics* 99, 173-193.

McFadden, D., (1973), "Conditional Logit Analysis of Qualitative Choice Behavior "in P. Zarembka (ed), *Frontiers in Econometrics* Academic Press, New York 105-142.

McFadden, D., (1981) "Econometrica Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds.), 198-272, MIT Press, Cambridge, MA.

McFadden, D., (1982), "Qualitative Response Models," in Hildenbrand (ed.), *Advances in Econometrics*, Econometric Society Monographs, Cambridge University Press.

McFadden, D., (1984), "Econometric Analysis of Qualitative Response Models,"in Griliches and Intriligator (eds), *Handbook of Econometrics*, Vol. 2, 1395- 1457, Amsterdam.

McFadden, D., (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57(5), 995-1026.

McFadden, D., and K. Train, (2000), "Mxed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15(5), 447-470.

Nevo, A. (2000), "A Practioner's Guide to Estimation of Random-Coefficient Logit Models of Demand," *Journal of Economics & Management Science,* Vol. 9, No. 4, 513-548.

Pakes, A., and D. Pollard, (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027-1057.

What's New in Econometrics? **NBER**, **Summer 2007**

Lecture 12, **Wednesday**, **August 1st**, **11-12.30 pm**

Missing Data

These notes discuss various aspects of missing data in both pure cross section and panel data settings. We begin by reviewing assumptions under which missing data can be ignored without biasing estimation or inference. Naturally, these assumptions are tied to "exogenous" sampling.

We then consider three popular solutions to missing data: inverse probability weighting, imputation, and Heckman-type selection corrections. The first to methods maintain "missing at random" or "ignorability" or "selection on observables" assumptions. Heckman corrections, whether applied to cross section data or panel data, linear models or (certain) nonlinear models, allow for "selection on unobservables." Unfortunately, their scope of application is limited. An important finding is that all methods can cause more harm than good if selection is on conditioning variables that are unobserved along with response variables.

1. When Can Missing Data be Ignored?

It is easy to obtain conditions under which we can ignore the fact that certain variables for some observations, or all variables for some observations, have missing values. Start with a linear model with possibly endogenous explanatory variables:

$$y_i = x_i\beta + u_i, \tag{1.1}$$

where x_i is $1 \times K$ and the instruments z_i are $1 \times L$, $L \geq K$. We model missing data with a selection indicator, drawn with each i. The binary variable s_i is defined as $s_i = 1$ if we can use observation i, $s_i = 0$ if we cannot (or do not) use observation i. In the $L = K$ case we use IV on the selected sample, which we can write as

$$\hat{\beta}_{IV} = \left(N^{-1} \sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i z_i' y_i \right) \tag{1.2}$$

$$= \beta + \left(N^{-1} \sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i z_i' u_i \right) \tag{1.3}$$

For consistency, we essentially need

$$rank\, E(s_i z_i' x_i) = K \tag{1.4}$$

and

$$E(s_i z_i' u_i) = 0, \tag{1.5}$$

which holds if $E(z_i'u_i|s_i) = 0,$ which in turn is implied by

$$E(u_i|z_i, s_i) = 0. \tag{1.6}$$

Sufficient for (1.6) is

$$E(u_i|z_i) = 0, \ s_i = h(z_i) \tag{1.7}$$

for some function $h(\cdot)$. Note that the zero covariance assumption, $E(z_i'u_i) = 0$, is not sufficient for consistency when $s_i = h(z_i)$. A special case is when $E(y_i|x_i) = x_i\beta$ and selection s_i is a function of x_i. Provided the selected sample has sufficient variation in x, can consistently estimate β by OLS on the selected sample.

We can use similar conditions for nonlinear models. What is sufficient for consistency on the selected sample?

(Linear or Nonlinear) Least Squares: $E(y|x, s) = E(y|x)$.

Least Absolute Deviations: $Med(y|x, s) = Med(y|x)$

Maximum Likelihood: $D(y|x, s) = D(y|x)$.

All of these allow selection on x but not generally on y (or unobservables that affect y).

In the statistics literature, just using the data for which we observe all of (y_i, x_i, z_i) (or just (y_i, x_i) without instruments) is called the "complete case method." In cases where we model some feature of $D(y|x)$, it is clear that the richer is x, the more likely ignoring selection will not bias the results. In the case of estimating unconditional moments, say $\mu = E(y_i)$, unbiasedness and consistency of the sample on the selected sample requires $E(y|s) = E(y)$.

Similar conditions can be obtained for panel data. For example, if we model $D(y_t|x_t)$, and s_t is the indicator equal to one if (x_t, y_t) is observed, then the condition sufficient to ignore selection is

$$D(s_t|x_t, y_t) = D(s_t|x_t), \ t = 1, \ldots, T. \tag{1.8}$$

If, for example, x_t contains y_{t-1}, then selection is allowed to depend on the lagged response under (1.8). To see that (1.8) suffices, let the true conditional density be $f_t(y_{it}|x_{it}, \gamma)$. Then the partial log-likelihood function for a random draw i from the cross section can be written as

$$\sum_{t=1}^{T} s_{it} \log f_t(y_{it}|x_{it}, g) \equiv \sum_{t=1}^{T} s_{it} l_{it}(g). \tag{1.9}$$

Except for ensuring identifiability of γ, it suffices to show that $E[s_{it}l_{it}(\gamma)] \geq E[s_{it}l_{it}(g)]$ for all $g \in \Gamma$ (the parameter space). But by a well-known result from MLE theory – the Kulback-Leibler information inequality – γ maximizes $E[l_{it}(g)|x_{it}]$ for all x_{it}. But

$$E[s_{it}l_{it}(g)|x_{it}] = E\{E[s_{it}l_{it}(g)|y_{it},x_{it}]|x_{it}\} = E\{E(s_{it}|y_{it},x_{it})l_{it}(g)|x_{it}\}$$
$$= E\{E(s_{it}|x_{it})l_{it}(g)|x_{it}\} = E(s_{it}|x_{it})E[l_{it}(g)|x_{it}],$$

where we used $E(s_{it}|y_{it},x_{it}) = E(s_{it}|x_{it})$ from (1.8). Because $E(s_{it}|x_{it}) = P(s_{it} = 1|x_{it}) \geq 0$, it follows that $E[s_{it}l_{it}(\gamma)|x_{it}] \geq E[s_{it}l_{it}(g)|x_{it}]$ for all $g \in \Gamma$. Taking expectations of this inequality and using iterated expectations gives the result. Thus, we have shown that γ maximizes the expected value of each term in the summand in (1.9) – often not uniquely – and so it also maximizes the expected value of the sum. For identification, we have to assume it is the unique maximizer, as is usually the case of the model is identified in an unselected population and the selection scheme selects out "enough" of the population. One implication of this finding is that selection is likely to be less of a problem in dynamic models where lags of *y* and lags of other covariates appear, because then selection is allowed to be an arbitrary function of them. But, what is ruled out by (1.8) is selection that depends on idiosyncratic shocks to *y* between $t - 1$ and t.

Methods to remove time-constant, unobserved heterogeneity deserve special attention. Suppose we have the linear model, written for a random draw i,

$$yi_t = \eta_t + x_{it}\beta + c_i + u_{it}. \tag{1.10}$$

Suppose that we have instruments, say z_{it}, for x_{it}, including the possibility that $z_{it} = x_{it}$. If we apply random effects IV methods on the unbalanced panel, sufficient for consistency (fixed T) are

$$E(u_{it}|z_{i1},\ldots,z_{iT},s_{i1},\ldots,s_{iT},c_i) = 0,\ t = 1,\ldots,T \tag{1.11}$$

and

$$E(c_i|z_{i1},\ldots,z_{iT},s_{i1},\ldots,s_{iT}) = E(c_i) = 0, \tag{1.12}$$

along with a suitable rank condition. Somewhat weaker conditions suffice, but the general point is that selection must be strictly exogenous with respect to the idiosyncratic errors as well as the unobserved effect, c_i. If we use the fixed effects estimator on the unbalanced panel, we can get by with the first assumption, but, of course, all the instruments and selection to be arbitrarily correlated with c_i. To see why, consider the just identified case and define, say, $\ddot{y}_{it} = y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ and similarly for and $\ddot{x}_{it}$ and $\ddot{z}_{it}$, where $T_i = \sum_{r=1}^{T} s_{ir}$ is the number of time periods for observation i (properly viewed as random). The FEIV estimator is

$$\hat{\beta}_{FEIV} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{z}_{it}' \ddot{x}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{z}_{it}' \ddot{y}_{it} \right)$$
$$= \beta + \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{z}_{it}' \ddot{x}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{z}_{it}' u_{it} \right).$$

Because $\ddot{z}_{it}$ is a function of $(z_{i1}, \ldots, z_{iT}, s_{i1}, \ldots, s_{iT})$, (1.11) implies $\sum_{t=1}^{T} E(s_{it} \ddot{z}_{it}' u_{it}) = 0$ (as do weaker assumptions). There is a set of second moment assumptions that makes the usual, nonrobust inference procedures valid, but these impose homoskedasticity and serial independence of the u_{it} conditional on (z_i, s_i, c_i).

There are some simple ways to test for selection bias in panel data applications. One important violation of (1.11) is when units drop out of the sample in period $t + 1$ because of shocks realized in time t. This generally induces correlation between $s_{i,t+1}$ and u_{it}. A simple test in the FE environment is to simply add $s_{i,t+1}$ to the equation at time t, and estimate the resulting model by fixed effects (or FEIV). A simple t test can be used (probably fully robust). Of course the test entails dropping the last time period, and it need not have power for detecting correlation between s_{it} and u_{it} – that is, contemporaneous selection.

The consistency of FE (and FEIV) on the unbalanced panel under (1.11) breaks down if the slope coefficients are random but one ignores this in estimatin. That is, replace β with b_i but still use the FE estimator. Then the error term contains the term $x_i d_i$ where $d_i = b_i - \beta$. If selection is a function of d_i, then the usual FE estimator will be inconsistent. (Recall that the FE estimator, on balanced panels, has some robustness to random slopes.) A simple test is to allow d_i to be correlated with selection through the number of available time periods, T_i. The idea is to consider alternatives with

$$\mathrm{E}(b_i | z_{i1}, \ldots, z_{iT}, s_{i1}, \ldots, s_{iT}) = E(b_i | s_{i1}, \ldots, s_{iT}) = E(b_i | T_i). \tag{1.13}$$

Then, add interaction terms of dummies for each possible sample size (with $T_i = T$ as the base group),

$$1[T_i = 2]x_{it}, 1[T_i = 3]x_{it}, \ldots, 1[T_i = T - 1]x_{it} \tag{1.14}$$

to the equation and estimate it by FE. Significance of these interaction terms indicates that random slopes are correlated with the available time periods, and suggests one might have to remove those random slopes (if possible).

If we first difference instead to remove c_i – a method that has important advantages for

attrition problems – we can apply the pooled IV results:

$$\Delta y_{it} = \varphi_t + \Delta x_{it} + \Delta u_{it}, \; t = 2, \ldots, T \tag{1.15}$$

and, if z_{it} is the set of IVs at time t, we can use

$$E(\Delta u_{it} | z_{it}, s_{it}) = 0 \tag{1.16}$$

as being sufficient to ignore the missingess. Again, can add $s_{i,t+1}$ to test for attrition.

Not suprisingly, nonlinear models with unosberved effects are considerably more difficult to handle, although certain conditional MLEs (logit, Poisson) can accomodate selection that is arbitrarily correlated with the unobserved effect.

2. Inverse Probability Weighting

2.1. Weighting with Cross-Sectional Data

A general solution to solving missing data problems when selection is not exogenous is based on probability weights. To illustrate, for simplicity, suppose y is a random variable whose population mean $\mu = E(y)$ we would like to estimate, but some observations are missing on y. Let $\{(y_i, s_i, z_i) : i = 1, \ldots, N\}$ indicate independent, identically distributed draws from the population, where z_i is a vector that, for now, we assume is always observed. Suppose we assume the "selection on observables" assumption

$$P(s = 1 | y, z) = P(s = 1 | z) \equiv p(z) \tag{2.1}$$

where $p(z) > 0$ for all possible values of z. Then we can solve the missing data problem by weighting the observed data points by $1/p(z_i)$:

$$\tilde{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{p(z_i)} \right) y_i, \tag{2.2}$$

where note that s_i selects out the observed data points. It is easy to show, using iterated expectations, that $\hat{\mu}_{IPW}$ is not only consistent for y_i, it is unbiased, too. (This same kind of estimator arises in treatment effect estimation.) Of course, except in special cases, we must estimate $p(z_i)$; when z_i is always observed along with s_i, flexible binary response models such as logit or probit, or nonparametric methods, can be used. Let $\hat{p}(z_i)$ denote the estimated selection probability (also called the propensity score). Then an operational estimator is

$$\hat{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(z_i)} \right) y_i. \tag{2.3}$$

As written, this estimator assumes we know the size of the random sample, N, which is not

necessarily the case for some sampling schemes, such as variable probability sampling. We can also write $\hat{\mu}_{IPW}$ as

$$\hat{\mu}_{IPW} = N_1^{-1}(N_1/N)\sum_{i=1}^{N}\left(\frac{s_i}{\hat{p}(z_i)}\right)y_i = N_1^{-1}\sum_{i=1}^{N} s_i\left(\frac{\hat{\rho}}{\hat{p}(z_i)}\right)y_i \tag{2.4}$$

where $N_1 = \sum_{i=1}^{N} s_i$ is the number of selected observations and $\hat{\rho} = N_1/N$ is a consistent estimate of $P(s_i = 1)$. The weights reported to account for missing data are often $\hat{\rho}/\hat{p}(z_i)$, which can be greater or less than unity. (By iterated expectatins, $\rho = E[p(z_i)]$.) Equation (2.4) shows that $\hat{\mu}_{IPW}$ is a weighted average of the observed data points with weights $\hat{\rho}/\hat{p}(z_i)$. Yet a different estimator is obtained by solving the least squares problem

$$\min_m \sum_{i=1}^{N}\left(\frac{s_i}{\hat{p}(z_i)}\right)(y_i - m)^2,$$

which results in

$$\breve{\mu}_{IPW} = \left(\sum_{h=1}^{N}\frac{s_h}{\hat{p}(z_h)}\right)^{-1}\left(\sum_{i=1}^{N}\left(\frac{s_i}{\hat{p}(z_i)}\right)y_i\right), \tag{2.5}$$

which is a different version a weighted.average.

Horowitz and Manski (1998) have considered the problem of estimating population means using IPW. Their main focus is on establishing bounds that do not rely on potentially strong, untestable assumptions such as the unconfoundedness assumption in (2.1). But they also note a particular problem with certain IPW estimators even when the conditioning variable, x, is always observed. They consider estimation of the mean $E[g(y)|x \in A]$ for some set A. If we define $d_i = 1[x_i \in A]$ then the problem is to estimate $E[g(y)|d = 1]$. HM point out that, if one uses the weights commonly reported with survey data – weights that do not condition on the event $d = 1$ – then the IPW estimate of the mean can lie outside the logically possible values of $E[g(y)|d = 1]$. HM note that this problem can be fixed by using probability weights $P(s = 1|d = 1)/P(s = 1|d = 1, z)$ unfortunately, this choice is not possible when data on x can also be missing.

Failure to condition on $d = 1$ when computing the probability weights when interest lies in $E[g(y)|d = 1]$ is related to a general problem that arises in estimating models of conditional means when data are missing on x. To see why, suppose the population regression function is linear:

$$E(y|x) = \alpha + x\beta. \tag{2.6}$$

Let z be a variables that are always observed and let $p(z)$ be the selection probability, as before. Now, suppose that at least part of x is not always observed, so that x is not a subset of z. This means that some elements of x cannot appear in $p(z)$ because $p(z)$ normally has to be estimated using the data on (s_i, z_i) for all i. The IPW estimator of β solves

$$\min_{a,b} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(z_i)} \right) (y_i - a - x_i b)^2. \tag{2.7}$$

Here is the problem: suppose that selection is exogenous in the sense that

$$P(s = 1|x,y) = P(s = 1|x). \tag{2.8}$$

Then we saw in Section 1 that using least squares on the selected sample results in a consistent estimator of $\theta = (\alpha, \beta')'$, which is also $\sqrt{N}$-asymptotically normal. What about the weighted estimator? The problem is that if (2.8) holds, and z does not include x, then it is very unlikely that

$$P(s = 1|x,y,z) = P(s = 1|z). \tag{2.9}$$

In other words, the key unconfoundedness assumption fails, and the IPW estimator of θ is generally inconsistent. We actually introduce inconsistency by weighting when a standard unweighted regression on the complete cases would be consistent. In effect, the IPW estimator uses weights that are functions of the wrong variables.

If x is always observed and can (and should) be included in z, then weighting is much more attractive. Typically, z might contain lagged information, or interview information that would not be included in x. If it turns out that selection is a function only of x, flexible estimation of the model $P(s = 1|z)$ will pick that up in large sample sizes.

If x is always observed and we know that $P(s = 1|x,y) = P(s = 1|x)$, is there any reason to weight by $1/p(x)$? If $E(y|x) = \alpha + x\beta$ and $Var(y|x)$, weighting is asymptotically inefficient. If $E(y|x) = \alpha + x\beta$ but $Var(y|x)$ is heteroskedastic, then weighting may or may not be more efficient than not. (The efficient estimator would be the WLS estimator that appropriately accounts for $Var(y|x)$, different issue than probability weighting.) But both weighting and not weighting are consistent. The advantage of weighting is that, if the population "model" is in fact just a linear projection, the IPW estimator consistently estimates that linear projection while the unweighted estimator does not. In other words, if we write

$$L(y|1,x) = \alpha^* + x\beta^* \tag{2.10}$$

where $L(\cdot|\cdot)$ denotes the linear projection, then under $P(s = 1|x,y) = P(s = 1|x)$, the IPW estimator is consistent for θ^*. The unweighted estimator has a probabilty limit that depends on $p(x)$.

One reason to be interested in the LP is that the parameters of the LP show up in certain treatment effect estimators. The notes on treatment effects contained a discussion of a "double robustness" result due to Robins and Ritov (1997); see also Wooldridge (2007). The idea is this. In treatment effect applications, the ATE requires estimation of $E(y_g)$ for the two counterfactual outcomes, $g = 0, 1$). The LP has the property that $E(y_1) = \alpha_1^* + E(x)\beta_1^*$, and so, if we consistently estimate α_1^* and β_1^* then we can estimate $E(y_1)$ by averaging across x. A similar statement holds for y_0. Now, the IPW estimator identifies α_1^* and β_1^* if the model for $p(x)$ is correctly specified. But if $E(y_1|x) = \alpha_1 + x\beta_1$ then the IPW estimator is consistent for α_1 and β_1 even if $p(x)$ is misspecified. And, of course, $E(y_1) = \alpha_1 + E(x)\beta_1$. So, regardless of whether we are estimating the conditional mean parameters or the LP parameters, we consistently estimate $E(y_1)$. The case where the IPW estimator does not consistently estimate $E(y_1)$ is when $E(y_1|x)$ is not linear and $p(x)$ is misspecified.

The double robustness result holds for certain nonlinear models, too, although one must take care in combining the conditional mean function with the proper objective function – which, in this case, means quasi-log-likelihood function. The two cases of particular interest are the logistic response function for binary or fractional responses coupled with the Bernoulli QLL, and the exponential response function coupled with the Poisson QLL.

Returning to the IPW regression estimator that solves (2.7), suppose we assume the ignorability assumption (2.9),

$$E(u) = 0, E(x'u) = 0,$$

$$p(z) = G(z,\gamma)$$

for a parametric function $G(\cdot)$ (such as flexible logit), and $\hat{\gamma}$ is the binary response MLE. Then, as shown by Robins, Rotnitzky, and Zhou (1995) and Wooldridge (2007), the asymptotic variance of $\hat{\theta}_{IPW}$, using the estimated probability weights, is

$$Avar\sqrt{N}\,(\hat{\theta}_{IPW} - \theta) = [E(x_i'x_i)]^{-1}E(r_i r_i')[E(x_i'x_i)]^{-1}, \tag{2.11}$$

where r_i is the $P \times 1$ vector of population residuals from the regression $(s_i/p(z_i))x_i'u_i$ on d_i', where d_i is the $M \times 1$ score for the MLE used to obtain $\hat{\gamma}$. The asymptotic variance of $\hat{\theta}_{IPW}$ is

easy to estimate:

$$\left(\sum_{i=1}^{N}[s_i/G(z_i,\hat{\gamma})]x_i'x_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{r}_i\hat{r}_i'\right)\left(\sum_{i=1}^{N}[s_i/G(z_i,\hat{\gamma})]x_i'x_i\right)^{-1}, \tag{2.12}$$

or, if x_i is always observed, the terms $s_i/G(z_i,\hat{\gamma})$ can be dropped in the outer parts of the sandwich. In the case that d_i is the score from a logit model of s_i on functions, say, $h(z_i)$, $\hat{d}_i$ has the simple form

$$\hat{d}_i = h_i'(s_i - \Lambda(h_i\hat{\gamma})), \tag{2.13}$$

where $\Lambda(a) = \exp(a)/[1+\exp(a)]$ and $h_i = h(z_i)$. This illustrates a very interesting finding of Robins, Rotnitzky, and Zhou (1995) and related to the Hirano, Imbens, and Ritter (2003) efficient estimator for means using IPW estimators. Suppose that, for a given set of functions h_{i1}, the logit model is correctly specified in the sense that there is a γ_1 such that $P(s_i = 1|z_i) = \Lambda(h_{i1}\gamma_1)$. Now suppose we take some additional functions of z_i, say $h_{i2} = h_2(z_i)$, and add them to the logit. Then, asymptotically, the coefficients on h_{i2} are zero, and so the adjustment to the asumptotic variance comes from regressing $(s_i/\Lambda(h_{i1}\gamma_1)x_i'u_i$ on $(h_{i1}, h_{i2})[s_i - \Lambda(h_{i1}\gamma_1)]$. Now, notice that, even though the coefficients on h_{i2} are zero in the logit model, the score vector depends on (h_{i1}, h_{i2}). Therefore, the residual variance from regressing $(s_i/\Lambda(h_{i1}\gamma_1)x_i'u_i$ on $(h_{i1}, h_{i2})[s_i - \Lambda(h_{i1}\gamma_1)]$ is generally smaller than that from using the correct logit model, which is obtained from regressing on $h_{i1}[s_i - \Lambda(h_{i1}\gamma_1)]$. By overspecifying the logit model for s_i, we generally reduce the asymptotic variance of the IPW estimator. And the process does not stop there. We can keep adding functions of z_i to the logit to reduce the asymptotic variance of the estimator of the IPW estimator. In the limit, if we have chosen the sequence of functions so that they approximate any well-behaved function, then we achieve asymptotic efficiency. This is precisely what the HIR estimator does by using a logit series estimator for the propensity score.

Wooldridge (2007) shows that the adjustment to the asymptotic variance in (2.12) carries over to general nonlinear models and estimation methods. One consequence is that ignoring the estimation in $\hat{p}(z)$ – as commercial software typically does when specifying probability weights – results in conservative inference. But the adjustment to obtain the correct asymptotic variance is fairly straightforward.

