"What's New in Econometrics"

Lecture 1

Estimation of Average Treatment Effects

Under Unconfoundedness

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction
2. Potential Outcomes
3. Estimands and Identification
4. Estimation and Inference
5. Assessing Unconfoundedness (not testable)
6. Overlap
7. Illustration based on Lalonde Data

1. Introduction

We are interested in estimating the average effect of a program or treatment, allowing for heterogenous effects, assuming that selection can be taken care of by adjusting for differences in observed covariates.

This setting is of great applied interest.

Long literature, in both statistics and economics. Influential economics/econometrics papers include Ashenfelter and Card (1985), Barnow, Cain and Goldberger (1980), Card and Sullivan (1988), Dehejia and Wahba (1999), Hahn (1998), Heckman and Hotz (1989), Heckman and Robb (1985), Lalonde (1986). In stat literature work by Rubin (1974, 1978), Rosenbaum and Rubin (1983).

Unusual case with many proposed (semi-parametric) estimators (matching, regression, propensity score, or combinations), many of which are actually used in practice.

We discuss implementation, and assessment of the critical assumptions (even if they are not testable).

In practice concern with overlap in covariate distributions tends to be important.

Once overlap issues are addressed, choice of estimators is less important. Estimators combining matching and regression or weighting and regression are recommended for robustness reasons.

Key role for analysis of the joint distribution of treatment indicator and covariates prior to using outcome data.

2. Potential Outcomes (Rubin, 1974)

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
$Y_i(1) - Y_i(0)$ is unit-level causal effect

Covariates X_i (not affected by treatment)
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 observe for each unit the triple (W_i, Y_i, X_i), where Y_i is the realized outcome:

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

Several additional pieces of notation.

First, the propensity score (Rosenbaum and Rubin, 1983) is defined as the conditional probability of receiving the treatment,

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

Also the two conditional regression and variance functions:

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

3. Estimands and Identification

Population average treatments

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

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 the sample average treatment effect (SATE):

$$\tau_S = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

τ_P versus τ_S does not matter for estimation, but matters for variance.

4. Estimation and Inference

Assumption 1 (Unconfoundedness, Rosenbaum and Rubin, 1983a)

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

"conditional independence assumption," "selection on observables." In missing data literature "missing at random."

To see the link with standard exogeneity assumptions, assume constant effect and linear regression:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i, \quad \Longrightarrow \quad Y_i = \alpha + \tau \cdot W_i + X_i'\beta + \varepsilon_i$$

with $\varepsilon_i \perp\!\!\!\perp X_i$. 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.

Motivation for Unconfoundeness Assumption (I)

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.

Motivation for Unconfoundeness Assumption (II)

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.

It is clear that settings where some of necessary covariates are not observed will require strong assumptions to allow for identification. E.g., instrumental variables settings Absent those assumptions, typically only bounds can be identified (e.g., Manski, 1990, 1995).

Motivation for Unconfoundeness Assumption (III)

Example of a model that is consistent with unconfoundedness: suppose we are interested in estimating the average effect of a binary input on a firm's output, or $Y_i = g(W, \varepsilon_i)$.

Suppose that profits are 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 output given inputs, then unconfoundedness will hold as

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

Overlap

Second assumption on 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."

Identification Given Assumptions

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$$

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

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

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 overlap is important.

Given identification of $\tau(x)$,

$$\tau_P = \mathbb{E}[\tau(X_i)]$$

Alternative Assumptions

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

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

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.

If we are interested in $\tau_{P,T}$ it is sufficient to assume

$$Y_i(0) \perp\!\!\!\perp W_i \mid X_i,$$

Propensity Score

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

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

Only need to condition on scalar function of covariates, which would be much easier in practice if X_i is high-dimensional.

(Problem is that the propensity score $e(x)$ is almost never known.)

Efficiency Bound

Hahn (1998): for any regular estimator for τ_P, denoted by $\widehat{\tau}$, with

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

the variance must satisfy:

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

Estimators exist that achieve this bound.