Nevo (2003) explicitly considers a generalized method of moments framework, and shows how to exploit known population moments to allow selection to depend on selected elements

of the data vector w. (Hellerstein and Imbens (1998) use similar methods to improve estimation when population moments are known.) In particular, Nevo assumes that, along with the moment conditions $E[r(w,\theta)] = 0$, the population moments of the vector $h(w)$, say μ_h, are known. Under the assumption that selection depends on $h(w)$, that is, $P(s = 1|w) = P(s = 1|h(w))$, Nevo obtains an expanded set of moment conditions that can be used to estimate θ and the parameters γ in the selection equation. Suppose we use a logit model for $P(s = 1|h(w))$. Then

$$E\left[\frac{s_i}{\Lambda(h(w_i)\gamma)} r(w_i,\theta) \right] = 0 \tag{2.14}$$

and

$$E\left[\frac{s_i h(w_i)}{\Lambda(h(w_i)\gamma)} \right] = \mu_h. \tag{2.15}$$

Equation (2.15) generally identifies γ, and then this $\hat{\gamma}$ can be used in a second step to choose $\hat{\theta}$ to make the weighted sample moments

$$N^{-1} \sum_{i=1}^{N} \left[\frac{s_i}{\Lambda(h(w_i)\hat{\gamma})} r(w_i,\hat{\theta}) \right] \tag{2.16}$$

as close to zero as possible. Because (2.15) adds as many moment restrictions as parameters, the GMM estimator using both sets of moment conditions is equivalent to the two-step estimator just described.

Another situation where the missing data problem can be solved via weighting is when data have been censored due to a censored duration. The response variable of interest may be the duration, or it may be a variable observed only if a duration or survival time is observed. Let y be a univariate response and x a vector of conditioning variables, and suppose we are interested in estimating $\mathrm{E}(y|x)$. A random draw i from the population is denoted (x_i, y_i). Let $t_i > 0$ be a duration and let $c_i > 0$ denote a censoring time (where $t_i = y_i$ is allowed). Assume that (x_i, y_i) is observed whenever $t_i \leq c_i$, so that $s_i = 1(t_i \leq c_i)$. Under the assumption that c_i is independent of (x_i, y_i, t_i),

$$\mathrm{P}(s_i = 1|x_i, y_i, t_i) = G(t_i), \tag{2.17}$$

where $G(t) \equiv \mathrm{P}(c_i \geq t)$. In order to use inverse probability weighting, we need to observe t_i whenever $s_i = 1$, which simply means that t_i is uncensored. Plus, we need only observe c_i when $s_i = 0$. of c_i. As shown in Wooldridge (2007), it is more efficient to estimate $G(\cdot)$ using

the density of $\min(c_i, t_i)$ given t_i. Generally, let $h(c, \gamma)$ denote a parametric model for the density of the censoring times, c_i, and let $G(t, \gamma)$ be the implied model for $P(c_i \geq t)$. The log likelihood is

$$\sum_{i=1}^{N} \{(1 - s_i) \log[h(c_i, \gamma)] + s_i \log[G(t_i, \gamma)]\}, \tag{2.18}$$

which is just the log-likelihood for a standard censored estimation problem but where t_i (the underlying duration) plays the role of the censoring variable. As shown by Lancaster (1990 for grouped duration data, where $h(c, \gamma)$ is piecewise constant, the solution to (2.18) gives a survivor function identical to the Kaplan-Meier estimator but where the roles of c_i and t_i are reversed; that is, we treat t_i as censoring c_i. The linear regression model has a long history, and has been studied recently by Honoré, Khan, and Powell (2002). See also Rotnitzky and Robins (2005) for a survey of how to obtain semiparametrically efficient estimators. The Koul-Susarla-van Ryzin (1981) estimator is an IPW least squares estimator, but their proposals for inference are very difficult to implement. As shown by Wooldridge (2007), this is another instance where estimating the selection probability by MLE is more efficient than using the known probability (if you could). Plus, obtaining the smaller variance matrix involves only a multivariate regression of the weighted score for the second stage problem – OLS, NLS, MLE, or IV – on the score for the first-stage Kaplan-Meier estimation. This simple procedure is valid when the distribution of c_i is taken to be discrete. Other authors undertake the asymptotics allowing for an underlying continuous censoring time, which makes estimating asymptotic variances considerably more difficult.

2.2 Attrition in Panel Data

Inverse probability weighting can be applied to solve, in some cases, the attrition problem in panel data. For concreteness, consider maximum pooled maximum likelihood, where we model a density $f_t(y_t|\mathbf{x}_t)$ for any conditioning variables $\mathbf{x}_t$. These need not be strictly exogenous or always observed. Let $f_t(y_t|\mathbf{x}_t, \theta)$ be the parametric model, and let s_{it} be the selection indicator. We assume that attrition is absorbing, so $s_{it} = 1 \Rightarrow s_{ir} = 1$ for $r < t$. The estimator that ignores attrition solves

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta}), \tag{2.19}$$

which is consistent if $P(s_{it} = 1|y_{it}, \mathbf{x}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$. This follows by showing

$E[P(s_{it} = 1|\mathbf{x}_{it})E[\log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}]$, and using the fact that the true value of θ maximizes $E[\log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta})|\mathbf{x}_{it}]$ for all t, and $P(s_{it} = 1|\mathbf{x}_{it}) \geq 0$. But, if selection depends on y_{it} even after conditioning on $\mathbf{x}_{it}$, the unweighted estimator is generally inconsistent. If $\mathbf{w}_{it} = (\mathbf{x}_{it}, y_{it})$, then perhaps we can find variables $\mathbf{r}_{it}$, such that

$$P(s_{it} = 1|\mathbf{w}_{it}, \mathbf{r}_{it}) = P(s_{it} = 1|\mathbf{r}_{it}) \equiv p_{it} > 0, t = 1, \ldots, T. \tag{2.20}$$

(The "obvious" set of variables $\mathbf{r}_{it} = \mathbf{w}_{it}$ is not usually available since we will have estimate the probabilities.) If we could observe the p_{it}, we could use the weighted MLE,

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \sum_{t=1}^{T} (s_{it}/p_{it}) \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta}), \tag{2.21}$$

which we call $\hat{\theta}_w$. The estimator $\hat{\theta}_w$ is generally consistent because

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it}, \boldsymbol{\theta})] \; = E[q_t(\mathbf{w}_{it}, \boldsymbol{\theta})], t = 1, \ldots, T, . \tag{2.22}$$

where $q_t(\mathbf{w}_{it}, \boldsymbol{\theta}) = \log f_t(y_{it}|\mathbf{x}_{it}, \boldsymbol{\theta})$ is the objective function.

How do we choose $\mathbf{r}_{it}$ to make (2.20) hold (if possible)? A useful strategy, considered by RRZ, is to build the p_{it} up in a sequential fashion. At time t, $\mathbf{z}_{it}$ is a set of variables observed for the subpopulation with $s_{i,t-1} = 1$. ($s_{i0} \equiv 1$ by convention). Let

$$\pi_{it} = P(s_{it} = 1|z_{it}, s_{i,t-1} = 1), t = 1, \ldots, T. \tag{2.23}$$

Typically, $\mathbf{z}_{it}$ contains elements from $(\mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i1})$, and perhaps variables dated at $t - 1$ or earlier that do not appear in the population model. Unfortunately, $\mathbf{z}_{it}$ rarely can depend on time-varying variables that are observed in period t (since we have to apply a binary response model for the sample with $s_{i,t-1} = 1$, and this includes units that have left the sample at time t!) Given the monotone nature of selection, we can estimate models for π_{it} sequential when the $\mathbf{z}_{it}$ are observed for every unit in the sample at time $t - 1$.

How do we obtain p_{it} from the π_{it}? Not without some assumptions. Let $\mathbf{v}_{it} = (\mathbf{w}_{it}, \mathbf{z}_{it}), t = 1, \ldots, T$. An ignorability assumption that works is

$$P(s_{it} = 1|\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT}, s_{i,t-1} = 1) = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1), t \geq 1. \tag{2.24}$$

That is, given the entire history $\mathbf{v}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT})$, selection at time t (given being still in the sample at $t - 1$) depends only on z_{it}; in practice, this means only on variables observed at $t - 1$. This is a strong assumption; RRZ (1995) show how to relax it somewhat in a regression framework with time-constant covariates. Using this assumption, we can show that

$$p_{it} \equiv P(s_{it} = 1|\mathbf{v}_i) = \pi_{it}\pi_{i,t-1} \cdot \cdot \cdot \pi_{i1}. \tag{2.25}$$

In the general framework, we have $\mathbf{r}_{it} = (\mathbf{z}_{it}, \ldots, \mathbf{z}_{i1})$ but, because of the ignorability assumption, it is as if we can take $\mathbf{r}_{it} = [(\mathbf{w}_{i1}, \mathbf{z}_{i1}), \ldots, (\mathbf{w}_{iT}, \mathbf{z}_{it})]$.

So, a consistent two-step method is:

(1) In each time period, estimate a binary response model for $P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1)$, which means on the group still in the sample at $t - 1$. The fitted probabilities are the $\hat{\pi}_{it}$. Form $\hat{p}_{it} = \hat{\pi}_{it}\hat{\pi}_{i,t-1} \cdot \cdot \cdot \hat{\pi}_{i1}$. Note that we are able to compute $\hat{p}_{it}$ only for units in the sample at time $t - 1$.

(2) Replace p_{it} with $\hat{p}_{it}$ in (2.21), and obtain the weighted M-estimator.

Consistency is straightforward – standard two-step estimation problem – if we have the correct functional form and the ignorability of selection assumption holds. As shown by RRZ (1995) in the regression case, it is more efficient to estimate the p_{it} than to use know weights, if we could. See RRZ (1995) and Wooldridge (2002) for a simple regression method for adjusting the score; it is similar to that used for the cross section case, but just pooled across t.

IPW for attrition suffers from a similar drawback as in the cross section case. Namely, if $P(s_{it} = 1|\mathbf{w}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$ then the unweighted estimator is consistent. If we use weights that are not a function of $\mathbf{x}_{it}$ in this case, the IPW estimator is generally inconsistent: weighting uncessesarily causes inconsistency.

Related to the previous point is that it would be rare to apply IPW in the case of a model with completely specified dynamics. Why? Suppose, for example, we have a model of $E(y_{it}|x_{it}, y_{i,t-1}, \ldots, x_{i1}, y_{i0})$. If our variables affecting attrition, z_{it}, are functions of $(y_{i,t-1}, \ldots, x_{i1}, y_{i0})$ – as they often must be – then selection is on the basis of conditioning variables, and so the unweighted estimator is also consistent. RRZ (1995) explicitly cover regressions that do not have correct dynamics.

3. Imputation

Section 1 discussed conditions under which dropping observations with any missing data results in consistent estimators. Section 2 showed that, under an unconfoundedness assumption, inverse probability weighting can be applied to the complete cases to recover population parameters. One problem with using IPW for models that contain covariates is that the weighting may actually hurt more than it helps if the covariates are sometimes missing and selection is largely a function of those covariates.

A different approach to missing data is to try to fill in the missing values, and then analyze the resulting data set as a complete data set. Imputation methods, and multiple imputation use

either means, fitted values, values or averages from "similar" observations, or draws from posterior distributions to fill in the missing values. Little and Rubin (2002) provides an accessible treatment with lots of references to work by Rubin and coauthors.

Naturally, such procedures cannot always be valid. Most methods depend on a *missing at random* (MAR) assumption. When data are missing on only one variable – say, the response variable, y – MAR is essentially the same as the unconfoundedness assumption $P(s = 1|y,x) = P(s = 1|x)$. (The assumption *missing completely at random* (MCAR) is when s is independent of $w = (x,y)$.) MAR can be defined for general missing data patterns. For example, in a bivariate case, let $w_i = (w_{i1}, w_{i2})$ be a random draw from the population, where data can be missing on either variable. Let $r_i = (r_{i1}, r_{i2})$ be the "retention" indicators for w_{i1} and w_{i2}, so $r_{ig} = 1$ implies w_{ig} is observed. The MCAR assumption is that r_i is independent of w_i, so $D(r_i|w_i) = D(r_i)$. The MAR assumption is that implies $P(r_{i1} = 0, r_{i2} = 0|w_i) = P(r_{i1} = 0, r_{i2} = 0) \equiv \pi_{00}$, $P(r_{i1} = 1, r_{i2} = 0|w_{i1})$, $P(r_{i1} = 0, r_{i2} = 1|w_{i2})$, and then $P(r_{i1} = 1, r_{i2} = 1|w_i) = 1 - \pi_{00} - P(r_{i1} = 1, r_{i2} = 0|w_{i1}) - P(r_{i1} = 0, r_{i2} = 1|w_{i2})$. Even with just two variables, the restrictions imposed by MAR are not especially appealing, unless, of course, we have good reason to just assume MCAR.

MAR is more natural with monotone missing data problems, which sometime apply in panel data situations with attrition. Order the w_{ig} so that if w_{ih} is observed the so is w_{ig}, $g < h$. Then the retention indicators satisfy $r_{ig} = 1 \Rightarrow r_{i,g-1} = 1$. Under MAR, the joint density $f(w_1, \ldots, w_G)$ is easy to estimate. Write $f(w_1, \ldots, w_G) = f(w_G|w_{G-1}, \ldots, w_1) \cdot f(w_{G-1}|w_{G-1}, \ldots, w_1) \cdots f(w_2|w_1) f(w_1)$. Given parametric models, we can write partial log likelihood as

$$\sum_{g=1}^{G} r_{ig} \log f(w_{ig}|w_{i,g-1}, \ldots, w_{i1}, \theta), \tag{3.1}$$

where $f(w_1|w_0, \theta) \equiv f(w_1|w_0, \theta)$, and it suffices to multiply only by r_{ig} because $r_{ig} = r_{ig} r_{i,g-1} \cdots r_{i2}$. Under MAR,

$$E(r_{ig}|w_{ig}, \ldots, w_{i1}) = E(r_{ig}|w_{i,g-1}, \ldots, w_{i1}), \tag{3.2}$$

and so by (3.2),

$$E[r_{ig} \log f(w_{ig}|w_i^{(g-1)}\theta)|w_i^{(g-1)}] = E(r_{ig}|w_i^{(g-1)}) E[\log f(w_{ig}|w_i^{(g-1)}\theta)|w_i^{(g-1)}]. \tag{3.3}$$

The first term on the RHS of (3.3) is $E(r_{ig}|w_i^{(g-1)}) = P(r_{ig} = 1|w_i^{(g-1)}) \geq 0$ and the true value of

θ maximizes the second part by the conditional Kullback-Leibler information inequality (for example, Wooldridge (2002, Chapter 13)). Therefore, the parameters of the conditional densities are generally identified, provided the missing data problem is not too severe.

Before briefly describing how multiple imputation works, a simple example helps illustrate the general idea behind imputation. Suppose y is a random variable in a population with mean μ_y, but data are missing on some y_i random drawn from the population. Let s_i be the binary selection indicator, and let $\mathbf{x}_i$ be a set of observed covariates. So, a random draw consists of $(\mathbf{x}_i, y_i, s_i)$ but where y_i is missing if $s_i = 0$. As we discussed earlier, unless s is independent of y – that is, the data are MCAR – the complete-case sample average,

$$\tilde{\mu}_y = \left(\sum_{i=1}^{N} s_i \right)^{-1} \sum_{i=1}^{N} s_i y_i, \tag{3.4}$$

is not unbiased or consistent for μ_y; its probability limit is, of course, $E(y|s = 1)$.

Suppose, however, that the selection is ignorable conditional on $\mathbf{x}$:

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}), \tag{3.5}$$

where $m(\mathbf{x}, \boldsymbol{\beta})$ is, for simplicity, a parametric function. As we discussed in Section 1, nonlinear least squares, and a variety of quasi-MLEs, are consistent for $\boldsymbol{\beta}$ using the selected sample. Now, because we observe $\mathbf{x}_i$ for all i, we can obtained fitted values, $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, for any unit it the sample. Let $\hat{y}_i = s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ be the imputed data. Then an imputation estimator of μ_y is

$$\hat{\mu}_y = N^{-1} \sum_{i=1}^{N} \{ s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \}. \tag{3.6}$$

The plim of $\hat{\mu}_y$ is easy to find by replacing $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}$ and sample average with the population average:

$$\begin{aligned} E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] &= E[E(s_i y_i | \mathbf{x}_i, s_i)] + E[(1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \\ &= E[s_i E(y_i | \mathbf{x}_i, s_i)] + E[(1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \\ &= E[s_i m(\mathbf{x}_i, \boldsymbol{\beta})] + E[(1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \\ &= E[m(\mathbf{x}_i, \boldsymbol{\beta})] = \mu_y. \end{aligned} \tag{3.7}$$

(Of course, we could average the $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ across all i, but that would throw away some information on the y_i that we observe.)

If $D(y|\mathbf{x}, s) = D(y|\mathbf{x})$ then we can use MLE on the complete cases, obtain estimates of the

parameters, say $\hat{\boldsymbol{\theta}}$, and then use $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ as above, where $m(\mathbf{x}, \boldsymbol{\beta})$ is the mean function implied by the model for $D(y|\mathbf{x})$. For example, y could be a corner solution response and then we use a Tobit or some flexible extension for $D(y|\mathbf{x})$.

One danger in using even simple imputation methods like the one just covered is that we will ignore the errors in the imputed values. $\hat{y}_i$ differs from y_i for two reasons. First, if we write

$$y_i = m(\mathbf{x}_i, \boldsymbol{\beta}) + u_i, \tag{3.8}$$

then, even if we knew $\boldsymbol{\beta}$, the error would be u_i. (In effect, we are replace y_i with its conditional expectation.) Having to estimate $\boldsymbol{\beta}$ further introduces estimation error. Analytical formulas can be worked out, but bootstrapping a standard error or confidence interval for $\hat{\mu}_y$ is also straightforward: we would draw observation indices at random, without replacement, and perform the imputation steps on each new bootstrap sample.

As an example of how just using the imputed values as if they were real data, suppose we run a linear regression using the complete data and obtain $\mathbf{x}_i\hat{\boldsymbol{\beta}}$. Again defining $\hat{y}_i = s_i y_i + (1 - s_i)\mathbf{x}_i\hat{\boldsymbol{\beta}}$, suppose we use the imputed data set to reestimate $\boldsymbol{\beta}$. It is well known that we just get $\hat{\boldsymbol{\beta}}$ back again. However, our estimated error variance will be too small because every residual for an imputed data point is identically zero. It follows that, while $SSR/(N_1 - K)$ is generally unbiased for σ_u^2 (under the Gauss-Markov assumptions), where N_1 is the number of complete cases, $SSR/(N - K)$ has a downward bias.

The previous method ignores the random error in (3.4); Little and Rubin (2002) call it the method of "conditional means." Generally, as they show in Table 4.1, the method of conditional means results in downward bias in estimating variances. Instead, LR propose adding a random draw to $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ to impute a value. Of course, this entails have a distribution from which to draw the u_i. If we assume that u_i is independent of $\mathbf{x}_i$ and normally distributed, then we can draw, say, $\check{u}_i$ from a Normal$(0, \hat{\sigma}_u^2)$, distribution, where $\hat{\sigma}_u^2$ is estimated using the complete case nonlinear regression residuals. This procedure works well for estimating σ_y^2 in the case where linear regression is used and $(\mathbf{x}_i, y_i)$ is jointly normal. LR refer to this as the "conditional draw" method of imputation, which is a special case of stochastic imputation.

Little and Rubin argue that the conditional draw approach, at least in the jointly normal case, works well when a covariate is missing. Suppose that $\mathbf{x} = (x_1, x_2)$ and data are missing on x_2 but not (x_1, y). One possibility for imputing x_{i2} when it is missing is to regress x_{i2} on x_{i1} using the complete cases, and then use fitted values, or conditional draws, to impute x_{i2}. LR show that the method of conditional draws (not conditional means) works well when y is

included along with x_1 in obtained the estimated conditional means from the complete-case regression.

Unfortunately, except in simple cases, it is difficult to quantity the uncertainty from single-imputation methods, where one imputed values is obtained for each missing variable. One possibility, which has been studied in the statistics literature, is to bootstrap the entire estimation method – assuming, of course, that the imputations eliminates the nonresponse bias (so that missing at random holds). In the example of conditional draws above, the imputation procedure is simply included in any subsequent estimation, and bootstrap samples are obtained over and over again. On each bootstrap replication, say *b*, an estimate of the parameters using the complete cases, $\hat{\boldsymbol{\theta}}_{complete}^{(b)}$ is obtained (which would be the beta hats and error variance estimate in the regression case), missing data values are imputed using conditional draws, and then an estimate of $\boldsymbol{\theta}$ using the imputed data, $\hat{\boldsymbol{\theta}}_{imputed}^{(b)}$, can be obtained. Of course, this can be computationally intensive for nonlinear estimation problems.

An alternative is the method of multiple imputation. Its justification is Bayesian, and based on obtaining the posterior distribution – in particular, mean and variance – of the parameters conditional on the observed data. For general missing data patterns, the computation required to impute missing values is quite complicated, and involves simulation methods of estimation. LR and Cameron and Trivedi (2005) provide discussion. But the idea is easily illustrated using the above example. Rather than just impute one set of missing values to create one "complete" data set, created several imputed data sets. (Often the number is fairly small, such as five or so.) Then, estimate the parameters of interest using each imputed data set, and then use an averaging to obtain a final parameter estimate and sampling error.

Briefly, let $\mathbf{W}_{mis}$ denote the matrix of missing data and $\mathbf{W}_{obs}$ the matrix of observations. Assume that MAR holds. Then multiple imputation is justified as a way to estimate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$, the posterier mean of $\boldsymbol{\theta}$ given $\mathbf{W}_{obs}$. But by iterated expectations,

$$E(\boldsymbol{\theta}|\mathbf{W}_{obs}) = E[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}]. \tag{3.9}$$

Now, if we can obtain estimates $\hat{\boldsymbol{\theta}}_d = E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis}^{(d)})$ for imputed data set *d*, then we can approximate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$ as

$$\bar{\boldsymbol{\theta}} = D^{-1} \sum_{d=1}^{D} \hat{\boldsymbol{\theta}}_d, \tag{3.10}$$

which is just the average of the parameter estimates across the imputed samples.

Further, we can obtain a "sampling" variance by estimating $Var(\theta|\mathbf{W}_{obs})$ using

$$Var(\theta|\mathbf{W}_{obs}) = E[Var(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}] + Var[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}], \tag{3.11}$$

which suggests

$$\begin{aligned} \widehat{Var}(\theta|\mathbf{W}_{obs}) &= D^{-1}\sum_{d=1}^{D}\hat{\mathbf{V}}_d + (D-1)^{-1}\sum_{d=1}^{D}(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})' \\ &\equiv \bar{\mathbf{V}} + \mathbf{B}, \end{aligned} \tag{3.12}$$

where $\bar{\mathbf{V}}$ is the average of the variance estimates across imputed samples and $\mathbf{B}$ is the between-imputation variance. For small a small number of imputations, a correction is usually made, namely, $\bar{\mathbf{V}} + (1 + D)^{-1}\mathbf{B}$. Therefore, assume that one trusts the MAR assumption, and the underlying distributions used to draw the imputed values, inference with multiple imputations is fairly straightforward. Because D need not be very large, estimation of nonlinear models using multiple imputations is not computationally prohibitive (once one has the imputed data, of course).

Like weighting methods, imputation methods have an important shortcoming when applied to estimation of models with missing conditioning variables. Suppose again that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, we are interested in some feature of the conditional distribution $D(y|\mathbf{x})$, data are missing on y and $\mathbf{x}_2$ – say, for the same units – and selection is a function of $\mathbf{x}_2$. Then, as we discussed in Section 1, MLE using the complete cases is consistent, asymptotically normal, and inference is standard. What about imputation methods? Because they generally rely on MAR, they would require that $D(s|y, \mathbf{x}_1, \mathbf{x}_2) = D(s|\mathbf{x}_1)$. Because this is false in this example, MI cannot be expected to produce convincing imputations.

Imputation for the monotone case discussed above is relatively straightforward under MAR, and MAR is at least plausible. Because the conditional densities are identified, imputation can proceed sequentially: given w_{i1} and $\hat{\theta}$, missing values on w_{i2} can be imputed by drawing from $f_2(\cdot|w_{i1}, \hat{\theta})$. Then, w_{i3} can be imputed by drawing from $f(\cdot|\hat{w}_{i2}, w_{i1}, \hat{\theta})$, where $\hat{w}_{i2}$ may or may not be imputed. And so on.

4. Heckman-Type Selection Corrections

4.1. Corrections with Instrumental Variables

Here we briefly cover the well-known Heckman selection correction with endogenous explanatory variables in a linear model. The model is

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{4.1}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2 \tag{4.2}$$

$$y_3 = 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0]. \tag{4.3}$$

where $\mathbf{z}$ is $1 \times L$ with first element unity (and also in $\mathbf{z}_1$). As usually, $L_1 < L$ for identification. The key point to be made here is, depending on how the Heckman correction is carried out in this case, (4.2) can just be a linear projection – in which case the nature of y_2 is unrestricted – or, effectively, v_2 must be normal and independent of $\mathbf{z}$. Intuitively, we need two elements in $\mathbf{z}$ not also in z_1: loosely, one to induce exogenous variation in y_2 and the other to induce exogenous variation in selection. If we assume (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) observed when $y_3 = 1$; (b) $E(u_1|\mathbf{z}, v_3) = \gamma_1 v_3$; (c) $v_3|\mathbf{z}$ ~Normal$(0, 1)$; (d) $E(\mathbf{z}'v_2) = \mathbf{0}$ and $\boldsymbol{\delta}_{22} \neq \mathbf{0}$, then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{4.4}$$

where $e_1 = u_1 - g(\mathbf{z}, y_3) = u_1 - E(u_1|\mathbf{z}, y_3)$. So, selection is exogenous in (4.4) because $E(e_1|\mathbf{z}, y_3) = 0$. Because y_2 is not exogenous, we estimate (4.4) by IV, using the selected sample, where the instruments are $(\mathbf{z}, \lambda(\mathbf{z}\boldsymbol{\delta}_3))$ because $g(\mathbf{z}, 1) = \lambda(\mathbf{z}\boldsymbol{\delta}_3)$. So, the two-step estimator is

(i) Probit of y_3 on $\mathbf{z}$ to (using all observations) to get $\hat{\lambda}_{i3} \equiv \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$

(ii) IV (2SLS if overidentifying restrictions) of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{\lambda}_{i3}$ using the selected sample and instruments $(\mathbf{z}_i, \hat{\lambda}_{i3})$.

If y_2 is always observed, it is tempting to obtain the fitted values $\hat{y}_{i2}$ from the reduced form y_{i2} on $\mathbf{z}_i$, and then use OLS of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}, \hat{\lambda}_{i3}$ in the second stage. But this effectively puts $\alpha_1 v_2$ in the error term, so we would need $u_1 + \alpha_2 v_2$ to be normal (or something similar); it would not be consistent for discrete y_2, for example. Implicitly, the reduced form estimated by the proper two-step procedure is, on the selected sample, $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + \eta_2\lambda(\mathbf{z}\boldsymbol{\delta}_3) + r_3$. But this is just a linear projection; generally, the rank condition on the selected sample should hold if $\mathbf{z}$ causes sufficient variation in y_2 and y_3 in the population.

This example raises another point: even if y_2 is exogenous in the full population, one should generally treat it as endogenous in the selected subsample. Why? Because y_2 cannot be included in the first-stage probit if it is not always observed, so consistency of the Heckman procedure would require $P(y_3 = 1|\mathbf{z}_1, y_2) = P(y_3 = 1|\mathbf{z}_1)$, a tenuous assumption. Unless we have an instrument for y_2, simply treating it as exogenous in the second stage after omitting it

from the first is tantamount to imposing an exclusion restriction on a reduced form.

In addition to the linear model, with or without endogenous variables, Heckman-type corrections are available for a limited set of nonlinear models. Terza (1998) contains the approach for exponential functions with exogenous explanatory variables, where the selection equation follows a probit; see also Wooldridge (2002, Chapter 19). A selection correction is also fairly easy to implement in probit models, too; see Wooldridge (2002, Chapter 17). As in trying to account for endogenous explanatory variables in such models, merely inserting an estimated inverse Mills ratio inside, say, an exponential model, or probit model, or Tobit model. One can always base a test on a variable-addition approaches, but they cannot be shown to solve the selection problem.

A very similar issue arises when using Heckman's method to correct for attrition in panel data (when selection on observables does not hold). With attrition as an absorbing state, it is common to estimate models in first differences to remove additive heterogeneity, say

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}, t = 2, \ldots, T. \tag{4.5}$$

We assume $s_{it} = 1 \Rightarrow s_{ir} = 1, r < t$. Let w_{it} be a set of variables that we always observe when $s_{i,t-1} = 1$ such that w_{it} is a good predictor of selection – in a sense soon to be made precise. We model the selection in time period t conditional on $s_{i,t=1} = 1$ as

$$s_{it} = 1[w_{it}\delta_t + v_{it} > 0] \tag{4.6}$$

$$v_{it}|(w_{it}, s_{i,t-1} = 1) \sim Normal(0,1), t = 2, 3, \ldots, T. \tag{4.7}$$

Since attrition is an absorbing state, $s_{it-1} = 0$ implies $s_{it} = 0$. This leads to a probit model for s_{it} conditional on $s_{i,t-1} = 1$:

$$P(s_{it} = 1|w_{it}, s_{i,t-1} = 1) = \Phi(w_{it}\delta_t), t = 2, \ldots, T. \tag{4.8}$$

Naturally, we need to estimate δ_t, which we do as a sequence of probits. For $t = 2$, we use the entire sample to estimate a probit for still being in the sample in the second period. For $t = 3$, we estimate a probit for those units still in the sample as of $t = 2$. And so on. When we reach $t = T$, we have the smallest group of observations because we only use units still in the sample as of $T - 1$. Where might the w_{it} come from? Since they have to be observed at time t for the entire subgroup with $s_{i,t-1} = 1$, w_{it} generally cannot contain variables dated at time t (unless some information is known at time t on people who attrit at time t). When the x_{it} are strictly exogenous, we can always include in w_{it} elements of $(x_{i,t-1}, x_{i,t-2}, \ldots, x_{i1})$. Note that the potential dimension of w_{it} grows as we move ahead through time. Unfortunately, $y_{i,t-1}$ cannot

be in w_{it} because $y_{i,t-1}$ is necessarily correlated with Δu_{it}. But, if we assume that

$$E(u_{it}|x_i, y_{i,t-1}, \ldots, y_{i1}, c_i) = 0, t = 2, \ldots, T, | \tag{4.9}$$

then elements from $(y_{i,t-2}, y_{i,t-3}, \ldots, y_{i1})$ can be in w_{it}. If we start with a model where x_{it} is strictly exogenous, as in standard panel data models, assumption (4.9) is very strong because in such models u_{it} tends to be serially correlated, and therefore correlated with lagged y_{it} in general. Still, since we are allowing for c_i, it might be that the errors $\{u_{it}\}$ are serially uncorrelated.

In what sense do we need the w_{it} to be good predictors of attrition? A sufficient condition is, given $s_{i,t-1} = 1$,

$$(\Delta u_{it}, v_{it}) \text{ is independent of } (\Delta x_{it}, w_{it}). \tag{4.10}$$

Now, Δu_{it} is independent of $(\Delta x_{it}, w_{it})$ holds if w_{it} contains only lags of x_{it} because we assume x_{it} is strictly exogenous. Unfortunately, v_{it} is independent of $(\Delta x_{it}, w_{it})$ can be very restrictive because Δx_{it} cannot be included in w_{it} in interesting cases (because x_{it} is not observed for everyone with $s_{i,t-1} = 1$). Therefore, when we apply a sequential Heckman method, we must omit at least some of the explanatory variables in the first-stage probits. If attrition is largely determined by changes in the covariates (which we do not see for everyone), using pooled OLS on the FD will be consistenty, whereas the Heckman correction would actually cause inconsistency.

As in the cross section case, we can "solve" this problem by using instrumental variables for any elements of Δx_{it} not observed at time *t*. Assume sequential exogeneity, that is

$$E(u_{it}|x_{it}, x_{i,t-1}, \ldots, x_{i1}, c_i) = 0, t = 1, \ldots, T. \tag{4.11}$$

(Recall that this condition does allow for lagged dependent variables in x_{it}). We now replace (4.10) with

$$(\Delta u_{it}, v_{it}) \text{ is independent of } (z_{it}, w_{it}) \tag{4.12}$$

conditional on $s_{i,t-1} = 1$. Choosing z_{it} to be a subset of w_{it} is attractive, because then (4.12) $E(\Delta u_{it}|z_{it}, w_{it}, v_{it}, s_{i,t-1} = 1) = E(\Delta u_{it}|w_{it}, v_{it}, s_{i,t-1} = 1)$, in which case (4.12) holds if $(\Delta u_{it}, v_{it})$ is independent of w_{it} given $s_{i,t-1} = 1$. Then, after a sequence of probits (where, in each time period, we use observations on all units available in the previous time periods), we can apply pooled 2SLS, say, on the selected sample, to the equation

$$\Delta y_{it} = \Delta x_{it}\beta + \rho_2 d2_t\hat{\lambda}_{it} + \rho_3 d3_t\hat{\lambda}_{it} + \ldots + \rho_T dT_t\hat{\lambda}_{it} + error_{it}. \tag{4.13}$$

with instruments $(z_{it}, d2_t\hat{\lambda}_{it}, d3_t\hat{\lambda}_{it}, \ldots, dT_t\hat{\lambda}_{it})$. Because $\hat{\lambda}_{it}$ depends on w_{it}, it is critical to have an element in w_{it} moving around selection separately from its correlation with Δx_{it}.

One can also test and correct for selection bias for any pattern of missing data on the response variable (or, generally, on endogenous explanatory variables). The key is that data are always observed on variables taken to be strictly exogenous, conditional on unobserved heterogeneity. Semykina and Wooldridge (2006) work through the details for the model

$$\begin{aligned} y_{it} &= x_{it}\beta + c_i + u_{it} \\ E(u_{it}|z_i, c_i) &= 0, \end{aligned} \tag{4.14}$$

where $z_i = (z_{i1}, \ldots, z_{iT})$, so that some elements of x_{it} are possibly endogenous, but the instruments, z_{it}, are strictly exogenous but allowed to be correlated with c_i. A simple test for correlation between s_{it} and the idiosyncratic error – which, recall from Section 1, causes inconsistency in the FE-IV estimator, is available using Heckman's approach. In the first stage, estimate a pooled probit, or separate probit models, on z_{it} and, say, the time averages, $\bar{z}_i$. Obtain estimated inverse Mills ratios. Then, estimate the equation

$$y_{it} = x_{it}\beta + \rho\hat{\lambda}_{it} + c_i + error_{it} \tag{4.15}$$

by FEIV, and use a standard (but robust) test of $\rho = 0$. This allows for endogeneity of x_{it} under H_0, and so is a pure selection bias test. Or, the $\hat{\lambda}_{it}$ can be interacted with year dummies. The usefulness of this test is that it maintains only $E(E(u_{it}|z_i, s_i, c_i) = 0$ under H_0. Unfortunately, as a correction procedure, it generally does not lead to consistent estimators. (See Semykina and Wooldridge (2006).) As it turns out, a procedure that does produce consistent estimates under certain assumptions is just to add the time-average of the instruments, $\bar{z}_i$, to (4.15) and use pooled IV, where $\bar{z}_i$ and $\hat{\lambda}_{it}$ act as their own instruments.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 13, Wednesday, Aug 1st, 2.00-3.00pm

Weak Instruments and Many Instruments

1. Introduction

In recent years a literature has emerged that has raised concerns with the quality of inferences based on conventional methods such as Two Stage Least Squares (TSLS) and Limited Information Maximum Likelihood (LIML) in instrumental variables settings when the instrument(s) is/are only weakly correlated with the endogenous regressor(s). Although earlier work had already established the poor quality of conventional normal approximations with weak or irrelevant instruments, the recent literature has been motivated by empirical work where *ex post* conventional large sample approximations were found to be misleading. The recent literature has aimed at developing better estimators and more reliable methods for inference.

There are two aspects of the problem. In the just-identified case (with the number of instruments equal to the number of endogenous regressors), or with low degrees of over-identification, the focus has largely been on the construction of confidence intervals that have good coverage properties even if the instruments are weak. Even with very weak, or completely irrelevant, instruments, conventional methods are rarely substantively misleading, unless the degree of endogeneity is higher than one typically encounters in studies using cross-section data. Conventional TSLS or LIML confidence intervals tend to be wide when the instrument is very weak, even if those intervals do not have the correct nominal coverage for all parts of the parameter space. In this case better estimators are generally not available. Improved methods for confidence intervals based on inverting test statistics have been developed although these do not have the simple form of an estimate plus or minus a constant times a standard error.

The second case of interest is that with a high degree of over-identification. These settings often arise by interacting a set of basic instruments with exogenous covariates in order to

improve precision. If there are many (weak) instruments, standard estimators can be severely biased, and conventional methods for inference can be misleading. In particular TSLS has been found to have very poor properties in these settings. Bootstrapping does not solve these problems. LIML is generally much better, although conventional LIML standard errors are too small. A simple to implement proportional adjustment to the LIML standard errors based on the Bekker many-instrument asymptotics or the Chamberlain-Imbens random coefficients argument appears to lead to substantial improvements in coverage rates.

2. Motivation

Much of the recent literature is motivated by a study by Angrist and Krueger (1991, AK). Subsequently Bound, Jaeger and Baker (1996, BJB) showed that for some specifications AK employed normal approximations were not appropriate despite very large sample sizes (over 300,000 observations).

2.1 The Angrist-Krueger Study

AK were interested in estimating the returns to years of education. Their basic specification is:

$$Y_i = \alpha + \beta \cdot E_i + \varepsilon_i,$$

where Y_i is log (yearly) earnings and E_i is years of education. Their concern, following a long literature in economics, e.g., Griliches, (1977), Card (2001), is that years of schooling may be endogenous, with pre-schooling levels of ability affecting both schooling choices and earnings given education levels. In an ingenuous attemp to address the endogeneity problem AK exploit variation in schooling levels that arise from differential impacts of compulsory schooling laws. School districts typically require a student to have turned six by January 1st of the year the student enters school. Since individuals are required to stay in school till they turn sixteen, individual born in the first quarter have lower required minimum schooling levels than individuals born in the last quarter. The cutoff dates and minimum school dropout age differ a little bit by state and over time, so the full picture is more

complicated but the basic point is that the compulsory schooling laws generate variation in schooling levels by quarter of birth that AK exploit.

One can argue that a more natural analysis of such data would be as a Regression Discontinuity (RD) design, where we focus on comparisons of individuals born close to the cutoff date. We will discuss such designs in a later lecture. However, in the census only quarter of birth is observed, not the actual date, so there is in fact little that can be done with the RD approach beyond what AK do. In addition, there are substantive arguments why quarter of birth need not be a valid instrument (e.g., seasonal patterns in births, or differential impacts of education by age at entering school). AK discuss many of the potential concerns. See also Bound, Jaeger and Baker (1996). We do not discuss these concerns here further.

Table 1 shows average years of education and average log earnings for individual born in the first and fourth quarter, using the 1990 census. This is a subset of the AK data.

Table 1: Summary Statistics Subset of AK Data

Variable	1st Quarter	4th Quarter	difference
Year of Education	12.688	12.840	0.151
Log Earnings	5.892	5.905	0.014
ratio			0.089

The sample size is 162,487. The last column gives the difference between the averages by quarter, and the last row the ratio of the difference in averages. The last number is the Wald estimate of the returns to education based on these data:

$$\hat{\beta} = \frac{\overline{Y}_4 - \overline{Y}_1}{\overline{E}_4 - \overline{E}_1} = 0.0893 \quad (0.0105),$$

where $\overline{Y}_t$ and $\overline{E}_t$ are the average level of log earnings and years of education for individuals

born in the t-th quarter. This is also equal to the Two-Stage-Least-Squares (TSLS) and Limited-Information-Maximum-Likelihood (LIML) estimates because there is only a single instrument and a single endogenous regressor. The standard error here is based on the delta method and asymptotic joint normality of the numerator and denominator.

AK also present estimates based on additional instruments. They take the basic instrument and interact it with 50 state and 9 year of birth dummies. Here we take this a bit further, and following Chamberlain and Imbens (2004) we interact the single binary instrument with state times year of birth dummies to get 500 instruments. Also including the state times year of birth dummies as exogenous covariates leads to the following model:

$$Y_i = X_i'\beta + \varepsilon_i, \qquad \mathbb{E}[Z_i \cdot \varepsilon_i] = 0,$$

where X_i is the 501-dimensional vector with the 500 state/year dummies and years of education, and Z_i is the vector with 500 state/year dummies and the 500 state/year dummies multiplying the indicator for the fourth quarter of birth. Let $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ be the $N \times 1$ vector of log earnings, the $N \times 501$ matrix with regressors, and the $N \times 1000$ matrix of instruments. The TSLS estimator for β is then

$$\hat{\beta}_{\text{TSLS}} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}\right).$$

For these data this leads to

$$\hat{\beta}_{\text{TSLS}} = 0.073 \quad (0.008).$$

The LIML estimator is based on maximization of the log likelihood function

$$L(\beta, \pi, \Omega) = \sum_{i=1}^{N}\left(-\frac{1}{2}\ln|\Omega| - \frac{1}{2}\begin{pmatrix} Y_i - X_i'\beta \\ E_i - Z_i'\pi \end{pmatrix}'\Omega^{-1}\begin{pmatrix} Y_i - X_i'\beta \\ E_i - Z_i'\pi \end{pmatrix}\right),$$

For this subset of the AK data we find, for the coefficient on years of education,

$$\hat{\beta}_{\text{LIML}} = 0.095 \quad (0.017).$$

In large samples the LIML and TSLS are equivalent under homoskedasticity.

2.2 The Bound-Jaeger-Baker critique

BJB found that are potential problems with the AK results. They suggested that despite the large samples used by AK large sample normal approximations may be very poor. The reason is that the instruments are only very weakly correlated with the endogenous regressor. The most striking evidence for this is based on the following calculations, that are based on a suggestion by Alan Krueger. Take the AK data and re-calculate their estimates after replacing the actual quarter of birth dummies by random indicators with the same marginal distribution. In principle this means that the standard (gaussian) large sample approximations for TSLS and LIML are invalid since they rely on non-zero correlations between the instruments and the endogenous regressor. Doing these calculations once for the single and 500 instrument case, for both TSLS and LIML, leads to the results in Table 2

Table 2: Real and Random QOB Estimates

	Single Instrument		500 Instruments			
			TSLS		LIML	
Real QOB	0.089	(0.011)	0.073	(0.008)	0.095	(0.017)
Random QOB	-1.958	(18.116)	0.059	(0.085)	-0.330	(0.1001)

With the single instrument the results are not so disconcertening. Although the confidence interval is obviously not valid, it is wide, and few researchers would be misled by the results. With many instruments the results are much more troubling. Although the instrument contains no information, the results suggest that the instruments can be used to infer precisely what the returns to education are. These results have provided the motivation for the recent weak instrument literature. Note that there is an earlier literature, e.g., Phillips (1984)

Rothenberg (1984), but it is the BJB findings that got the attention of researchers doing empirical work.

2.3 Simulations with Weak Instruments and Varying Degrees of Endogeneity

Here we provide slightly more systematic simulation evidence of the weak instrument problems in the AK setting. We create 10,000 artificial data sets, all of size 160,000, designed to mimic the key features of the AK data. In each of these data sets half the units have quarter of birth (denoted by Q_i) equal to 0 and 1 respectively. Then we draw the two reduced form residuals ν_i and η_i from a joint normal distribution

$$\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.446 & \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} \\ \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} & 10.071 \end{pmatrix} \right).$$

The variances of the reduced form errors mimic those in the AK data. The correlation between the reduced form residuals in the AK data is 0.318. The implied OLS coefficient is $\rho \cdot \sqrt{0.446}/\sqrt{10.071}$. Then years of education is equal to

$$E_i = 12.688 + 0.151 \cdot Q_i + \eta_i,$$

and log earnings is equal to

$$Y_i = 5.892 + 0.014 \cdot Q_i + \nu_i.$$

Now we calculate the IV estimator and its standard error, using either the actual qob variable or a random qob variable as the instrument. We are interested in the size of tests of the null that coefficient on years of education is equal to $0.089 = 0.014/0.151$. We base the test on the t-statistic. Thus we reject the null if the ratio of the point estimate minus 0.089 and the standard error is greater than 1.96 in absolute value. We repeat this for 12 different values of the reduced form error correlation. In Table 3 we report the proportion of rejections and the median and 0.10 quantile of the width of the estimated 95% confidence intervals.

Table 3: Coverage Rates of Conv. TSLS CI by Degree of Endogeneity

ρ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
implied OLS	0.00	0.02	0.04	0.06	0.08	0.11	0.13	0.15	0.17	0.19	0.20	0.21
Real QOB	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.95
Med Width 95% CI	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.06	0.05	0.05	0.05
0.10 quant Width	0.08	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.05	0.04	0.04	0.04
Random QOB	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.82	0.53
Med Width 95% CI	1.82	1.81	1.78	1.73	1.66	1.57	1.45	1.30	1.09	0.79	0.57	0.26
0.10 quant Width	0.55	0.55	0.5403	0.53	0.51	0.48	0.42	0.40	0.33	0.24	0.17	0.08

In this example, unless the reduced form correlations are very high, e.g., at least 0.95, with irrelevant the conventional confidence intervals are wide and have good coverage. The amount of endogeneity that would be required for the conventional confidence intervals to be misleading is higher than one typically encounters in cross-section settings. It is likely that these results extend to cases with a low degree of over-identification, using either TSLS, or preferably LIML. Put differently, although formally conventional confidence intervals are not valid uniformly over the parameter space (e.g., Dufour, 1997), there are no examples we are aware of where they have substantively misleading in just-identified examples. This in contrast to the case with many weak instruments where especially TSLS can be misleading in empirically relevant settings.

3. Weak Instruments

Here we discuss the weak instrument problem in the case of a single instrument, a single endogenous regressor, and no additional exogenous regressors beyond the intercept. More generally the qualitative features of these results by and large apply to the case with a few

weak instruments. We consider the model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i,$$

$$X_i = \pi_0 + \pi_1 \cdot Z_i + \eta_i,$$

with $(\varepsilon_i, \eta_i) \perp\!\!\!\perp Z_i$, and jointly normal with covariance matrix Σ. (The normality is mainly for some of the exact results, and it does not play an important role.) The reduced form for the first equation is

$$Y_i = \alpha_0 + \alpha_1 \cdot Z_i + \nu_i,$$

where the parameter of interest is $\beta_1 = \alpha_1/\pi_1$. Let

$$\Omega = \mathbb{E}\left[\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix}' \right], \qquad \text{and} \quad \Sigma = \mathbb{E}\left[\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix}' \right],$$

be the covariance matrix of the reduced form and stuctural disturbances respectively. Many of the formal results in the literature are for the case of known Ω, and normal disturbances. This is largely innocuous, as Ω can be precisely estimated in typical data sets. Note that this it not the same as assuming that Σ is known, which is not innocuous since it depends on Ω and β, and cannot be precisely estimated in settings with weak instruments

$$\Sigma = \begin{pmatrix} \Omega_{11} - 2\beta\Omega_{12} + \beta^2\Omega_{22} & \Omega_{12} - \beta\Omega_{22} \\ \Omega_{12} - \beta\Omega_{22} & \Omega_{22} \end{pmatrix}.$$

The standard estimator for β_1 is

$$\hat{\beta}_1^{\text{IV}} = \frac{\frac{1}{N}\sum_{i=1}^N \left(Y_i - \overline{Y}\right)\left(Z_i - \overline{Z}\right)}{\frac{1}{N}\sum_{i=1}^N \left(X_i - \overline{X}\right)\left(Z_i - \overline{Z}\right)},$$

where $\overline{Y} = \sum_i Y_i/N$, and similarly for $\overline{X}$ and $\overline{Z}$.

A simple interpretation of the weak instrument is that with the concentration parameter

$$\lambda = \pi_1^2 \cdot \sum_{i=1}^{N} \left(Z_i - \overline{Z}\right)^2 \Big/ \sigma_\eta^2.$$

close to zero, both the covariance in the numerator and the covariance in the denominator are close to zero. In reasonably large samples both are well approximated by normal distributions:

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(Z_i - \overline{Z}\right) - \text{Cov}(Y_i, Z_i)\right) \approx \mathcal{N}\left(0, V(Y_i \cdot Z_i)\right),$$

and

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Z_i - \overline{Z}\right) - \text{Cov}(X_i, Z_I)\right) \approx \mathcal{N}\left(0, V(X_i \cdot Z_i)\right).$$

These two normal approximations tend to be accurate in applications with reasonable sample sizes, irrespective of the population values of the covariances. If $\pi_1 \neq 0$, as the sample size gets large, then the ratio will eventually be well approximated by a normal distribution as well. However, if $\text{Cov}(X_i, Z_i) \approx 0$, the ratio may be better approximated by a Cauchy distribution, as the ratio of two normals centered close to zero.

The weak instrument literature is concerned with inference for β_1 when the concentration parameter λ is too close to zero for the normal approximation to the ratio to be accurate.

Staiger and Stock (1997, SS) formalize the problem by investigating the distribution of the standard IV estimator under an alternative asymptotic approximation. The standard asymptotics (strong instrument asymptotics in the SS terminology) is based on fixed parameters and the sample size getting large. In their alternative asymptotic sequence SS model π_1 as a function of the sample size, $\pi_{1N} = c/\sqrt{N}$, so that the concentration parameter converges to a constant:

$$\lambda \longrightarrow c^2 \cdot V(Z_i).$$

SS then compare coverage properties of various confidence intervals under this (weak instrument) asymptotic sequence.

The importance of the SS approach is not in the specific sequence. The concern is more that if a particular confidence interval does not have the appropriate coverage asymptotically under the SS asymptotics, then there are values of the (nuisance) parameters in a potentially important part of the parameter space (namely around $\pi_i = 0$) such that the actual coverage is substantially away from the nominal coverage for any sample size. More recently the issue has therefore been reformulated as requiring confidence intervals to have asymptotically the correct coverage probabilities uniformly in the parameter space. See for a discussion from this perspective Mikusheva (2007). For estimation this perspective is not helpful: there cannot be estimators that are consistent for β^* uniformly in the parameter space since if $\pi_1 = 0$, there are no consistent estimators for β_1. However, for testing there are generally confidence intervals that are uniformly valid, but they are not of the conventional form, that is, a point estimate plus or minus a constant times a standard error.

3.1 Tests and Confidence Intervals in the Just-identified Case

Let the instrument $\tilde{Z}_i = Z_i - \overline{Z}$ be measured in deviations from its mean. Then define the statistic

$$S(\beta_1) = \frac{1}{N}\sum_{i=1}^{N} \tilde{Z}_i \cdot (Y_i - \beta_1 \cdot X_i)\,.$$

Then, under the null hypothesis that $\beta_1 = \beta_1^*$, and conditional on the instruments, the statistic $\sqrt{N} \cdot S(\beta_1^*)$ has an exact normal distribution

$$\sqrt{N} \cdot S(\beta_1^*) \sim \mathcal{N}\left(0, \sum_{i=1}^{N} \tilde{Z}_i^2 \cdot \sigma_\varepsilon^2\right).$$

Importantly, this result does not depend on the strength of the instrument. Anderson and Rubin (1949, AR) propose basing tests for the null hypothesis

$$H_0: \quad \beta_1 = \beta_1^0, \qquad \text{against the alternative hypothesis} \quad H_a: \quad \beta_1 \neq \beta_1^0,$$

on this idea, through the statistic

$$\text{AR}\left(\beta_1^0\right) = \frac{N \cdot S(\beta_1^0)^2}{\sum_{i=1}^N \tilde{Z}_i^2} \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix} \right)^{-1}.$$

This statistic has an exact chi-squared distribution with degrees of freedom equal to one. In practice, of course, one does not know the reduced form covariance matrix Ω, but substituting an estimated version of this matrix based on the average of the estimated reduced form residuals does not affect the large sample properties of the test.

A confidence interval can be based on this test statistic by inverting it. For example, for a 95% confidence interval for β_1, we would get

$$\text{CI}_{0.95}^{\beta_1} = \left\{ \beta_1 \,|\text{AR}(\beta_1) \leq 3.84 \right\}.$$

Note that this AR confidence interval cannot be empty, because at the standard IV estimator $\hat{\beta}_1^{\text{IV}}$ we have $\text{AR}(\hat{\beta}_1^{\text{IV}}) = 0$, and thus $\hat{\beta}_1^{\text{IV}}$ is always in the confidence interval. The confidence interval can be equal to the entire real line, if the correlation between the endogenous regressor and the instrument is close to zero. This is not surprising: in order to be valid even if $\pi_1 = 0$, the confidence interval must include all real values with probability 0.95.

3.3 Tests and Confidence Intervals in the Over-identified Case

The second case of interest is that with a single endogenous regressor and multiple instruments. We deal separately with the case where there are many (similar) instrument, so this really concerns the case where the instruments are qualitatively different. Let the number of instrumens be equal to K, so that the reduced form is

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i,$$

with Z_i a k-dimensional column vector. There is still only a single endogenous regressor, and no exogenous regressors beyond the intercept. All the results generalize to the case with additional exogenous covariates at the expense of additional notatio. The AR approach can

be extended easily to this over-identified case, because the statistic $\sqrt{N} \cdot S(\beta_1^*)$ still has a normal distribution, but now a multivariate normal distribution. Hence one can base tests on the AR statistic

$$\text{AR}\left(\beta_1^0\right) = N \cdot S\left(\beta_1^0\right)' \left(\sum_{i=1}^{N} \tilde{Z}_i \cdot \tilde{Z}_i'\right)^{-1} S\left(\beta_1^0\right) \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)^{-1}.$$

Under the same conditions as before this has an exact chi-squared distribution with degrees of freedom equal to the number of instruments, k. A practical problem arises if we wish to construct confidence intervals based on this statistic. Suppose we construct a confidence interval, analogously to the just-identified case, as

$$\text{CI}_{0.95}^{\beta_1} = \left\{\beta_1 \,\middle|\, \text{AR}(\beta_1) \leq \mathcal{X}_{0.95}^2(K)\right\},$$

where $\mathcal{X}_{0.95}^2(k)$ is the 0.95 quantile of the chi-squared distribution with degrees of freedom equal to k. The problem is that this confidence interval can be empty. The interpretation is that the test does not only test whether $\beta_1 = \beta_1^0$, but also tests whether the instruments are valid. However, one generally may not want to combine those hypotheses.

Kleibergen (2002) modifies the AR statistic and confidence interval construction. Instead of the statistic $S(\beta_1)$, he considers a statistic that looks at the correlation between a particular linear combination of the instruments (namely the estimated endogenous regressor) and the residual:

$$\tilde{S}\left(\beta_1^0\right) = \frac{1}{N} \sum_{i=1}^{N} \left(\tilde{Z}_i' \hat{\pi}_1(\beta_1^0)\right) \cdot \left(Y_i - \beta_1^0 \cdot X_i\right),$$

where $\hat{\pi}$ is the maximum likelihood estimator for π_1 under the restriction $\beta_1 = \beta_1^0$. The test is then based on the statistic

$$K\left(\beta_1^0\right) = \frac{N \cdot S(\beta_1^0)^2}{\sum_{i=1}^{N} \tilde{Z}_i^2} \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)^{-1}.$$

This statistic has no longer an exact chi-squared distribution, but in large samples it still has an approximate chi-square distribution with degrees of freedom equal to one. Hence the test is straightforward to implement using standard methods.

Moreira (2003) proposes a method for adjusting the critical values that applies to a number of tests, including the Kleibergen test. His idea is to focus on *similar* tests, test that have the same rejection probability for all values of the nuisance parameter. The nuisance parameter is here the vector of reduced form coefficients π, since we assume the residual covariance matrix is known. The way to adjust the critical values is to consider the distribution of a statistic such as the Kleibergen statistic conditional on a complete sufficient statistic for the nuisance parameter. In this setting a complete sufficient statistic is readily available in the form of the maximum likelihood estimator under the null, $\hat{\pi}_1(\beta_1^0)$. Moreira's preferred test is based on the likelihood ratio. Let

$$LR\left(\beta_1^0\right) = 2 \cdot \left(L\left(\hat{\beta}_1, \hat{\pi}\right) - L\left(\beta_1^0, \hat{\pi}(\beta_1^0)\right)\right),$$

be the likelihood ratio. Then let $c_{LR}(p, 0.95)$, be the 0.95 quantile of the distribution of $LR(\beta_1^0)$ under the null hypothesis, conditional on $\hat{\pi}(\beta_1^0) = p$. The proposed test is to reject the null hypothesis at the 5% level if

$$LR\left(\beta_1^0\right) > c_{LR}(\hat{\pi}(\beta_1^0), 0.95),$$

where conventional test would use critical values from a chi-squared distribution with a single degree of freedom. This test can then be converted to construct a 95% confidence intervals. Calculation of the (large sample) critical values is simplified by the fact that they only depend on the number of instruments k, and a scaled version of the $\hat{\pi}(\beta_1^0)$. Tabulations of these critical values are in Moreira (2003) and have been programmed in STATA (See Moreira's website).

3.4 Conditioning on the First Stage

The AR, Kleibergen and Moreira proposals for confidence intervals are asymptotically

valid irrespective of the strength of the first stage (the value of π_1). However, they are not valid if one first inspects the first stage, and conditional on the strength of that, decides to proceed. Specifically, if in practice one first inspects the first stage, and decide to abandon the project if the first stage F-statistic is less than some fixed value, and otherwise proceed by calculating an AR, Kleibergen or Moreira confidence interval, the large sample coverage probabilities would not necessarily be the nominal ones. In practice researchers do tend to inspect and report the strength of the first stage. This is particularly true in recent instrumental variables literature where researchers argue extensively for the validity of the instrumental variables assumption. This typically involves detailed arguments supporting the alleged mechanism that leads to the correlation between the endogenous regressor and the instruments. For example, Section I in AK (page 981-994) is entirely devoted to discussing the reasons and evidence for the relation between their instruments (quarter of birth) and years of education. In such cases inference conditional on this may be more appropriate.

Chioda and Jansson (2006) propose a clever alternative way to construct a confidence interval that is valid conditional on the strength of the first stage. Their proposed confidence interval is based on inverting a test statistic similar to the AR statistic. It has a non-standard distribution conditional on the strength of the first stage, and they suggest a procedure that involves numerically approximating the critical values. A caveat is that because the first stage F-statistic, or the first stage estimates are not ancillary, conditioning on them involves loss of information, and as a result the Chioda-Jansson confidence intervals are wider than confidence intervals that are not valid conditional on the first stage.

4. Many Weak Instruments

In this section we discuss the case with many weak instruments. The problem is both the bias in the standard estimators, and the misleadingly small standard errors based on conventional procedures, leading to poor coverage rates for standard confidence intervals in many situations. The earlier simulations showed that especially TSLS, and to a much lesser extent LIML, have poor properties in this case. Note first that resampling methods such as bootstrapping do not solve these problems. In fact, if one uses the standard bootstrap with

TSLS in the AK data, one finds that the average of the bootstrap estimates is very close to the TSLS point estimat, and that the bootstrap variance is very close to the TSLS variance.

The literature has taken a number of approaches. Part of the literature has focused on alternative confidence intervals analogues to the single instrument case. In addition a variety of new point estimators have been proposed.

4.1 Bekker Asymptotics

In this setting alternative asymptotic approximations play a bigger role than in the single instrument case. In an important paper Bekker (1995) derives large sample approximations for TSLS and LIML based on sequences where the number of instruments increases proportionally to the sample size. He shows that TSLS is not consistent in that case. LIML is consistent, but the conventional LIML standard errors are not valid. Bekker then provides LIML standard errors that are valid under this asymptotic sequence. Even with relatively small numbers of instruments the differences between the Bekker and conventional asymptotics can be substantial. See also Chao and Swanson (2005) for extensions.

For the simple case with a single endogenous regressor, and no exogenous regressors beyond the intercept, the adjustment to the variance is multiplicative. Thus, one can simply multiply the standard LIML variance by

$$1+\frac{K/N}{1-K/N}\cdot\left(\sum_{i=1}\left(\pi_1'\tilde{Z}_i\right)^2\Big/N\right)^{-1}\cdot\left(\begin{pmatrix}1\\ \beta_1\end{pmatrix}'\Omega^{-1}\begin{pmatrix}1\\ \beta_1\end{pmatrix}\right)^{-1}.$$

Substituting estimated values for the unknown parameters is likely to work fairly well in practice. One can see from this expression why the adjustment can be substantial even if K is small. The second factor can be large if the instruments are weak, and the third factor can be large if the degree of endogeneity is high. If the instruments are strong, then $\sum_{i=1}(\pi_1'\tilde{Z}_i)^2/K$ will diverge, and the adjustment factor will converge to one.

4.2 Random Effects Estimators

Chamberlain and Imbens (2004, CI) propose a random effects quasi maximum likelihood

estimator. They propose modelling the first stage coefficients π_k, for $k = 1, \ldots, K$, in the regression

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i = \pi_0 + \sum_{k=1}^{K} \pi_k \cdot Z_{ik} + \eta_i,$$

(after normalizing the instruments to have mean zero and unit variance,) as independent draws from a normal $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ distribution. (More generally CI allow for the possibility that only some of the first stage coefficients come from this common distribution, to take account of settings where some of the instruments are qualitatively different from the others.) The idea is partly that in most cases with many instruments, as for example in the AK study, the instruments arise from interacting a small set of distinct instruments with other covariates. Hence it may be natural to think of the coefficients on these instruments in the reduced form as exchangeable. This notion is captured by modelling the first stage coefficients as independent draws from the same distribution. In addition, this set up parametrizes the many-weak instrument problem in terms of a few parameters: the concern is that the values of both μ_π and σ_π^2 are close to zero.

Assuming also joint normality for (ε_i, η_i), one can derive the likelihood function

$$\mathcal{L}(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2, \Omega).$$

In contrast to the likelihood function in terms of the original parameters $(\beta_0, \beta_1, \pi_0, \pi_1, \Omega)$, this likelihood function depends on a small set of parameters, and a quadratic approximation to its logarithms is more likely to be accurate.

CI discuss some connections between the REQML estimator and LIML and TSLS in the context of this parametric set up. First they show that in large samples, with a large number of instruments, the TSLS estimator corresponds to the restricted maximum likelihood estimator where the variance of the first stage coefficients is fixed at a large number, or $\sigma_\pi^2 = \infty$:

$$\hat{\beta}_{\text{TSLS}} \approx \arg \max_{\beta_0, \beta_1, \pi_0, \mu_\pi} = L(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2 = \infty, \Omega).$$

From a Bayesian perspective, TSLS corresponds approximately to the posterior mode given a flat prior on all the parameters, and thus puts a large amount of prior mass on values of the parameter space where the instruments are jointly powerful.

In the same setting with a large number of instruments, no exogenous covariates, and a known reduced form covariance matrix, the LIML estimator corresponds approximately to the REQML estimator where we fix $\sigma_\pi^2 \cdot (1 \ \beta_1)'\Omega^{-1}(1 \ \beta_1)'$ at a large number. In the special case where we fix $\mu_\pi = 0$ and the random effects specification applies to all isntruments, CI show that the REQML estimator is identical to LIML. However, like the Bekker asymptotics, the REQML calculations suggests that the standard LIML variance is too small: the variance of the REQML estimator is approximately equal to the standard LIML variance times

$$1 + \sigma_\pi^{-2} \cdot \left(\begin{pmatrix} 1 \\ \beta_1 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} 1 \\ \beta_1 \end{pmatrix} \right)^{-1}.$$

This is similar to the Bekker adjustment if we replace σ_π^2 by $\sum_{i=1}(\pi_1'\tilde{Z}_i)^2(K \cdot N)$ (keeping in mind that the instruments have been normalized to have unit variance). In practice the CI adjustment will be bigger than the Bekker adjustment because the ml estimator for σ_π^2 will take into account noise in the estimates of the $\hat{\pi}$, and so $\hat{\sigma}_\pi^2 < \sum_{i=1}(\hat{\pi}_1'\tilde{Z}_i)^2(K \cdot N)$.

4.3 Choosing Subsets of the Instruments

In an interesting paper Donald and Newey (2001) consider the problem of choosing a subset of an infinite sequence of instruments. They assume the instruments are ordered, so that the choice is the number of instruments to use. Given the set of instruments they consider a variety of estimators including TSLS and LIML. The criterion they focus on is based on an approximation to the expected squared error. This criterion is not feasible because it depends on unknown parameters, but they show that using an estimated version of this leads to approximately the same expected squared error as using the infeasible criterion. Although in its current form not straightforward to implement, this is a very promising approach that can apply to many related problems such as generalized method of moments settings with many moments.

4.4 Other Estimators

Other estimators have also been investigated in the many weak instrument settings. Hansen, Hausman and Newey (2006), and Hausman, Newey and Woutersen (2007) look at Fuller's estimator, which is modification of LIML that has finite moments. Phillips and Hale (1977) (and later Angrist, Imbens and Krueger, 1999) suggest a jackknive estimator. Hahn, Hausman and Kuersteiner (2004) look at jackknife versions of TSLS.

4.5 Flores' simulations

Many simulations exercises have been carried out for evaluating the performance of testing procedures and point estimators. In general it is difficult to assess the evidence of these experiments. They are rarely tied to actual data sets, and so the choices for parameters, distributions, sample sizes, and number of instruments are typically arbitrary.

In one of the more extensive simulation studies Flores-Lagunes (2007) reports results comparing TSLS, LIML, Fuller, Bias corrected versions of TSLS, LIML and Fuller, a Jacknife version of TSLS (Hahn, Hausman and Kuersteiner, 2004), and the REQML estimator, in settings with 100 and 500 observations, and 5 and 30 instruments for the single endogenous variable. He looks at median bias, median absolute error, inter decile range, coverage rates, and He concludes that "our evidence indicates that the random-effects quasi-maximum likelihood estimator outperforms alternative estimators in terms of median point estimates and coverage rates."

References

Anderson, T., and H. Rubin, (1949), "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics* 21, 570-582-.

Andrews, D., M. Moreira, and J. Stock, (2006), "Optimal Two-sided Invariant Similar Tests for Instrumental Variables Regresion," *Econometrica* 74, 715-752-.

Andrews, D., and J. Stock, (2007), "Inference with Weak Instruments," *Advances in Economics and Econometrics*, Vol III, Blundel,, Newey and Persson (eds.), 122-173.

Angrist, J., G. Imbens, and A. Krueger, (1999), "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics,* 14, 57-67.

Angrist, J., and A. Krueger, (1991), "Does Compulsory Schooling Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics* 106, 979-1014.

Bekker, P., (1994), "Alternative Approximations to the Distribution of Instrumental Variables Estimators," *Econometrica* 62, 657-681.

Bound, J., A. Jaeger, and R. Baker, (1996), "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association* 90, 443-450.

Card, D., (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69(5), 1127-1160.

Chamberlain, G., and G. Imbens, (2004), "Random Effects Estimators with Many Instrumental Variables," *Econometrica* 72(1), 295-306.

Chao, J., and N. Swanson, (2005), "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica* 73(5), 1673-1692.

Dufour, J.-M., (1997), "Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica* 65, 1365-1387.

Chioda, L., and M. Janson, (1998), "Optimal Conditional Inference for Instrumental Variables Regression," unpublished manuscript, department of economics, UC Berkeley.

Donald, S., and W. Newey, (2001), "Choosing the Number of Instruments," *Econometrica* 69, 1161-1191.

Flores-Lagunes, A., (2007), "Finite Sample Evidence of IV Estimators Under Weak Instruments," *Journal of Applied Econometrics,* 22, 677-694.

Fuller, W., (1977), "Some Properties of a Modification of the Limited Information Estimator," *Econometrica* 45(), 939-954.

Griliches, Z., (1977), "Estimating the Returns to Schooling – Some Econometric Problems," *Econometrica* 45(1), 1-22.

Hahn, J., and J. Hausman, (2003), "Weak Instruments: Diagnosis and Cures in Empirical Ecnometrics," *American Economic Review, Papers and Proceedings* 93, 118-115.

Hahn, J., J. Hausman, and G. Kuersteiner, (2004), "Estimation with Weak Instruments: Accuracy of Higher Order Bias and MSE Approxmiations," *Econometrics Journal*.

Hansen, C., J. Hausman, and W. Newey, (2006), "Estimation with Many Instrumental Variables," Unpublished Manuscript, Department of Economics, MIT.

Hausman, J., W. Newey, and T. Woutersen, (2006), "IV Estimation with Heteroskedasticity and Many Instruments," Unpublished Manuscript, MIT.

Kleibergen, F., (2002), "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica* 70(5), 1781-1803.

Mikusheva, A., (2007), "Uniform Inferences in Econometrics," Chapter 3, PhD Thesis, Harvard University, Department of Economics.

Moreira, M., (2001), "Tests with Correct Size when Instruments can be Arbitrarily Weak," Unpublished Paper, Department of Economics, Harvard University.

Moreira, M., (2003), "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica* 71(4), 1027-1048.

Philips, P., (1984), "Exact Small Sample Theory in the Simultaneous Equations Model," *Handbook of Econometrics*, (Griliches and Intrilligator, eds), Vol 2, North Holland.

Phillips, G., and C. Hale, (1977), "The Bias of Instrumental Variables Estimators of Simultaneous Equations Systems," *International Economic Review*, 18, 219-228.

Rothenberg, T., (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics," *Handbook of Econometrics*, (Griliches and Intrilligator, eds), Vol 2, Amsterdam, North Holland.

Staiger, D., and J. Stock, (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 68, 1055-1096.

What's New in Econometrics? **NBER**, **Summer 2007**
Lecture 14, **Wednesday**, **August 1st**, **3**:**15**-**4**:**15 pm**

Quantile Methods

These notes review quantile estimation in a variety of situations, including models with endogenous explanatory variables – including endogenous treatment effects – and panel data models with unobserved heterogeneity. Recent work on interpreting quantile estimators when the quantile is misspecified is also covered.

1. Reminders About Means, Medians, and Quantiles

Consider the standard linear model in a population, with intercept α and $K \times 1$ slopes $\boldsymbol{\beta}$:

$$y = \alpha + \mathbf{x}\boldsymbol{\beta} + u. \tag{1.1}$$

Assume $E(u^2) < \infty$, so that the distribution of u is not too spread out. Given a large random sample, when should we expect ordinary least squares, which solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} (y_i - a - \mathbf{x}_i\mathbf{b})^2, \tag{1.2}$$

and least absolute deviations (LAD), which solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} |y_i - a - \mathbf{x}_i\mathbf{b}|, \tag{1.3}$$

to provide similar parameter estimates? There are two important cases. If

$$D(u|\mathbf{x}) \text{ is symmetric about zero} \tag{1.4}$$

then OLS and LAD both consistently estimate α and $\boldsymbol{\beta}$. If

$$u \text{ is independent of } \mathbf{x} \text{ with } E(u) = 0, \tag{1.5}$$

where $E(u) = 0$ is the normalization that identifies α, then OLS and LAD both consistently estimate the slopes, $\boldsymbol{\beta}$. If u has an asymmetric distribution, then $Med(u) \equiv \eta \neq 0$, and $\hat{\alpha}_{LAD}$ converges to $\alpha + \eta$ because $Med(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta} + Med(u|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta} + \eta$. Of course, independence between u and $\mathbf{x}$ rules out heteroskedasticity in $Var(u|\mathbf{x})$.

In many applications, neither (1.4) nor (1.5) is likely to be true. For example, y may be a measure of wealth, in which case the error distribution is probably asymmetric and $Var(u|\mathbf{x})$ not constant. Therefore, it is important to remember that if $D(u|\mathbf{x})$ is asymmetric and changes with $\mathbf{x}$, then we should not expect OLS and LAD to deliver similar estimates of $\boldsymbol{\beta}$, even for "thin-tailed" distributions. In other words, it is important to separate discussions of resiliency to outliers from the different quantities identified by least squares (the conditional mean,

$E(y|\mathbf{x})$) and least absolution deviations (the conditional median, $Med(y|\mathbf{x})$). Of course, it is true that LAD is much more resilient to changes in extreme values because, as a measure of central tendency, the median is much less sensitive than the mean to changes in extreme values. But a significant difference between OLS and LAD should not lead one to somehow prefer LAD. It is possible that $E(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta}$, $Med(y|\mathbf{x})$ is not linear, and therefore LAD does not consistently estimate $\boldsymbol{\beta}$. Generally, if we just use linear models as approximations to underlying nonlinear functions, we should not be surprised if the linear approximation to the conditional mean, and that for the median, can be very different. (Warning: Other so-called "robust" estimators, which are intended to be insensitive to outliers or influential data, usually require symmetry of the error distribution for consistent estimation. Thus, they are not "robust" in the sense of delivering consistency under a wide range of assumptions.)

Sometimes one can use a transformation to ensure conditional symmetry or the independence assumption in (1.5). When $y_i > 0$, the most common transformation is the natural log. Often, the linear model $\log(y) = \alpha + \mathbf{x}\boldsymbol{\beta} + u$ is more likely to satisfy symmetry or independence. Suppose that symmetry about zero holds in the linear model for $\log(y)$. Then, because the median passes through monotonic functions (unlike the expectation), $\text{Med}(y|\mathbf{x}) = \exp(\text{Med}[\log(y)|\mathbf{x}]) = \exp(\alpha + \mathbf{x}\boldsymbol{\beta})$, and so we can easily recover the partial effects on the median of y itself. By contrast, we cannot generally find $\text{E}(y|\mathbf{x}) = \exp(\alpha + \mathbf{x}\boldsymbol{\beta})\text{E}[\exp(u)|\mathbf{x}]$. If, instead, we assume $D(u|\mathbf{x}) = D(u)$, then $\text{Med}(y|\mathbf{x})$ and $\text{E}(y|\mathbf{x})$ are both exponential functions of $\mathbf{x}\boldsymbol{\beta}$, but with different "intercepts" inside the exponential function.

The fact that the median passes through monotonic functions is very handy for applying LAD to a variety of problems, particularly corner solution responses where an outcome has nonnegative support and a mass point at zero. But the expectation operator has useful properties that the median does not: linearity and the law of iterated expectations. To see how these help to identify interesting quantities, suppose we begin with a random coefficient model

$$y_i = a_i + \mathbf{x}_i\mathbf{b}_i, \tag{1.6}$$

where a_i is the heterogeneous intercept and $\mathbf{b}_i$ is a $1 \times K$ matrix of heterogeneous slopes ("random coefficients"). If we assume that $(a_i, \mathbf{b}_i)$ is independent of $\mathbf{x}_i$, then

$$\text{E}(y_i|\mathbf{x}_i) = \text{E}(a_i|\mathbf{x}_i) + \mathbf{x}_i\text{E}(\mathbf{b}_i|\mathbf{x}_i) \equiv \alpha + \mathbf{x}_i\boldsymbol{\beta}, \tag{1.7}$$

where $\alpha = \text{E}(a_i)$ and $\boldsymbol{\beta} = \text{E}(\mathbf{b}_i)$. Because OLS consistently estimates the parameters of a

conditional mean linear in those parameters, OLS consistently estimates the population averaged effects, or average partial effects, $\boldsymbol{\beta}$. Even under independence, there is no way to derive $\text{Med}(y_i|\mathbf{x}_i)$ without imposing more restrictions. In general, LAD of y_i on $1, \mathbf{x}_i$ does not consistently estimate $\boldsymbol{\beta}$ or the medians of the elements of b_{ij}. Are there any reasonable assumptions that imply LAD consistenty estimates something of interest in (1.7)? Yes, although multivariate symmetry is involved. With multivariate distributions there is no unique definition of symmetry. A fairly strong restriction is the notion of a *centrally symmetric* distribution (Serfling (2006)). If $\mathbf{u}_i$ is a vector, then its distribution conditional on $\mathbf{x}_i$ is centrally symmetric if

$$\text{D}(\mathbf{u}_i|\mathbf{x}_i) = \text{D}(-\mathbf{u}_i|\mathbf{x}_i). \tag{1.8}$$

This condition implies that, for any $\mathbf{g}_i$ a function of $\mathbf{x}_i$, $\text{D}(\mathbf{g}_i'\mathbf{u}_i|\mathbf{x}_i)$ has a univariate distribution that is symmetric about zero. Of course, (1.8) implies that $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$.

We can apply this to the random coefficient model as follows. Write $\mathbf{c}_i = (a_i, \mathbf{b}_i)$ with $\boldsymbol{\gamma} = E(\mathbf{c}_i)$, and let $\mathbf{d}_i = \mathbf{c}_i - \boldsymbol{\gamma}$. Then we can write

$$\begin{aligned} y_i &= \alpha + \mathbf{x}_i\boldsymbol{\beta} + (a_i - \alpha) + \mathbf{x}_i(\mathbf{b}_i - \boldsymbol{\beta}) \\ &\equiv \alpha + \mathbf{x}_i\boldsymbol{\beta} + \mathbf{g}_i'\mathbf{d}_i \end{aligned} \tag{1.9}$$

with $\mathbf{g}_i = (1, \mathbf{x}_i)$. Therefore, if $\mathbf{c}_i$ has a centrally symmetric distribution about $\boldsymbol{\gamma}$, then $Med(\mathbf{g}_i'\mathbf{d}_i|\mathbf{x}_i) = 0$, and LAD applied to the usual model $y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + u_i$ consistently estimates α and $\boldsymbol{\beta}$. Because a_i and $\mathbf{b}_i$ have centrally symmetric distributions about their α and $\boldsymbol{\beta}$, respectively, it is clear that these are the only sensible measures of central tendency in the distribution of $\mathbf{c}_i$.

Usually, we are interested in how covariates affect quantiles other than the median, in which case quantile estimation is applied to a sequence of linear models. Write the τ^{th} quantile in the distribution $D(y_i|\mathbf{x}_i)$ as $\text{Quant}_\tau(y_i|\mathbf{x}_i)$. Under linearity,

$$\text{Quant}_\tau(y_i|\mathbf{x}_i) = \alpha(\tau) + \mathbf{x}_i\boldsymbol{\beta}(\tau) \tag{1.10}$$

where, in general, the intercept and slopes depend on the quantile, τ. Under (1.10), consistent estimators of $\alpha(\tau)$ and $\boldsymbol{\beta}(\tau)$ are obtained by minimizing the *asymmetric absolute loss function* or the "check" function:

$$\min_{\alpha\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^K} \sum_{i=1}^{N} c_\tau(y_i - \alpha - \mathbf{x}_i\boldsymbol{\beta}), \tag{1.11}$$

where

$$c_\tau(u) = (\tau 1[u \geq 0] + (1-\tau)1[u < 0])|u| = (\tau - 1[u < 0])u \tag{1.12}$$

and $1[\cdot]$ is the "indicator function." Consistency is relatively easy to establish because the objective function is continuous in its parameters. Asymptotic normality is more difficult because any sensible definition of the Hessian of the objective function, away from the nondifferentiable kink, is identically zero. But it has been worked out under a variety of conditions; see Koenker (2005) for a recent treatment.

2. Some Useful Asymptotic Results

2.1. What Happens if the Quantile Function is Misspecified?

When we use OLS to estimate the parameters of a linear model, we always have a simple characterization of the plim of the OLS estimator when the mean is not linear: If α^* and $\boldsymbol{\beta}^*$ are the plims from the OLS regression y_i on $1, \mathbf{x}_i$ then these provide the smallest mean squared error approximation to $E(y|\mathbf{x}) = \mu(\mathbf{x})$. In other words, $(\alpha^*, \boldsymbol{\beta}^*)$ solves

$$\min_{a,\mathbf{b}} E[(\mu(\mathbf{x}) - \alpha - \mathbf{x}\boldsymbol{\beta})^2], \tag{2.1}$$

where, of course, the expectation is over the distribution of $\mathbf{x}$. Under some restrictions, (albeit restrictive), β_j^* is the average partial effect $E_{\mathbf{x}}[\partial\mu(\mathbf{x})/\partial x_j]$ – multivariate normality of $\mathbf{x}$ is sufficient – and under less restrictive (but still restrictive) assumptions, the β_j^* estimate the average partial effects up. These follow from the work of Chung and Goldberger (1984), Ruud (1984), and Stoker (1986).

Although the linear formulation of quantiles has been viewed by some – for example, Buchinsky (1991) and Chamberlain (1991) – as a linear approximation to the true conditional quantile, most of the the linear model is treated as being correctly specified. In some ways, this is strange because usually many quantiles are estimated. Yet assuming that different quantiles are linear in the same functions of $\mathbf{x}$ might be unrealistic.

Angrist, Chernozhukov, and Fernandez-Val (2006) provide a treatment of quantile regression under misspecification of the quantile function and characterize the probability limit of the LAD estimator. To describe the result, absorb the intercept into $\mathbf{x}$ and, rather than assume a correctly specified conditional quantile, let $\boldsymbol{\beta}(\tau)$ be the solution to the population quantile regression problem. Therefore, $\mathbf{x}\boldsymbol{\beta}(\tau)$ is the plim of the estimated quantile function. ACF have a couple of different ways to characterize $\boldsymbol{\beta}(\tau)$. One result is that $\boldsymbol{\beta}(\tau)$ solves

$$\min_{\beta} E\{w_\tau(\mathbf{x},\boldsymbol{\beta})[q_\tau(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta})]^2\}, \tag{2.2}$$

where the weight function $w_\tau(\mathbf{x},\boldsymbol{\beta})$ is

$$w_\tau(\mathbf{x},\boldsymbol{\beta}) = \int_0^1 (1-u) f_{y|x}(u\mathbf{x}\boldsymbol{\beta} + (1-u)q_\tau(\mathbf{x})|\mathbf{x})du \geq 0. \tag{2.3}$$

In other words, $\boldsymbol{\beta}(\tau)$ is the best weighted mean square approximation to the true quantile function, where the weights are the average of the conditional density of y_i over a line from the candidate approximation, $\mathbf{x}\boldsymbol{\beta}$, to the true quantile function, $q_\tau(\mathbf{x})$. The multiplication of the density by $(1-u)$ gives more weight to points closer to the true conditional quantile. It is interesting that the ACF characterization is in terms of a weighted mean squared error, a concept we usually associate with conditional mean approximation. ACF also show an approximation where the weighting function does not depend on $\boldsymbol{\beta}$, and use it characterize a "partial" regression quantiles, and to characterize omitted variables bias with quantile regression.

2.2. Computing Standard Errors

First consider the case where we want to estimate the parameters in a linear quantile model, for a given quantile, τ. For a random draw, write

$$y_i = \mathbf{x}_i\boldsymbol{\theta} + u_i, \text{ Quant}_\tau(u_i|\mathbf{x}_i) = 0, \tag{2.4}$$

where we include unity in $\mathbf{x}_i$ so that contains an intercept and the slopes. Let $\hat{\boldsymbol{\theta}}$ be the quantile estimators, and define the quantile regression residuals, $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\theta}}$. Under weak conditions (see, for example, Koenker (2005)), $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with asymptotic variance

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \tag{2.5}$$

where

$$\mathbf{A} \equiv \text{E}[f_u(0|\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i] \tag{2.6}$$

and

$$\mathbf{B} \equiv \tau(1-\tau)\text{E}(\mathbf{x}_i'\mathbf{x}_i). \tag{2.7}$$

Expression (2.5) is the now familiar standard "sandwich" form of asymptotic variances. It is fully robust in the sense that it is valid without further assumptions on $D(u_i|\mathbf{x}_i)$. The matrix $\mathbf{B}$ is simple to estimate as

$$\hat{\mathbf{B}} = \tau(1-\tau)\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right), \tag{2.8}$$

where $0 < \tau < 1$ is the chosen quantile. This estimator is consistent under the weak assumption of finite second moments for $\mathbf{x}_i$. The matrix $\mathbf{A}$ is harder to estimate because of the presence of $f_u(0|\mathbf{x}_i)$, and we do not have a parametrix model for the density of u_i given $\mathbf{x}_i$. But we only have to estimate this conditional density at $u = 0$, so we could use a nonparametric density estimator (based on the $\hat{u}_i$). Powell (1986, 1991) proposed a simpler approach, which leads to

$$\hat{\mathbf{A}} = (2Nh_N)^{-1}\sum_{i=1}^{N}1[|\hat{u}_i| \leq h_N]\mathbf{x}_i'\mathbf{x}_i, \tag{2.9}$$

where $\{h_N > 0\}$ is a nonrandom sequence shrinking to zero as $N \to \infty$ with $\sqrt{N}\,h_N \to \infty$. are sufficient for consistency. The second condition controls how quickly h_N shrinks to zero. For example, $h_N = aN^{-1/3}$ for any $a > 0$ satisfies these conditions. The practical problem in choosing a (or choosing h_N more generally) is discussed by Koenker (2005), who also discusses some related estimators. In particular, in equation (2.9), observation i does not contribute if $|\hat{u}_i| > h_N$. Other methods allow each observation to enter the sum but with a weight that declines as $|\hat{u}_i|$ increases. (As an interesting aside, the derivation of (2.9) involves the simple equality $\mathrm{E}\{(1[|u_i| \leq h_N]|\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i\} = \mathrm{E}(1[|u_i| \leq h_N]\mathbf{x}_i'\mathbf{x}_i)$, which is analogous to the key step in the regression frameworks for justifying the heteroskedasticity-robust variance matrix estimator.)

The nonparametric bootstrap can be applied to quantile regression, but if the data set is large, the computation using several hundred bootstrap samples can be costly.

If we assume that u_i is independent of $\mathbf{x}_i$ then $f_u(0|\mathbf{x}_i) = f_u(0)$ and equation (2.5) simplifies to

$$\frac{\tau(1-\tau)}{[f_u(0)]^2}[\mathrm{E}(\mathbf{x}_i'\mathbf{x}_i)]^{-1} \tag{2.10}$$

and its estimator has the general form

$$\frac{\tau(1-\tau)}{[\hat{f}_u(0)]^2}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}, \tag{2.11}$$

and a simple, consistent estimate of $f_u(0)$ is the histogram estimator

$$\hat{f}_u(0) = (2Nh_N)^{-1} \sum_{i=1}^{N} 1[|\hat{u}_i| \leq h_N]. \tag{2.12}$$

Of course, one can use other kernel estimators for $\hat{f}_u(0)$. This nonrobust estimator is the one commonly reported as the default be statistical packages, including Stata.

If the quantile function is misspecified, even the "robust" form of the variance matrix, based on the estimate in (2.9), is not valid. In the generalized linear models and generalized estimating equations literature, the distinction is sometimes made between a "fully robust" variance estimator and a "semi-robust" variance estimator. In the GLM and GEE literatures, the semi-robust estimator assumes $E(y_i|\mathbf{x}_i)$, or the panel version of it, is correctly specified, but does not impose restrictions on $Var(y_i|\mathbf{x}_i)$ or other features of $D(y_i|\mathbf{x}_i)$. On the other hand, a fully robust variance matrix estimator is consistent for the asymptotic variance even if the mean function is misspecified. For, say, nonlinear least squares, or quasi-MLE in the linear exponential family, one needs to include the second derivative matrix of the conditional mean function to have a fully robust estimator. For some combinations of mean functions and objective functions, the Hessian of the mean function disappears, and the fully robust and semi-robust estimators are the same. For two-step methods, such as GEE, analytical formulas for fully robust estimators are very difficult to obtain, and almost all applications use the semi-robust form. This is a long-winded way to say that there is precedent for worrying about how to estimate asymptotic variances when the main feature being estimated is misspecified. In GEE terminology, $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ where $\hat{\mathbf{A}}$ is given by (2.9), is only semi-robust.

Kim and White (2002) and Angrist, Chernozhukov, and Fernández-Val (2006) provide a fully robust variance matrix estimator when the linear quantile function is possibly misspecified. The estimator of $\mathbf{A}$ in (2.9) is still valid, but the estimator of $\mathbf{B}$ needs to be extended. If we use the outer product of the score we obtain

$$\hat{\mathbf{B}} = \left(N^{-1} \sum_{i=1}^{N} (\tau - 1[\hat{u}_i < 0])^2 \mathbf{x}_i' \mathbf{x}_i \right), \tag{2.13}$$

where the $\hat{u}_i$ are the residuals from the (possibly) misspecified quantile regression, is generally consistent.

As shown by Hahn (1995, 1997), the nonparametric bootstrap (and the Bayesian bootstrap) generally provides consistent estimates of the fully robust variance without claims about the conditional mean being correct. It does not, however, provide asymptotic refinements for

testing and confidence intervals compared with those based on first-order asymptotics. See Horowitz (2001) for a discussion, and on how to smooth the problem so that refinements are possible.

ACF actually provide the covariance function for the process $\{\hat{\boldsymbol{\theta}}(\tau) : \varepsilon \leq \tau \leq 1 - \varepsilon\}$ for some $\varepsilon > 0$, which can be used to test hypotheses jointly across multiple quantiles (including all quantiles at once).

As an example of quantile regression, we use the data from Abadie (2003). Stata was used to do the estimation and obtain the standard errors; these are the nonrobust standard errors that use

Dependent Variable:	*nettfa*					
	(1)	(2)	(3)	(4)	(5)	(6)
Explanatory Variable	Mean (OLS)	.10 Quantile	.25 Quantile	Median (LAD)	.75 Quantile	.90 Quantile
inc	.783	−.0179	.0713	.324	.798	1.291
	(.104)	(.0177)	(.0072)	(.012)	(.025)	(.048)
age	−1.568	−.0663	.0336	−.244	−1.386	−3.579
	(1.076)	(.2307)	(.0955)	(.146)	(.287)	(.501)
age^2	.0284	.0024	.0004	.0048	.0242	.0605
	(.0138)	(.0027)	(.0011)	(.0017)	(.0034)	(.0059)
e401k	6.837	.949	1.281	2.598	4.460	6.001
	(2.173)	(.617)	(.263)	(.404)	(.801)	(1.437)
N	2,017	2,017	2,017	2,017	2,017	2,017

The effect of income is very different across quantiles, with its largest effect at upper quantiles. Similarly, eligibility for a 401(k) plan has a much larger effect on financial wealth at the upper end of the wealth distribution. The mean and median slope estimates are very different, implying that the model with an additive error that is either independent of the covariates, or has a symmetric distribution given the covariates, is not a good characterization.

3. Quantile Regression with Endogenous Explanatory Variables

Recently, there has been much interest in using quantile regression in models with endogenous explanatory variables. Some strategies are fairly simple, others are more complicated. Suppose we start with the model

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{3.1}$$

where the full vector of exogenous variables is $\mathbf{z}$ and y_2 is potential endogenous – whatever

that means in the context of quantile regression. The most straightforward case to handle is least absolute deviations, because median restrictions are easier to justify when joint distributions are involved.

Amemiya's (1982) two-stage LAD estimator, whose asymptotic properties were derived by Powell (1986), adds a reduced form for y_2, say

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2. \tag{3.2}$$

While (3.2) can be estimated by OLS to obtain $\hat{\boldsymbol{\pi}}_2$, using LAD in the first stage to estimate $\boldsymbol{\pi}_2$ is more in the spirit of 2SLAD. In the second step, the fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, are inserted in place of y_{i2} to given LAD of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$. By replacing $\hat{\boldsymbol{\pi}}_2$ with $\boldsymbol{\pi}_2$, it is clear that the 2SLAD estimator essentially requires symmetry of the composite error $\alpha_1 v_2 + u_1$. While the properties of 2SLAD were originally worked out for nonstochastic $\mathbf{z}_i$ – so that (u_{i1}, v_{i2}) is independent of $\mathbf{z}_i$ – it is clear that symmetry of $\alpha_1 v_2 + u_1$ given $\mathbf{z}$ is sufficient.

We might as well assume $D(u_1, v_2|\mathbf{z})$ is centrally symmetric, in which case a control function approach can be used, too. Write

$$u_1 = \rho_1 v_2 + e_1, \tag{3.3}$$

where e_1 given $\mathbf{z}$ would have a symmetric distribution. Because $Med(v_2|\mathbf{z}) = 0$, the first stage estimator can be LAD. Given the LAD residuals $\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, these residuals can be added to second-stage LAD. So, we do LAD of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$. It seems likely that a t test on $\hat{v}_{i2}$ is valid as a test for the null that y_2 is exogenous.

There can be problems of interpretation in just applying either 2SLAD or the CF approach. Suppose we view this as an omitted variable problem, where a_1 is the omitted variable, and interest lies in the "structual" median

$$Med(y_1|\mathbf{z}, y_2, a_1) = Med(y_1|\mathbf{z}_1, y_2, a_1) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + a_1. \tag{3.4}$$

Then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + a_1 + e_1 \tag{3.5}$$

$$Med(e_1|\mathbf{z}, y_2, a_1) = 0. \tag{3.6}$$

If (3.4) was stated in terms of means, then $E(e_1|\mathbf{z}) = 0$ by construction, and a very sensible exogeneity condition is $E(a_1|\mathbf{z}) = E(a_1) = 0$ (as a normalization) or that $Cov(\mathbf{z}, a_1) = \mathbf{0}$. But here we cannot even assert that $Med(e_1|\mathbf{z}) = Med(e_1)$ because (3.6) does not imply this; there is no law of iterated medians. To further compound the problem, the median of the sum is not the sum of the medians; so, even if we stated exogeneity as $Med(a_1|\mathbf{z}) = Med(a_1)$ and just

asserted $Med(e_1|\mathbf{z}) = Med(e_1)$, $a_1 + e_1 = u_1$ would not generally satisfy $Med(u_1|\mathbf{z}) = Med(u_1)$. Of course, we can make enough multivariate symmetric assumptions so that all linear combinations of errors have symmetric distributions. But then LAD methods are purely to guard against outliers; usual 2SLS will provide consisten, asymptotically normal estimates of the parameters under symmetry (and, of course, weaker assumptions).

With quantile estimation, such two-step estimators are even more difficult to justify. The Angrist, Chernozhukov, and Fernandez-Val (2006) partialling out representations can provide some sort of interpretation of netting out the control, v_2, but it is difficult to know whether the parameters are ultimately interesting.

Abadie (2003) and Abadie, Angrist, and Imbens (2002) show how to define and estimate policy parameters with a binary endogenous treatment, say *D*, and binary instrumental variable, say *Z*. The outcome is *Y* with observed covariates, *X*. The potential outcomes on *Y* are Y_d, $d = 0, 1$ – that is, without treatment and with treatment, respectively. The counterfactuals for treatment are D_z, $z = 0, 1$. Thus, D_0 is what treatment status would be if the instrument (often, randomized eligibility) equals zero, and D_1 is treatment status if $Z = 1$. The data we observe are $X, Z, D = (1 - Z)D_0 + ZD_1$, and $Y = (1 - D)Y_0 + DY_1$. As discussed in AAI, identification of average treatment effects, and ATE on the treated, is difficult. Instead, they focus on treatment effects for *compliers*, that is, the (unobserved) subpopulation with $D_1 > D_0$. This is the group of subjects who do not participate if ineligible but do participate if eligible.

AAI specify the linear equation

$$Quant_\tau(Y|X, D, D_1 > D_0) = \alpha_\tau D + X\beta_\tau, \tag{3.7}$$

and define α_τ as the *quantile treatment effect* (QTE) for compliers. If we observed the event $D_1 > D_0$, then (3.7) could be estimated by standard quantile regression using the subsample of compliers. But, in effect, the binary variable $1[D_1 > D_0]$ is an omitted variable. But *Z* is an available instrument for *D*. As discussed by AAI, (3.7) identifies differences in quantiles on the potential outcomes, Y_1 and Y_0, and not the quantile of the difference, $Y_1 - Y_0$. The latter effects are harder to identify. (Of course, in the case of mean effects, there is no difference in the two effects.)

The assumptions used by AAI to identify α_τ are

$$(Y_1, Y_0, D_1, D_0) \text{ is independent of } Z \text{ conditional on } X \tag{3.8}$$

$$0 < P(Z = 1|X) < 1 \tag{3.9}$$

$$P(D_1 = 1|X) \neq P(D_0 = 1|X) \tag{3.10}$$

$$P(D_1 \geq D_0|X) = 1. \tag{3.11}$$

Under these assumptions, AAI show that a weighted quantile estimation identifies α_τ. The estimator that is computationally most convenient is obtained as follows. Define

$$\kappa_v(U) = 1 - \frac{D(1 - v(U))}{1 - \pi(X)} - \frac{(1 - D)v(U)}{\pi(X)}, \tag{3.12}$$

where $U = (Y, D, X)$, $v(U) = P(Z = 1|U)$, and $\pi(X) = P(Z = 1|X)$. AAI show that $\kappa_\tau(u) = P(D_1 > D_0|U = u)$, and so this weighting function is nonnegative. They also show that α_τ and β_τ in (3.7) solve

$$\min_{\alpha,\beta} E[\kappa_\tau(U)c_\tau(Y - \alpha D - X\beta)], \tag{3.13}$$

where $c_\tau(\cdot)$ is the check function defined earlier. To operationalize the estimate, $\kappa_\tau(\cdot)$ needs to be estimated, which means estimating $P(Z = 1|Y, D, X)$ and $P(Z = 1|X)$. AAI use linear series estimators to approximate $P(Z = 1|Y, D, X)$ and $P(Z = 1|X)$, and derive the asymptotic variance of the two-step estimator that solves

$$\min_{\delta} \sum_{i=1}^{N} 1[\hat{\kappa}_v(U_i) \geq 0]\hat{\kappa}_v(U_i)c_\tau(Y_i - W_i\delta), \tag{3.14}$$

where $W_i = (D_i, X_i)$ and δ contains α and β. The indicator function $1[\hat{\kappa}_v(U_i) \geq 0]$ ensures that only observations with nonnegative weights are used. Asymptotically, $\hat{\kappa}_v(u) \geq 0$, and this trimming of observations becomes less and less necessary. To ensure that $\hat{v}(u)$ and $\hat{\pi}(x)$ act like probabilities, series estimation using logit functions, as in Hirano, Imbens, and Ridder (2003), might be preferred (although that still would not ensure nonnegativity of $\hat{\kappa}_v(U_i)$ for all i).

Other recent work has looked at quantile estimation with endogenous treatment effects. Chernozhukov and Hansen (2005, 2006) consider indentification and estimation of QTEs in a model with endogenous treatment and without imposing functional form restrictions. Let $q(d, x, \tau)$ denote the τ^{th} quantile function for treatment level $D = d$ and covariates x. In the binary case, CH define the QTE as

$$QTE_\tau(x) = q(1, x, \tau) - q(0, x, \tau). \tag{3.15}$$

Using a basic result from probability, the average treatment effect, again conditional on x, can be obtained by integrating (3.15) over $0 < \tau < 1$.

The critical representation used by CH is that each potential outcome, Y_d, conditional on $X = x$, can be expressed as

$$Y_d = q(d, x, U_d) \tag{3.16}$$

where

$$U_d|Z \sim \text{Uniform}(0, 1), \tag{3.17}$$

and Z is the instrumental variable for treatment assignment, D. Thus, D is allowed to be correlated with U_d. Key assumptions are that $q(d, x, u)$ is strictly increasing in u and a "rank invariance" condition. The simplest form of the condition is that , conditional on $X = x$ and $Z = z$, U_d does not depend on d. The CH show that, with the observed Y defined as $Y = q(D, X, U_D)$,

$$P[Y \leq q(D, X, \tau)|X, Z] = P[Y < q(D, X, \tau)|X, Z] = \tau. \tag{3.18}$$

Equation (3.18) acts as a nonparametric conditional moment condition which, under certain assumptions, allows identification of $q(d, x, \tau)$. If we define $R = Y - q(D, X, \tau)$, then (3.18) implies that the τ^{th} quantile of R, conditional on (X, Z), is zero. This is similar to the more common situation where we have a conditional moment condition of the form $E(R|X, Z) = 0$. See Chernozhukov and Hansen (2005) for details concerning identification – they apply results of Newey and Powell (2003) – and Chernozhukov and Hansen (2005) for estimation methods, where they assume a linear form for $q(d, x, \tau)$ and obtain what they call the *quantile regression instrumental variables estimator*.

Other work that uses monotonicity assumptions and identifies structural quantile functions is Chesher (2003) and Imbens and Newey (2006).

4. Quantile Regression for Panel Data

Quantile regression methods can be applied to panel data, too. For a given quantile $0 < \tau < 1$, suppose we specify

$$\text{Quant}_\tau(y_{it}|\mathbf{x}_{it}) = \mathbf{x}_{it}\boldsymbol{\theta}, \; t = 1, \ldots, T, \tag{4.1}$$

where $\mathbf{x}_{it}$ probably allows for a full set of time period intercepts. Of course, we can write $y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + u_{it}$ where $\text{Quant}_\tau(u_{it}|\mathbf{x}_{it}) = 0$. The natural estimator of $\boldsymbol{\theta}_o$ is the pooled quantile regression estimator. Unless we assume that (3.1) has correctly specified dynamics, the variance matrix needs to be adjusted for serial correlation in the resulting score of the objective

function. These scores have the form

$$\mathbf{s}_{it}(\boldsymbol{\theta}) = -\mathbf{x}_{it}'\{\tau 1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} \geq 0] - (1-\tau)1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} < 0]\}, \tag{4.2}$$

which can be shown to have zero mean (at the "true" parameter), conditional on $\mathbf{x}_{it}$, under (4.1). The serial dependence properties are not restricted, nor is heterogeneity in the distributions across t. A consistent estimator of $\mathbf{B}$ (with T fixed and $N \to \infty$) is

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{ir}(\hat{\boldsymbol{\theta}})'. \tag{4.3}$$

This estimator is not robust to misspecification of the conditional quantiles, but the extension of Angrist, Chernozhukov, and Fernandez-Val (2006) should work in the pooled panel data case as well..

Estimation of $\mathbf{A}$ is similar to the cross section case. A robust estimator, that does not assume independence between u_{it} and $\mathbf{x}_{it}$, and allows the distribution of u_{it} to change across t, is

$$\hat{\mathbf{A}} = (2Nh_N)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}1[|\hat{u}_{it}| \leq h_N]\mathbf{x}_{it}'\mathbf{x}_{it}, \tag{4.4}$$

or, we can replace the indicator function with a smoothed version. Rather than using $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$ as the estimate of the asymptotic variance of $\hat{\boldsymbol{\theta}}$, the bootstrap can be applied by resampling cross section units.

Allowing explicitly for unobserved effects in quantile regression is trickier. For a given quantile $0 < \tau < 1$, a natural specification, which incorporates strict exogeneity conditional on c_i, is

$$\text{Quant}_\tau(y_{it}|\mathbf{x}_i, c_i) = \text{Quant}_\tau(y_{it}|\mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\boldsymbol{\theta} + c_i,\ t = 1, \ldots, T, \tag{4.5}$$

which is reminiscent of the way we specified the conditional mean in Chapter 10. Equivalently, we can write

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + c_i + u_{it},\ \text{Quant}_\tau(u_{it}|\mathbf{x}_i, c_i) = 0,\ t = 1, \ldots, T. \tag{4.6}$$

Unfortunately, unlike in the case of estimating effects on the conditional mean, we cannot proceed without further assumptions. A "fixed effects" approach, where we allow $\text{D}(c_i|\mathbf{x}_i)$ to be unrestricted, is attractive. Generally, there are no simple transformations to eliminate c_i and estimate $\boldsymbol{\theta}$. If we treat the c_i as parameters to estimate along with $\boldsymbol{\theta}$, the resulting estimator generally suffers from an incidental parameters problem. Briefly, if we try to estimate c_i for

each i then, with large N and small T, the poor quality of the estimates of c_i causes the accompanying estimate of $\boldsymbol{\theta}$ to be badly behaved. Recall that this was *not* the case when we used the FE estimator for a conditional mean: treating the c_i as parameters led us to the within estimator. Koenker (2004) derives asymptotic properties of this estimation procedure when T grows along with N, but also adds the assumptions that the regressors are fixed and $\{u_{it} : t = 1, \ldots, T\}$ is serially independent.

An alternative approach is suggested by Abrevaya and Dahl (2006) for $T = 2$. They are motivated by Chamberlain's correlated random effects linear model. In the $T = 2$ case, Chamberlain (1982) specifies

$$E(y_t|\mathbf{x}_1, \mathbf{x}_2) = \psi_t + \mathbf{x}_t\boldsymbol{\beta} + \mathbf{x}_1\boldsymbol{\xi}_1 + \mathbf{x}_2\boldsymbol{\xi}_2, \; t = 1, 2. \tag{4.7}$$

Notice that $\partial E(y_1|\mathbf{x})/\partial \mathbf{x}_1 = \boldsymbol{\beta} + \boldsymbol{\xi}_1$ and $\partial E(y_2|\mathbf{x})/\partial \mathbf{x}_1 = \boldsymbol{\xi}_1$. Therefore,

$$\boldsymbol{\beta} = \frac{\partial E(y_1|\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial E(y_2|\mathbf{x})}{\partial \mathbf{x}_1}, \tag{4.8}$$

and similarly if we reverse the roles of $\mathbf{x}_1$ and $\mathbf{x}_2$. Abrevaya and Dahl use this motivation to estimate separate linear quantile regressions $\text{Quant}_\tau(y_t|\mathbf{x}_1, \mathbf{x}_2)$ – reminiscent of Chamberlain's method – and then define the partial effect as

$$\boldsymbol{\beta}_\tau = \frac{\partial \text{Quant}_\tau(y_1|\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial \text{Quant}_\tau(y_2|\mathbf{x})}{\partial \mathbf{x}_1}. \tag{4.9}$$

For quantile regression, CRE approaches are generically hampered because finding quantiles of sums of random variables is difficult. For example, suppose we impose the Mundlak representation $c_i = \psi_o + \bar{\mathbf{x}}_i\boldsymbol{\xi}_o + a_i$. Then we can write $y_{it} = \psi_o + \mathbf{x}_{it}\boldsymbol{\theta}_o + \bar{\mathbf{x}}_i\boldsymbol{\xi}_o + a_i + u_{it} \equiv y_{it} = \psi_o + \mathbf{x}_{it}\boldsymbol{\theta}_o + \bar{\mathbf{x}}_i\boldsymbol{\xi}_o + v_{it}$, where v_{it} is the composite error. Now, if we assume v_{it} is independent of $\mathbf{x}_i$, then we can estimate $\boldsymbol{\theta}_o$ and $\boldsymbol{\xi}_o$ using pooled quantile regression of y_{it} on $1, \mathbf{x}_{it}$, and $\bar{\mathbf{x}}_i$. (The intercept does not estimate a quantity of particular interest.) But independence is very strong, and, if we truly believe it, then we probably believe all quantile functions are parallel. Of course, we can always just assert that the effect of interest is the set of coefficients on $\mathbf{x}_{it}$ in the pooled quantile estimation, and we allow these, along with the intercept and coefficients on $\bar{\mathbf{x}}_i$, to change across quantile. The asymptotic variance matrix estimator discussed for pooled quantile regression applies directly once we define the explanatory variables at time t to be $(1, \mathbf{x}_{it}, \bar{\mathbf{x}}_i)$.

We have more flexibility if we are interested in the median, and a few simple approaches suggest themselves. Write the model $\text{Med}(y_{it}|\mathbf{x}_i, c_i) = \text{Med}(y_{it}|\mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\boldsymbol{\theta} + c_i$ in error form

as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + c_i + u_{it}, \text{Med}(u_{it}|\mathbf{x}_i, c_i) = 0, \quad t = 1, .., T \tag{4.10}$$

and consider the multivariate conditional distribution $D(\mathbf{u}_i|\mathbf{x}_i)$. Above we discussed the centrally symmetric assumption, conditional on $\mathbf{x}_i$: $D(\mathbf{u}_i|\mathbf{x}_i) = D(-\mathbf{u}_i|\mathbf{x}_i)$. If we make this assumption, then the time-demeaned errors $\ddot{u}_{it}$ have (univariate) conditional (on $\mathbf{x}_i$) distributions symmetric about zero, which means we can consistently estimate $\boldsymbol{\theta}$ by applying pooled least absolute deviations to the time-demeaned equation $\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\theta} + \ddot{u}_{it}$, being sure to obtain fully robust standard errors for pooled LAD.

Alternatively, under the centrally symmetric assumption, the difference in the errors, $\Delta u_{it} = u_{it} - u_{i,t-1}$ have symmetric distributions about zero, so one can apply pooled LAD to $\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\theta} + \Delta u_{it}, \ t = 2, \ldots, T$. From Honoré (1992) applied to the uncensored case, LAD on the first differences is consistent when $\{u_{it} : t = 1, \ldots, T\}$ is an i.i.d. sequence conditional on $(\mathbf{x}_i, c_i)$, even if the common distribution is not symmetric – and this may afford robustness for LAD on the first differences rather than on the time-demeaned data. Interestingly, it follows from the discussion in Honoré (1992, Appendix 1) that when $T = 2$, applying LAD on the first differences is equivalent to estimating the c_i along with $\boldsymbol{\theta}_o$. So, in this case, there is no incidental parameters problem in estimating the c_i as long as $u_{i2} - u_{i1}$ has a symmetric distribution. Although not an especially weak assumption, central symmetry of $D(\mathbf{u}_i|\mathbf{x}_i)$ allows for serial dependence and heteroskedasticity in the u_{it} (both of which can depend on $\mathbf{x}_i$ or on t). As always, we should be cautious in comparing the pooled OLS and pooled LAD estimates of $\boldsymbol{\theta}$ on the demeaned or differenced data because they are only expected to be similar under the conditional symmetry assumption.

If we impose the Mundlak-Chamberlain device, we can get by with conditional symmetry of a sequence of bivariate distributions. Write $y_{it} = \psi_t + \mathbf{x}_{it}\boldsymbol{\theta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it}$, where $\text{Med}(u_{it}|\mathbf{x}_i, a_i) = 0$. If $D(a_i, u_{it}|\mathbf{x}_i)$ has a symmetric distribution around zero then $D(a_i + u_{it}|\mathbf{x}_i)$ is symmetric about zero, and, if this holds for each t, pooled LAD of y_{it} on $1, \mathbf{x}_{it}$, and $\bar{\mathbf{x}}_i$ consistently estimates $(\psi_t, \boldsymbol{\theta}, \boldsymbol{\xi})$. (Therefore, we can estimate the partial effects on $\text{Med}(y_{it}|\mathbf{x}_{it}, c_i)$ and also test if c_i is correlated with $\bar{\mathbf{x}}_i$.) The assumptions used for this approach are not as weak as we would like, but, like using pooled LAD on the time-demeaned data, adding $\bar{\mathbf{x}}_i$ to pooled LAD gives a way to compare with the usual FE estimate of $\boldsymbol{\theta}$.. (Remember, if we use pooled OLS with $\bar{\mathbf{x}}_i$ included, we obtain the FE estimate.) Fully robust inference can be obtained by computing $\hat{\mathbf{B}}$ and $\hat{\mathbf{A}}$ in (4.3) and (4.4).

5. Quantile Methods for "Censored" Data

As is well known, the statistical structure of parametric models for data that have truly been censored – such as top-coded wealth, or a right-censored duration – is essentially the same as models for corner solution responses – that is, variables that have a mass point, or pile up, at one or couple of values (usually, zero). Examples are labor supply, charitable contributions, and amount of life insurance. But an important point is that the interpretation of the estimates is different in these two cases. In the data censoring case, there is an underlying linear model (usually) whose coefficients we are interested in. For example, we are interested in the conditional distribution of wealth given covariates. That wealth has been top-coded means that we do not observe underlying wealth over its entire range. In effect, it is a missing data problem. The same is true with duration models.

In the corner solution case, we observe the response of interest over its entire range. We use models such as Tobit simply because we want to recognize the mass point or points. Linear functional forms for the mean, say, can miss important nonlinearities. When we apply, say, standard Tobit to a corner solution, y, we are interested in features of $D(y|\mathbf{x})$, such as $P(y > 0|\mathbf{x})$, $E(y|\mathbf{x}, y > 0)$, and $E(y|\mathbf{x})$. While the parameters in the model are important, they do not directly provide partial effects on the quantities of interest. Of course, if we use a linear model approximation for, say, $E(y|\mathbf{x})$, then the coefficients are approximate partial effects. A related point is: if we modify standard models for corner responses, say, consider heteroskedasticity in the latent error of a Tobit, we should consider how it affects $D(y|\mathbf{x})$, and not just the paramter estimates. In the case of censored data, it is the parameters of the underlying linear model we are interested in, and then it makes much more sense to focus on parameter sensitivity.

In applying LAD methods to "censored" outcomes, we should also be aware of the difference between true data censoring and corner solution responses. With true data censoring we clearly have an interest in obtaining estimates of, say,

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + u_i, \tag{5.1}$$

where y_i^* is the variable we would like to explain. If y_i^* is top coded at, say, r_i, then we observe $y_i = \min(y_i^*, r_i)$. If we assume $D(u_i|\mathbf{x}_i, r_i) = Normal(0, \sigma^2)$, then we can apply censored normal regression (also called type I Tobit). This method applies even if r_i is observed only when y_i^* has been censored, which happens sometimes in duration studies. As shown by Powell (1986), we can estimate (5.1) under much weaker assumptions than normality:

$$Med(u_i|\mathbf{x}_i, r_i) = 0 \tag{5.2}$$

suffices, provided the censoring values value, r_i, are always observed. Because the median passes through monotonic functions,

$$\begin{aligned} Med(y_i|\mathbf{x}_i, r_i) &= Med[\min(\mathbf{x}_i\boldsymbol{\beta} + u_i, r_i)|\mathbf{x}_i, r_i] \\ &= \min[Med(\mathbf{x}_i\boldsymbol{\beta} + u_i|\mathbf{x}_i, r_i), r_i] \\ &= \min(\mathbf{x}_i\boldsymbol{\beta}, r_i). \end{aligned} \tag{5.3}$$

Because LAD consistently estimates the parameters of a conditional median, at least under certain regularity conditions, (5.3) suggest estimate $\boldsymbol{\beta}$ as the solution to

$$\min_{\mathbf{b}} \sum_{i=1}^{N} |y_i - \min(\mathbf{x}_i\mathbf{b}, r_i)|. \tag{5.4}$$

Powell (1986) showed that, even though the objective function has a corner it it, the *censored least absolute deviations* (CLAD) estimator is $\sqrt{N}$-asymptotically normal. Honoré, Khan, and Powell (2002) provide methods that can be used when r_i is not always observed.

CLAD can also be applied to corner solution responses. Suppose the variable of interest, y_i, has a corner at zero, and is determined by

$$y = \max(0, \mathbf{x}\boldsymbol{\beta} + u). \tag{5.5}$$

If $D(u|\mathbf{x})$ is Normal$(0, \sigma^2)$, then the MLE is the type I Tobit. Given $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, we can compute partial effects on the mean and various probabilities. The partial effects on $Med(y|\mathbf{x})$ depend only on $\boldsymbol{\beta}$, because

$$Med(y|\mathbf{x}) = \max(0, \mathbf{x}\boldsymbol{\beta}). \tag{5.6}$$

Of course, (5.6) provides a way to estimate $\boldsymbol{\beta}$ by CLAD under just

$$Med(u|\mathbf{x}) = 0. \tag{5.7}$$

The β_j measure the partial effects on $Med(y|\mathbf{x})$ once $Med(y|\mathbf{x}) > 0$.

Once we recognize in corner solution applications that it is features of $D(y|\mathbf{x})$ that are of interest, (5.6) becomes just a particular feature of $D(y|\mathbf{x})$ that we can identify, and it is no better or worse than other features of $D(y|\mathbf{x})$ that we might want to estimate, such as a quantile other than the median, or the mean $E(y|\mathbf{x})$, or the "conditional" mean, $E(y|\mathbf{x}, y > 0)$. Emphasis is often given on the fact that the functional form for the median in (5.6) holds very generally when (5.5) holds; other than (5.7), no restrictions are made on the shape of the distribution $D(u|\mathbf{x})$ or of its dependence on $\mathbf{x}$. But for corner solution responses, there is nothing sacred

about (5.5). In fact, it is pretty restrictive because y depends on only one unobservable, u. Two-part models, summarized recently in Wooldridge (2007), allow more flexibility.

A model that is no more or less restrictive than (5.5) is

$$y = a \cdot \exp(\mathbf{x}\boldsymbol{\beta}), \tag{5.8}$$

where the only assumption we make is

$$E(a|\mathbf{x}) = 1, \tag{5.9}$$

where $D(a|\mathbf{x})$ is otherwise unrestricted. In particular, we do not know $P(a = 0|\mathbf{x})$, which is positive if y has mass point at zero, or $Med(a|\mathbf{x})$. Under (5.9),

$$E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}), \tag{5.10}$$

which means we can consistently estimate $\boldsymbol{\beta}$ using nonlinear regression or a quasi-MLE in the linear exponential family (such as Poisson or Gamma); it does not matter that y is a corner if its mean is given by (5.10). The point here is that, if we simply focus on assumptions and what can be identified under those assumptions, the model in (5.8) and (5.9) identifies just as many features of $D(y|\mathbf{x})$ as the model in (5.5) and (5.7). They are different features, but neither is inherently better than the other.

Continuing with this point, we can modify (5.5) rather simply and see that CLAD breaks down. Suppose we add multiplicative heterogeneity:

$$y = a \cdot \max(0, \mathbf{x}\boldsymbol{\beta} + u), \tag{5.11}$$

where $a \geq 0$, and even make the strong assumption that a is independent of $(\mathbf{x}, u)$. The distribution $D(y|\mathbf{x})$ now depends on the distribution of a, and does not follow a type I Tobit model; generally, finding its distribution would be difficult, even if we specify a simple distribution for a. Nevertheless, if we normalize $\mathrm{E}(a) = 1$, then $\mathrm{E}(y|\mathbf{x}, u) = \mathrm{E}(a|\mathbf{x}, u) \cdot \max(0, \mathbf{x}\boldsymbol{\beta} + u) = \max(0, \mathbf{x}\boldsymbol{\beta} + u)$ (because $\mathrm{E}(a|\mathbf{x}, u) = 1$). It follows immediately by iterated expectations that if assumption (17.3) holds, then $E(y|\mathbf{x})$ has exactly the same form as the type I Tobit model:

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma). \tag{5.12}$$

Therefore, the parameters $\boldsymbol{\beta}$ and σ^2 are identified and could be estimate by nonlinear least squares or weighted NLS, or a quasi-MLE using the mean function (5.12). Note that $D(y|\mathbf{x})$ does not follow the type I Tobit distribution, so MLE is not available.

On the other hand, if we focus on the median, we have

$$Med(y|\mathbf{x}, a) = a \cdot \max(0, \mathbf{x}\boldsymbol{\beta}). \tag{5.13}$$

But there is no "law of iterated median," so, generaly, we cannot determine $Med(y|\mathbf{x})$ without further assumptions. One might argue that we are still interested in the β_j because they measure the average partial effects on the median. But the do not appear to be generally identified under this variation on the standard Tobit model.

The issues in applying CLAD to corners gets even trickier in panel data applications. Suppose

$$y_{it} = \max(0, \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}) \tag{5.14}$$

$$Med(u_{it}|\mathbf{x}_i, c_i) = 0, \tag{5.15}$$

so that (5.15) embodies strict exogeneity of $\mathbf{x}_{it}$ conditional on c_i. Under (5.14) and (5.15),

$$Med(y_{it}|\mathbf{x}_i, c_i) = \max(0, \mathbf{x}_{it}\boldsymbol{\beta} + c_i). \tag{5.16}$$

Honoré (1992) and Honoré and Hu (2004) provide methods of estimating $\boldsymbol{\beta}$ without making any assumptions about the distribution of c_i, or restricting its dependence on $\mathbf{x}_i$. They do assume conditional exchangeability assumptions on the u_{it}; sufficient is independence with $\mathbf{x}_i$ and $\{u_{it}\}$ i.i.d. over t. Given estimates of the β_j, we can estimate the partial effects of the x_{tj} on $Med(y_t|\mathbf{x}_t, c)$ for $Med(y_t|\mathbf{x}_t, c) > 0$. Unfortunately, because we do not observe c_i, or know anything about its distribution, we do not know when the nonzero effect kicks in. We can write the partial effect of x_{tj} as

$$\theta_{tj}(\mathbf{x}_t, c) = 1[\mathbf{x}_t\boldsymbol{\beta} + c > 0]\beta_j. \tag{5.17}$$

We might be interested in averaging these across the distribution of unobserved heterogeneity, but this distribution is not identified. (Interesting, if c_i has a $Normal(\mu_c, \sigma_c^2)$ distribution, then it is easy to show that the average of (5.17) across the heterogeneity is $E_{c_i}[\theta_{tj}(\mathbf{x}_t, c_i)] = \Phi[(\mu_c - \mathbf{x}_t\boldsymbol{\beta})/\sigma_c]\beta_j$, and we can see immediatly that it depends on the location and scale of c_i.)

We can compare the situation of the median with the mean. Using the Altonji and Matkin (2005), suppose we assume $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$. Then $E(y_{it}|\mathbf{x}_i) = g_t(\mathbf{x}_{it}\boldsymbol{\beta}, \bar{\mathbf{x}}_i)$ for some unknown function $g_t(\cdot, \cdot)$, and $\boldsymbol{\beta}$ is identified (usually only up to scale) and the average partial effects on the mean are generally identified.

References

(To be added.)

What's New in Econometrics **NBER, Summer 2007**

Lecture 15, Wednesday, Aug 1st, 4.30-5.30pm

Generalized Method of Moments and Empirical Likelihood

1. Introduction

Generalized Method of Moments (henceforth GMM) estimation has become an important unifying framework for inference in econometrics in the last twenty years. It can be thought of as nesting almost all the common estimation methods such as maximum likelihood, ordinary least squares, instrumental variables and two–stage–least–squares and nowadays it is an important part of all advanced econometrics text books (Gallant, 1987; Davidson and McKinnon, 1993; Hamilton, 1994; Hayashi, 2000; Mittelhammer, Judge, and Miller, 2000; Ruud, 2000; Wooldridge, 2002). Its formalization by Hansen (1982) centers on the presence of known functions, labelled "moment functions", of observable random variables and unknown parameters that have expectation zero when evaluated at the true parameter values. The method generalizes the "standard" method of moments where expectations of known functions of observable random variables are equal to known functions of the unknown parameters. The "standard" method of moments can thus be thought of as a special case of the general method with the unknown parameters and observed random variables entering additively separable. The GMM approach links nicely to economic theory where orthogonality conditions that can serve as such moment functions often arise from optimizing behavior of agents. For example, if agents make rational predictions with squared error loss, their prediction errors should be orthogonal to elements of the information set. In the GMM framework the unknown parameters are estimated by setting the sample averages of these moment functions, the "estimating equations," as close to zero as possible.

The framework is sufficiently general to deal with the case where the number of moment functions is equal to the number of unknown parameters, the so–called "just–identified case", as well as the case where the number of moments exceeding the number of parameters to be estimated, the "over–identified case." The latter has special importance in economics where

the moment functions often come from the orthogonality of potentially many elements of the information set and prediction errors. In the just-identified case it is typically possible to estimate the parameter by setting the sample average of the moments exactly equal to zero. In the over-identified case this is not feasible. The solution proposed by Hansen (1982) for this case, following similar approaches in linear models such as two– and three–stage–least–squares, is to set a linear combination of the sample average of the moment functions equal to zero, with the dimension of the linear combination equal to the number of unknown parameters. The optimal linear combination of the moments depends on the unknown parameters, and Hansen suggested to employ initial, possibly inefficient, estimates to estimate this optimal linear combination. Chamberlain (1987) showed that this class of estimators achieves the semiparametric efficient bound given the set of moment restrictions. The Chamberlain paper is not only important for its substantive efficiency result, but also as a precursor to the subsequent empirical likelihood literature by the methods employed: Chamberlain uses a discrete approximation to the joint distribution of all the variables to show that the information matrix based variance bound for the discrete parametrization is equal to the variance of the GMM estimator if the discrete approximation is fine enough.

Th empirical likelihood literature developed partly in response to criticisms regarding the small sample properties of the two-step GMM estimator. Researchers found in a number of studies that with the degree of over-identification high, these estimators had substantial biases, and confidence intervals had poor coverage rates. See among others, Altonji and Segal (1996), Burnside and Eichenbaum (1996) , and Pagan and Robertson (1997). These findings are related to the results in the instrumental variables literature that with many or weak instruments two-stage-least squares can perform very badly (e.g., Bekker, 1994; Bound, Jaeger, and Baker, 1995; Staiger and Stock, 1997). Simulations, as well as theoretical results, suggest that the new estimators have LIML-like properties and lead to improved large sample properties, at the expense of some computational cost.

2. Examples

First the generic form of the GMM estimation problem in a cross–section context is

presented. The parameter vector θ^* is a K dimensional vector, an element of Θ, which is a subset of $\mathbb{R}^K$. The random vector Z has dimension P, with its support $\mathcal{Z}$ a subset of $\mathbb{R}^P$. The moment function, $\psi : \mathcal{Z} \times \Theta \to \mathbb{R}^M$, is a known vector valued function such that $E\left[\psi(Z, \theta^*)\right] = 0$, and $E\left[\psi(Z, \theta)\right] \neq 0$ for all $\theta \in \Theta$ with $\theta \neq \theta^*$. The researcher has available an independent and identically distributed random sample $Z_1, Z_2, \ldots, Z_N$. We are interested in the properties of estimators for θ^* in large samples.

Many, if not most models considered in econometrics fit this framework. Below are some examples, but this list is by no means exhaustive.

I. Maximum Likelihood

If one specifies the conditional distribution of a variable Y given another variable X as $f_{Y|X}(y|x, \theta)$, the score function satisfies these conditions for the moment function:

$$\psi(Y, X, \theta) = \frac{\partial \ln f}{\partial \theta}(Y|X, \theta).$$

By standard likelihood theory the score function has expectation zero only at the true value of the parameter. Interpreting maximum likelihood estimators as generalized method of moments estimators suggests a way of deriving the covariance matrix under misspecification (e.g., White, 1982), as well as an interpretation of the estimand in that case.

II. Linear Instrumental Variables

Suppose one has a linear model

$$Y = X'\theta^* + \varepsilon,$$

with a vector of instruments Z. In that case the moment function is

$$\psi(Y, X, Z, \theta) = Z' \cdot (Y - X'\theta).$$

The validity of Z as an instrument, together with a rank condition implies that θ^* is the unique solution to $E[\psi(Y, X, Z, \theta)] = 0$. This is a case where the fact that the methods allow for more moments than unknown parameters is of great importance as often instruments are independent of structural error terms, implying that any function of the basic instruments is orthogonal to the errors.

III. A Dynamic Panel Data Model

Consider the following panel data model with fixed effects:

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel data models studied in Keane and Runkle (1992), Chamberlain (1992), and Blundell and Bond (1998). This specific model has previously been studied by Bond, Bowsher, and Windmeijer (2001). One can construct moment functions by differencing and using lags as instruments, as in Arellano and Bond (1991), and Ahn and Schmidt, (1995):

$$\psi_{1t}(Y_{i1}, \ldots, Y_{iT}, \theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3, \ldots, T$, leading to a total of $(T-1) \cdot (T-2)/2$ moments, with only a single parameter. One would typically expect that the long lags do not necessarily contain much information, but they are often used to improve efficiency. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1}, \ldots, Y_{iT}, \theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

Despite the different nature of the two sets of moment functions, which makes them potentially very useful in the case that the autoregressive parameter is close to unity, they can all be combined in the GMM framework.

3. Two-step GMM Estimation

3.1 Estimation and Inference

In the just-identified case where M, the dimension of ψ, and K, the dimension of θ are identical, one can generally estimate θ^* by solving

$$0 = \frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \hat{\theta}_{\text{gmm}}). \tag{1}$$

If the sample average is replaced by the expectation, the unique solution is equal to θ^*, and under regularity conditions (e.g., Hansen, 1982, Newey and McFadden, 1994), solutions to (1) will be unique in large samples and consistent for θ^*. If $M > K$ the situation is more complicated as in general there will be no solution to (1).

Hansen's (1982) solution was to generalize the optimization problem to the minimization of the quadratic form

$$Q_{C,N}(\theta) = \frac{1}{N}\left[\sum_{i=1}^{N}\psi(z_i,\theta)\right]' \cdot C \cdot \left[\sum_{i=1}^{N}\psi(z_i,\theta)\right], \tag{2}$$

for some positive definite $M \times M$ symmetric matrix C. Under the regularity conditions given in Hansen (1982) and Newey and McFadden (1994), the minimand $\hat{\theta}_{\rm gmm}$ of (2) has the following large sample properties:

$$\hat{\theta}_{\rm gmm} \xrightarrow{p} \theta^*,$$

$$\sqrt{N}(\hat{\theta}_{\rm gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C \Gamma)^{-1}\Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}),$$

where

$$\Delta = \mathbb{E}\left[\psi(Z_i,\theta^*)\psi(Z_i,\theta^*)'\right] \qquad \text{and} \quad \Gamma = \mathbb{E}\left[\frac{\partial}{\partial\theta'}\psi(Z_i,\theta^*)\right].$$

In the just–identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial, as $\hat{\theta}_{\rm gmm}$ will, at least in large samples, be equal to the value of θ that sets the average moments exactly equal to zero. In that case Γ is a square matrix, and because it is full rank by assumption, Γ is invertible and the asymptotic covariance matrix reduces to $(\Gamma'\Delta^{-1}\Gamma)^{-1}$, irrespective of the choice of C. In the overidentified case with $M > K$, however, the choice of the weight matrix C is important. The optimal choice for C in terms of minimizing the asymptotic variance is in this case the inverse of the covariance of the moments, Δ^{-1}. Using the optimal weight matrix, the asymptotic distribution is

$$\sqrt{N}(\hat{\theta}_{\rm gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1}). \tag{3}$$

This estimator is generally not feasible because typically Δ^{-1} is not known to the researcher. The feasible solution proposed by Hansen (1982) is to obtain an initial consistent, but generally inefficient, estimate of θ^* by minimizing $Q_{C,N}(\theta)$ using an arbitrary positive definite $M \times M$ matrix C, e.g., the identity matrix of dimension M. Given this initial estimate, $\tilde{\theta}$, one can estimate the optimal weight matrix as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N} \sum_{i=1}^{N} \psi(z_i, \tilde{\theta}) \cdot \psi(z_i, \tilde{\theta})' \right]^{-1}.$$

In the second step one estimates θ^* by minimizing $Q_{\hat{\Delta}^{-1},N}(\theta)$. The resulting estimator $\hat{\theta}_{\rm gmm}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Hansen (1982) also suggested a specification test for this model. If the number of moments exceeds the number of free parameters, not all average moments can be set equal to zero, and their deviation from zero forms the basis of Hansen's test, similar to tests developed by Sargan (1958). See also Newey (1985a, 1985b). Formally, the test statistic is

$$T = Q_{\hat{\Delta},N}(\hat{\theta}_{\rm gmm}).$$

Under the null hypothesis that all moments have expectation equal to zero at the true value of the parameter, θ^*, the distribution of the test statistic converges to a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions, $M - K$.

One can also interpret the two–step estimator for over–identified GMM models as a just–identified GMM estimator with an augmented parameter vector (e.g., Newey and McFadden, 1994; Chamberlain and Imbens, 1995). Define the following moment function:

$$h(x,\delta) = h(x,\theta,\Gamma,\Delta,\beta,\Lambda) = \begin{pmatrix} \Lambda - \frac{\partial \psi}{\partial \theta'}(x,\beta) \\ \Lambda' C \psi(x,\beta) \\ \Delta - \psi(x,\beta)\psi(x,\beta)' \\ \Gamma - \frac{\partial \psi}{\partial \theta'}(x,\theta) \\ \Gamma' \Delta^{-1} \psi(x,\theta) \end{pmatrix}. \tag{4}$$

Because the dimension of the moment function $h(\cdot)$, $M\times K+K+(M+1)\times M/2+M\times K+K = (M+1)\times(2K+M/2)$, is equal to the combined dimensions of its parameter arguments, the

estimator for $\delta = (\theta, \Gamma, \Delta, \beta, \Lambda)$ obtained by setting the sample average of $h(\cdot)$ equal to zero is a just–identified GMM estimator. The first two components of $h(x, \delta)$ depend only on β and Λ, and have the same dimension as these parameters. Hence β^* and Λ^* are implicitly defined by the equations

$$E\left[\left(\begin{array}{c} \Lambda - \frac{\partial \psi}{\partial \theta'}(X,\beta) \\ \Lambda' C \psi(X,\beta) \end{array}\right)\right] = 0.$$

Given β^* and Λ^*, Δ^* is defined through the third component of $h(x, \delta)$, and given β^*, Λ^* and Δ^* the final parameters θ^* and Γ^* are defined through the last two moment functions.

This interpretation of the over-identified two-step GMM estimator as a just-identified GMM estimator in an augmented model is interesting because it also emphasizes that results for just–identified GMM estimators such as the validity of the bootstrap can directly be translated into results for over–identified GMM estimators. In another example, using the standard approach to finding the large sample covariance matrix for just–identified GMM estimators one can use the just-identified representation to find the covariance matrix for the over–identified GMM estimator that is robust against misspecification: the appropriate submatrix of

$$\left(E\left[\frac{\partial h}{\partial \delta}(X,\delta^*)\right]\right)^{-1} E[h(Z,\delta^*)h(Z,\delta^*)'] \left(E\left[\frac{\partial h}{\partial \delta}(Z,\delta^*)\right]\right)^{-1},$$

estimated by averaging at the estimated values. This is the GMM analogue of the White (1982) covariance matrix for the maximum likelihood estimator under misspecification.

3.2 Efficiency

Chamberlain (1987) demonstrated that Hansen's (1982) estimator is efficient, not just in the class of estimators based on minimizing the quadratic form $Q_{N,C}(\theta)$, but in the larger class of semiparametric estimators exploiting the full set of moment conditions. What is particularly interesting about this argument is the relation to the subsequent empirical likelihood literature. Many semiparametric efficiency bound arguments (e.g., Newey, 1991; Hahn, 1994) implicitly build fully parametric models that include the semiparametric one and then search for the least favorable parametrization. Chamberlain's argument is qualitatively different.

He proposes a specific parametric model that can be made arbitrarily flexible, and thus arbitrarily close to the model that generated the data, but does not typically include that model. The advantage of the model Chamberlain proposes is that it is in some cases very convenient to work with in the sense that its variance bound can be calculated in a straightforward manner. The specific model assumes that the data are discrete with finite support $\{\lambda_1, \ldots, \lambda_L\}$, and unknown probabilities $\pi_1, \ldots, \pi_L$. The parameters of interest are then implicitly defined as functions of these points of support and probabilities. With only the probabilities unknown, the variance bound on the parameters of the approximating model are conceptually straightforward to calculate. It then sufficies to translate that into a variance bound on the parameters of interest. If the original model is over-identified, one has restrictions on the probabilities. These are again easy to evaluate in terms of their effect on the variance bound.

Given the discrete model it is straightforward to obtain the variance bound for the probabilities, and thus for any function of them. The remarkable point is that one can rewrite these bounds in a way that does not involve the support points. This variance turns out to be identical to the variance of the two-step GMM estimator, thus proving its efficiency.

4. Empirical Likelihood

4.1 Background

To focus ideas, consider a random sample $Z_1, Z_2, \ldots, Z_N$, of size N from some unknown distribution. If we wish to estimate the common distribution of these random variables, the natural choice is the empirical distribution, that puts weight $1/N$ on each of the N sample points. However, in a GMM setting this is not necessarily an appropriate estimate. Suppose the moment function is

$$\psi(z, \theta) = z,$$

implying that the expected value of Z is zero. Note that in this simple example this moment function does not depend on any unknown parameter. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$ as $E_{\hat{F}_{emp}}[Z] = \sum z_i/N \neq 0$.

The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\hat{F}$ does satisfy the restriction. In other words, the approach is to look for the distribution function closest to $\hat{F}_{\rm emp}$, within the set of distribution functions satisfying $E_F[Z] = 0$. Empirical likelihood provides an operationalization of the concept of closeness here. The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i,$$

for $0 \leq \pi_i \leq 1$, $\sum_{i=1}^{N} \pi_i = 1$. This is not a likelihood function in the standard sense, and thus does not have all the properties of likelihood functions. The empirical likelihood estimator for the distribution function is

$$\max_{\pi} \sum_{i=1}^{N} \pi_i \qquad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i z_i = 0.$$

Without the second restriction the π's would be estimated to be $1/N$, but the second restriction forces them slightly away from $1/N$ in a way that ensures the restriction is satisfied. In this example the solution for the Lagrange multiplier is the solution to the equation

$$\sum_{i=1}^{N} \frac{z_i}{1 + t \cdot z_i} = 0,$$

and the solution for π_i is:

$$\hat{\pi}_i = 1/(1 + t \cdot z_i).$$

More generally, in the over-identified case a major focus is on obtaining point estimates through the following estimator for θ:

$$\max_{\theta, \pi} \sum_{i=1}^{N} \ln \pi, \qquad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0. \tag{5}$$

Qin and Lawless (1994) and Imbens (1997) show that this estimator is equivalent, to order $O_p(N^{-1/2})$, to the two-step GMM estimator. This simple discussion illustrates that for some, and in fact many, purposes the empirical likelihood has the same properties as a parametric likelihood function. This idea, first proposed by Owen (1988), turns out to be very powerful

with many applications. Owen (1988) shows how one can construct confidence intervals and hypothesis tests based on this notion.

Related ideas have shown up in a number of places. Cosslett's (1981) work on choice-based sampling can be interpreted as maximizing a likelihood function that is the product of a parametric part coming from the specification of the conditional choice probabilities, and an empirical likelihood function coming from the distribution of the covariates. See Imbens (1992) for a connection between Cosslett's work and two-step GMM estimation. As mentioned before, Chamberlain's (1987) efficiency proof essentially consists of calculating the distribution of the empirical likelihood estimator and showing its equivalence to the distribution of the two-step GMM estimator. See Back and Brown (1990) and Kitamura and Stutzer (1997) for a discussion of the dependent case.

4.2 Cressie-Read Discrepancy Statistics and Generalized Empirical Likelihood

In this section we consider a generalization of the empirical likelihood estimators based on modifications of the objective function. Corcoran (1998) (see also Imbens, Spady and Johnson, 1998), focus on the Cressie-Read discrepancy statistic, for fixed λ, as a function of two vectors p and q of dimension N (Cressie and Read 1984):

$$I_\lambda(p,q) = \frac{1}{\lambda\cdot(1+\lambda)}\sum_{i=1}^N p_i\left[\left(\frac{p_i}{q_i}\right)^\lambda - 1\right].$$

The Cressie-Read minimum discrepancy estimators are based on minimizing this difference between the empirical distribution, that is, the N-dimensional vector with all elements equal to $1/N$, and the estimated probabilities, subject to all the restrictions being satisfied.

$$\min_{\pi,\theta} I_\lambda(\iota/N,\pi) \qquad \text{subject to } \sum_{i=1}^N \pi_i = 1, \quad \text{and } \sum_{i=1}^N \pi_i\cdot\psi(z_i,\theta) = 0.$$

If there are no binding restrictions, because the dimension of $\psi(\cdot)$ and θ agree (the just-identified case), the solution for π is the empirical distribution it self, and $\pi_i = 1/N$. More generally, if there are over-identifying restrictions, there is no solution for θ to $\sum_i \psi(z_i,\theta)/N = 0$, and so the solution for π_i is as close as possible to $1/N$ in a way that ensures there is

an exact solution to $\sum_i \pi_i \psi(z_i, \theta) = 0$. The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Three special cases of this class have received most attention. First, the empirical likelihood estimator itself, which can be interpreted as the case with $\lambda \to 0$. This has the nice interpretation that it is the exact maximum likelihood estimator if Z has a discrete distribution. It does not rely on the discreteness for its general properties, but this interpretation does suggest that it may have attractive large sample properties.

The second case is the exponential tilting estimator with $\lambda \to -1$ (Imbens, Spady and Johnson, 1998), whose objective function is equal to the empirical likelihood objective funtion with the role of π and ι/N reversed. It can also be written as

$$\min_{\pi,\theta} \sum_{i=1}^N \pi_i \ln \pi_i \qquad \text{subject to } \sum_{i=1}^N \pi_i = 1, \quad \text{and } \sum_{i=1}^N \pi_i \psi(z_i, \theta) = 0.$$

Third, the case with $\lambda = -2$. This case was originally proposed by Hansen, Heaton and Yaron (1996) as the solution to

$$\min_\theta \frac{1}{N}\left[\sum_{i=1}^N \psi(z_i,\theta)\right]' \cdot \left[\frac{1}{N}\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\right]^{-1} \cdot \left[\sum_{i=1}^N \psi(z_i,\theta)\right],$$

where the GMM objective function is minimized over the θ in the weight matrix as well as the θ in the average moments. Hansen, Heaton and Yaron (1996) labeled this the continuously updating estimator. Newey and Smith (2004) pointed out that this estimator fits in the Cressie-Read class.

Smith (1997) considers a more general class of estimators, which he labels generalized empirical likelihood estimators, starting from a different perspective. For a given function $g(\cdot)$, normalized so that it satisfied $g(0) = 1$, $g'(0) = 1$, consider the saddle point problem

$$\max_\theta \min_t \sum_{i=1}^N g(t'\psi(z_i, \theta)).$$

This representation is more attractive from a computational perspective, as it reduces the dimension of the optimization problem to $M + K$ rather than a constrained optimization

problem of dimension $K+N$ with $M+1$ restrictions. There is a direct link between the t parameter in the GEL representation and the Lagrange multipliers in the Cressie-Read representation. Newey and Smith (2004) how to choose $g(\cdot)$ for a given λ so that the corresponding GEL and Cressie-Read estimators agree.

In general the differences between the estimators within this class is relatively small compared to the differences between them and the two-step GMM estimators. In practice the choice between them is largely driven by computational issues, which will be discussed in more detail in Section 5. The empirical likelihood estimator does have the advantage of its exact likelihood interpretation and the resulting optimality properties for its bias-corrected version (Newey and Smith, 2004). On the other hand, Imbens, Spady and Johnson (1998) argue in favor of the exponential tilting estimator as its influence function stays bounded where as denominator in the probabilities in the empirical likelihood estimator can get large. In simulations researcher have encountered more convergence problems with the continuously updating estimator (e.g., Hansen, Heaton and Yaron, 1996; Imbens, Johnson and Spady, 1998).

4.3 Testing

Associated with the empirical likelihood estimators are three tests for over-identiyfing restrictions, similar to the classical trinity of tests, the likelihood ratio, the Wald, and the Lagrange multiplier tests. Here we briefly review the implementation of the three tests in the empirical likelihood context. The leading terms of all three tests are identical to that of the test developed by Hansen (1982) based on the quadratic form in the average moments.

The first test is based on the value of the empirical likelihood function. The test statistic compares the value of the empirical likelihood function at the restricted estimates, the $\hat{\pi}_i$ with that at the unrestricted values, $\pi_i = 1/N$:

$$LR = 2 \cdot (L(\iota/N) - L(\hat{\pi})), \qquad \text{where} \quad L(\pi) = \sum_{i=1}^{N} \ln \pi_i.$$

As in the parametric case, the difference between the restricted and unrestricted likelihood function is multiplied by two to obtain, under regularity conditions, e.g., Newey and Smith

(2004), a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions for the test statistic under the null hypothesis.

The second test, similar to Wald tests, is based on the difference between the average moments and their probability limit under the null hypothesis, zero. As in the standard GMM test for overidentifying restrictions (Hansen, 1982), the average moments are weighted by the inverse of their covariance matrix:

$$Wald = \frac{1}{N}\left[\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\right]' \hat{\Delta}^{-1}\left[\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\right],$$

where $\hat{\Delta}$ is an estimate of the covariance matrix

$$\Delta = E[\psi(Z,\theta^*)\psi(Z,\theta)'],$$

typically based on a sample average at some consistent estimator for θ^*:

$$\hat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\psi(z_i,\hat{\theta})',$$

or sometimes a fully efficient estimator for the covariance matrix,

$$\hat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_i\psi(z_i,\hat{\theta})\psi(z_i,\hat{\theta})',$$

The standard GMM test uses an initial estimate of θ^* in the estimation of Δ, but with the empirical likelihood estimators it is more natural to substitute the empirical likelihood estimator itself. The precise properties of the estimator for Δ do not affect the large sample properties of the test, and like the likelihood ratio test, the test statistic has in large samples a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions.

The third test is based on the Lagrange multipliers t. In large samples their variance is

$$V_t = \Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}.$$

This matrix is singular, with rank equal to $M-K$. One option is therefore to compare the Lagrange multipliers to zero using a generalized inverse of their covariance matrix:

$$LM_1 = t'\left(\Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\right)^{-g} t.$$

This is not very attractive, as it requires the choice of a generalized inverse. An alternative is to use the inverse of Δ^{-1} itself, leading to the test statistic

$$LM_2 = t'\Delta t.$$

Because

$$\sqrt{N}\cdot t = V_t \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta^*) + o_p(1),$$

and $V_t\Delta V_t = V_t V_t^{-g} V_t = V_t$, it follows that

$$LM_2 = LM_1 + o_p(1).$$

Imbens, Johnson and Spady (1998) find in their simulations that tests based on LM_2 perform better than those based on LM_1. In large samples both have a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions. Again we can use this test with any efficient estimator for t, and with the Lagrange multipliers based on any of the discrepancy measures.

Imbens, Spady and Johnson (1998), and Bond, Bowsher and Windmeijer (2001) investigate through simulations the small sample properties of various of these tests. It appears that the Lagrange multiplier tests are often more attractive than the tests based on the average moments, although there is so far only limited evidence in specific models. One can use the same ideas for constructing confidence intervals that do not directly use the normal approximation to the sampling distribution of the estimator. See for discussions Smith (1998) and Imbens and Spady (2002).

6. Computational Issues

The two-step GMM estimator requires two minimizations over a K-dimensional space. The empirical likelihood estimator in its original likelihood form (5) requires maximization over a space of dimension K (for the parameter θ) plus N (for the N probabilities), subject to $M+1$ restrictions (on the M moments and the adding up restriction for the probabilities). This is in general a much more formidable computational problem than two optimizations

in a K-dimensional space. A number of approaches have been attempted to simplify this problem. Here we disuss three of them in the context of the exponential tilting estimator, although most of them directly carry over to other members of the Cressie-Read or GEL classes.

6.1 Solving the First Order Conditions

The first approach we discuss is focuses on the first order conditions and then concentrates out the probabilities π. This reduces the problem to one of dimension $K+M$, K for the parameters of interest and M for the Lagrange multipliers for the restrictions, which is is clearly a huge improvement, as the dimension of the problem no longer increases with the sample size. Let μ and t be the Lagrange multipliers for the restrictions $\sum \pi_i = 1$ and $\sum \pi_i \psi(z_i,\theta) = 0$. The first order conditions for the π's and θ and the Lagrange multipliers are

$$
\begin{aligned}
0 &= \ln \pi_i - 1 - \mu + t'\psi(z_i,\theta), \\
0 &= \sum_{i=1}^N \pi_i \frac{\partial \psi}{\partial \theta'}(z_i,\theta), \\
0 &= \exp(\mu - 1) \sum_{i=1}^N \exp\left(t'\psi(z_i,\theta)\right), \\
0 &= \exp(\mu - 1) \sum_{i=1}^N \psi(z_i,\theta) \cdot \exp\left(t'\psi(z_i,\theta)\right).
\end{aligned}
$$

The solution for π is

$$
\pi_i = \exp(\mu - 1 + t'\psi(z_i,\theta)).
$$

To determine the Lagrange multipliers t and the parameter of interest θ we only need π_i up to a constant of proportionality, so we can solve

$$
0 = \sum_{i=1}^N \psi(z_i,\theta) \exp(t'\psi(z_i,\theta)), \tag{6}
$$

and

$$
0 = \sum_{i=1}^N t' \frac{\partial \psi}{\partial \theta}(z_i,\theta) \exp(t'\psi(z_i,\theta)) \tag{7}
$$

Solving the system of equations (6) and (7) is not straightforward. Because the probability limit of the solution for t is zero, the derivative with respect to θ of both first order conditions converges zero. Hence the matrix of derivatives of the first order conditions converges to a singular matrix. As a result standard approaches to solving systems of equations can behave erratically, and this approach to calculating $\hat{\theta}$ has been found to have poor operating characteristics.

6.2 Penalty Function Approaches

Imbens, Spady and Johnson (1998) characterize the solution for θ and t as

$$\max_{\theta,t} K(t,\theta) \qquad \text{subject to } K_t(t,\theta)=0, \tag{8}$$

where $K(t,\theta)$ is the empirical analogue of the cumulant generating function:

$$K(t,\theta)=\ln\left[\frac{1}{N}\sum_{i=1}^{N}\exp(t'\psi(z_i,\theta)\right].$$

They suggest solving this optimization problem by maximizing the unconstrained objective function with a penalty term that consists of a quadratic form in the restriction:

$$\max_{\theta,t} K(t,\theta)-0.5\cdot A\cdot K_t(t,\theta)'W^{-1}K_t(t,\theta), \tag{9}$$

for some positive definite $M\times M$ matrix W, and a positive constant A. The first order conditions for this problem are

$$0=K_\theta(t,\theta)-A\cdot K_{t\theta}(t,\theta)W^{-1}K_t(t,\theta),$$

$$0=K_t(t,\theta)-A\cdot K_{tt}(t,\theta)W^{-1}K_t(t,\theta).$$

For A large enough the solution to this unconstrained maximization problem is identical to the solution to the constrained maximization problem (8). This follows from the fact that the constraint is in fact the first order condition for $K(t,\theta)$. Thus, in contrast to many penalty function approaches, one does not have to let the penalty term go to infinity to obtain the solution to the constrained optimization problem, one only needs to let the penalty term

increase sufficiently to make the problem locally convex. Imbens, Spady and Johnson (1998) suggest choosing

$$W = K_{tt}(t,\theta) + K_t(t,\theta)K_t(t,\theta)',$$

for some initial values for t and θ as the weight matrix, and report that estimates are generally not sensitive to the choices of t and θ.

6.3 Concentrating out the Lagrange Multipliers

Mittelhammer, Judge and Schoenberg (2001) suggest concentrating out both probabilities and Lagrange multipliers and then maximizing over θ without any constraints. As shown above, concentrating out the probabilities π_i can be done analytically. Although it is not in general possible to solve for the Lagrange multipliers t analytically, other than in the continuously updating case, for given θ it is easy to numerically solve for t. This involves solving, in the exponential tilting case,

$$\min_t \sum_{i=1}^N \exp(t'\psi(z_i,\theta)).$$

This function is strictly convex as a function of t, with the easy to calculate first and second derivatives equal to

$$\sum_{i=1}^N \psi(z_i,\theta)\exp(t'\psi(z_i,\theta)),$$

and

$$\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\exp(t'\psi(z_i,\theta)),$$

respectively. Therefore concentrating out the Lagrange multipliers is computationally fast using a Newton-Raphson algorithm. The resulting function $t(\theta)$ has derivatives with respect to θ equal to:

$$\frac{\partial t}{\partial \theta'}(\theta) = -\left(\frac{1}{N}\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\exp(t(\theta)'\psi(z_i,\theta)\right)^{-1}$$

$$\cdot \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \psi}{\partial \theta'}(z_i, \theta) \exp(t(\theta)'\psi(z_i, \theta) + \psi(z_i, \theta) t(\theta)' \frac{\partial \psi}{\partial \theta'}(z_i, \theta) \exp(t(\theta)'\psi(z_i, \theta)) \right)$$

After solving for $t(\theta)$, one can solve

$$\max_{\theta} \sum_{i=1}^{N} \exp(t(\theta)'\psi(z_i, \theta)). \tag{10}$$

Mittelhammer, Judge, and Schoenberg (2001) use methods that do not require first derivatives to solve (10). This is not essential. Calculating first derivatives of the concentrated objective function only requires first derivatives of the moment functions, both directly and indirectly through the derivatives of $t(\theta)$ with respect to θ. In general these are straightforward to calculate and likely to improve the performance of the algorithm.

In this method in the end the researcher only has to solve one optimization in a K-dimensional space, with the provision that for each evaluation of the objective function one needs to numerically evaluate the function $t(\theta)$ by solving a convex maximization problem. The latter is fast, especially in the exponential tilting case, so that although the resulting optimization problem is arguably still more difficult than the standard two-step GMM problem, in practice it is not much slower. In the simulations below I use this method for calculating the estimates. After concentrating out the Lagrange multipliers using a Newton-Rahpson algorithm that uses both first and second derivatives, I use a Davidon-Fletcher-Powell algorithm to maximize over θ, using analytic first derivatives. Given a direction I used a line search algoritm based on repeated quadratic approximations.

7. A Dynamic Panel Data Model

To get a sense of the finite sample properties of the empirical likelihood estimators we compare some of the GMM methods in the context of the panel data model briefly discussed in Section 2, using some simulation results from Imbens. The model is

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel

data models extensively studied in the literature. Bond, Bowsher and Windmeijer (2001) study this and similar models to evaluate the performance of test statistics based on different GMM and gel estimators. We use the moments

$$\psi_{1t}(Y_{i1},\ldots,Y_{iT},\theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3,\ldots,T$, leading to a total of $(T-1)\cdot(T-2)/2$ moments. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1},\ldots,Y_{iT},\theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

It is important to note, given the results discussed in Section 4, that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. As a result there is potentially a substantial difference between the different estimators, especially when the degree of overidentification is high.

We report some simulations for a data generating process with parameter values estimated on data from Abowd and Card (1989) taken from the PSID. See also Card (1994). This data set contains earnings data for 1434 individuals for 11 years. The individuals are selected on having positive earnings in each of the eleven years, and we model their earnings in logarithms. We focus on estimation of the autoregressive coefficient θ.

We then generate artificial data sets to investigate the repeated sampling properties of these estimators. Two questions are of most interest. First, how do the median bias and median-absolute-error deteriorate as a function of the degree of over-identification? Here, unlike in the theoretical discussion in Section 4, the additional moments, as we increase the number of years in the panel, do contain information, so they may in fact increase precision, but at the same time one would expect based on the theoretical calculations that the accuracy of the asymptotic approximations for a fixed sample size deteriorates with the

number of years. Second, we are interested in the performance of the confidence intervals for the parameter of interest. In two-stage-least-squares settings it was found that with many weak instruments the performance of standard confidence intervals varied widely between liml and two-stage-least-squares estimators. Given the analogy drawn by Hansen, Heaton and Yaron (1996) between the continuously updating estimator and liml, the question arises how the confidence intervals differ between two-step GMM and the various Cressie-Read and GEL estimators.

Using the Abowd-Card data we estimate θ and the variance of the fixed effect and the idiosyncratic error term. The latter two are estimated to be around 0.3. We then consider data generating processes where the individual effect η_i has mean zero and standard deviation equal to 0.3, and the error term has mean zero and standard deviation 0.3. We $\theta = 0.9$ in the simulations. This is larger than the value in estimated from the Abowd-Card data. We compare the standard Two-Step GMM estimator and the Exponential Tilting Estimator. Table 1 contains the results. With the high autoregressive coefficient, $\theta = 0.9$, the two-step GMM estimator has substantial bias and poor coverage rates.The exponential tilting estimator does much better with the high autoregressive coefficient. The bias is small, on the order of 10% of the standard error, and the coverage rate is much closer to the nominal one.

References

Abowd, J. and D. Card, (1989), "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57 (2), 441-445.

Ahn, S., and P. Schmidt, (1995), "Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, 68, 5-28.

Altonji, J., and L. Segal, (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, Vol 14, No. 3, 353–366.

Back, K., and D. Brown, (1990), "Estimating Distributions from Moment Restrictions", working paper, Graduate School of Business, Indiana University.

Bekker, P., (1994), "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 62, 657-681.

Bond, S., C. Bowsher, and F. Windmeijer, (2001), "Criterion-based Inference for GMM in Linear Dynamic Panel Data Models", IFS, London.

Bound, J., D. Jaeger, and R. Baker, (1995), "Problems with Instrumental Variables Estimation when the Correlation between Instruments and the Endogenous Explanatory Variable is Weak", forthcoming, em Journal of the American Statistical Association.

Burnside, C., and M., Eichenbaum, (1996), "Small Sample Properties of Generalized Method of Moments Based Wald Tests", *Journal of Business and Economic Statistics*, Vol. 14, 294-308.

Card, D., (1994) "Intertemporal Labour Supply: an Assessment", in: *Advances in Econometrics*, Simms (ed), Cambridge University Press.

Chamberlain, G., (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, vol. 34, 305–334, 1987

Corcoran, S., (1998), "Bartlett Adjustment of Empirical Discrepancy Statistics", *Biometrika*.

Cosslett, S. R., (1981), "Maximum Likelihood Estimation for Choice–based Samples", *Econometrica*, vol. 49, 1289–1316,

Cressie, N., and T. Read, (1984), "Multinomial Goodness-of-Fit Tests", *Journal of the Royal Statistical Society*, Series B, 46, 440-464.

Hall, A., (2005), *Generalized Method of Moments*, Oxford University Press.

Hansen, L–P., (1982), "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, vol. 50, 1029–1054.

Hansen, L.–P., J. Heaton, and A. Yaron, (1996), "Finite Sample Properties of Some Alternative GMM Estimators", *Journal of Business and Economic Statistics*, Vol 14, No. 3, 262–280.

Imbens, G. W., (1992), "Generalized Method of Moments and Empirical Likelihood," *Journal of Business and Economic Statistics*, vol. 60.

Imbens, G. W., R. H. Spady, and P. Johnson, (1998), "Information Theoretic Approaches to Inference in Moment Condition Models", *Econometrica*.

Imbens, G., and R. Spady, (2002), "Confidence Intervals in Generalized Method of Moments Models," *Journal of Econometrics*, 107, 87-98.

Imbens, G., and R. Spady, (2005), "The Performance of Empirical Likelihood and Its Generalizations," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Andrews and Stock (eds).

Kitamura, Y., and M. Stutzer, (1997), "An Information-theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica*, Vol. 65, 861-874.

Mittelhammer, R., G. Judge, and R. Schoenberg, (2005), "Empirical Evidence Concerning the Finite Sample Performance of EL-Type Structural Equation Estimation and Inference Methods," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Andrews and Stock (eds).

Mittelhammer, R., G. Judge, and D. Miller, (2000), *Econometric Foundations*, Cambridge University Press, Cambridge.

Newey, W., (1985), "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, vol. 29, 229–56.

Newey, W., and D. McFadden, (1994) "Estimation in Large Samples", in: McFadden and Engle (Eds.), *The Handbook of Econometrics*, Vol. 4.

Newey, W., and R. Smith, (2004), "Higher Order Properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 573-595.

Owen, A., (1988), "Empirical Likelihood Ratios Confidence Intervals for a Single Functional", *Biometrika*, 75, 237-249.

Owen, A., (2001), *Empirical Likelihood*, Chapman and Hall, London.

Pagan, A., and J. Robertson, (1997), "GMM and its Problems", unpublished manuscript, Australian National University.

Qin, and J. Lawless, (1994), "Generalized Estimating Equations", *Annals of Stat.*

Smith, R., (1997), "Alternative Semiparametric Likelihood Approaches to Generalized Method of Moments Estimation", *Economic Journal*, 107, 503-19.

Staiger, D., and J. Stock, (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557-586.

White, H., (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, vol. 50, 1–25.

Wooldridge, J., (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples", *Econometrica* 67, No. 6, 1385-1406.

Table 1: SIMULATIONS, $\theta = 0.9$

	Number of time periods								
	3	4	5	6	7	8	9	10	11
Two-Step GMM									
median bias	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
relative median bias	-0.02	0.08	0.03	0.08	0.03	0.11	0.08	0.13	0.11
median absolute error	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.88	0.85	0.82	0.80	0.80	0.79	0.78	0.79	0.76
covarage rate 95% ci	0.92	0.91	0.89	0.87	0.85	0.86	0.86	0.88	0.84
Exponential Tilting									
median bias	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00
relative median bias	0.04	0.09	0.02	-0.00	0.01	0.01	-0.02	0.08	0.13
median absolute error	0.05	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.87	0.86	0.84	0.86	0.88	0.86	0.87	0.88	0.87
covarage rate 95% ci	0.91	0.90	0.90	0.91	0.93	0.92	0.91	0.93	0.93

The relative median bias reports the bias divided by the large sample standard error. All results based on 10,000 replications.