Estimators

A. Regression Estimators

B. Matching

C. Propensity Score Estimators

D. Mixed Estimators **(recommended)**

A. Regression Estimators

Estimate $\mu_w(x)$ consistently and estimate τ_P or τ_S as

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

Simple implementations include

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

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

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

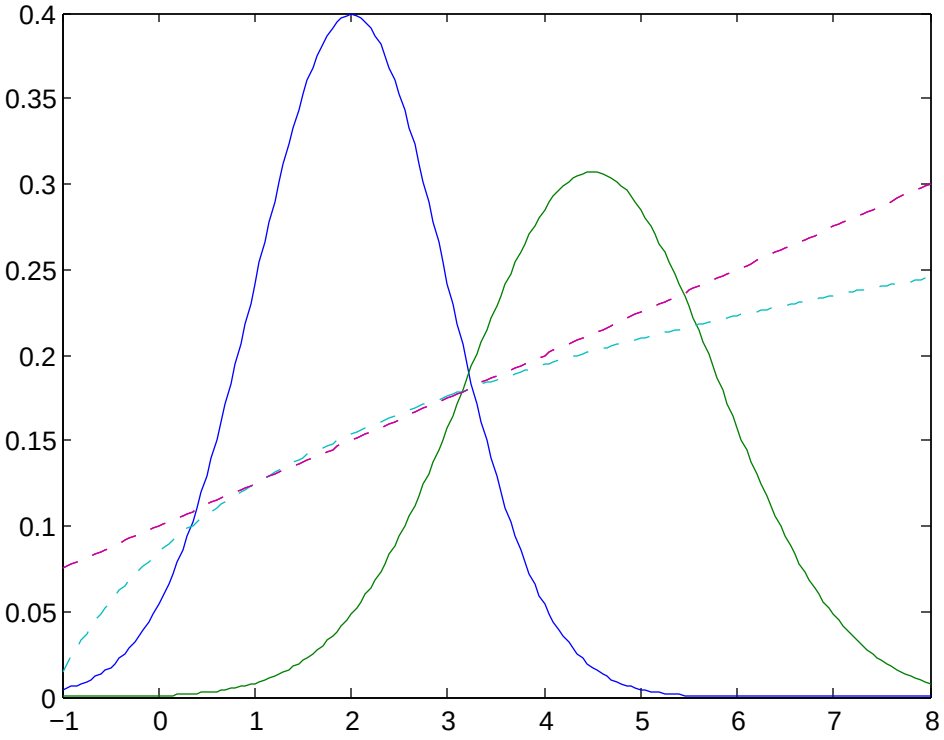
More generally, one can specify separate regression functions for the two regimes, $\mu_w(x) = \beta_w' x$.

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.

Note that $\mu_0(x)$ 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 $\bar{Y}_C + \widehat{\beta}'(\overline{X}_T - \overline{X}_C)$.

If $\overline{X}_T$ and $\overline{X}_C$ are close, the precise specification of the regression function will not matter much for the average prediction. With the two averages very different, the prediction based on a linear regression function can be sensitive to changes in the specification.


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−1
0
1
2
3
4
5
6
7
8

B. Matching

let $\ell_m(i)$ is the mth closest match, that is, the index l that satisfies $W_l \neq W_i$ and

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

Then

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

The simple matching estimator is

$$\widehat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right). \tag{2}$$

Issues with Matching

Bias is of order $O(N^{-1/K})$, where K is dimension of covariates. Is important in large samples if $K \geq 2$ (and dominates variance asymptotically if $K \geq 3$)

Not Efficient (but efficiency loss is small)

Easy to implement, robust.

C.1 Propensity Score Estimators: Weighting

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

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

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i\cdot Y_i}{e(X_i)} - \frac{(1-W_i)\cdot Y_i}{1-e(X_i)}\right). \tag{3}$$

Implementation of Horvitz-Thompson Estimator

Estimate $e(x)$ flexibly (Hirano, Imbens and Ridder, 2003)

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

$$- \sum_{i=1}^{N} \frac{(1-W_i) \cdot Y_i}{1-\widehat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{(1-W_i)}{1-\widehat{e}(X_i)}$$

Is efficient given nonparametric estimator for $e(x)$.

Potentially sensitive to estimator for propensity score.

Matching or Regression on the Propensity Score

Not clear what advantages are.

Large sample properties not known.

Simulation results not encouraging.

D.1 Mixed Estimators: Weighting and Regression

Interpret Horvitz-Thompson estimator as weighted regression estimator:

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i, \quad \text{with weights } \lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1 - W_i}{1 - e(X_i)}}.$$

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 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.

Matching and Regression

First match observations.

Define

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

Then adjust within pair difference for the within-pair difference in covariates $\widehat{X}_i(1) - \widehat{X}_i(0)$:

$$\widehat{\tau}_M^{adj} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\beta} \cdot \left(\widehat{X}_i(1) - \widehat{X}_i(0) \right) \right),$$

using regression estimate for β.

Can eliminate bias of matching estimator given flexible specification of regression function.

Estimation of the Variance

For efficient estimator of τ_P:

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

Estimate all components nonparametrically, and plug in.

Alternatively, use bootstrap.

(Does not work for matching estimator)

Estimation of the Variance

For all estimators of τ_S, for some known $\lambda_i(\mathbf{X}, \mathbf{W})$

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

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

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.

The sample variance of the outcome variable for these 2 units can then be used to estimate $\sigma^2_{W_i}(X_i)$:

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

5.I Assessing Unconfoundedness: Multiple Control Groups

Suppose we have 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

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

Suppose we extend the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates,

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

Now 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.

5.II Assessing Unconfoundedness: Estimate Effects on Pseudo Outcomes

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 $X_i = (Y_{i,-1}, \ldots, Y_{i,-T}, Z_i)$.

Now consider the following two assumptions. The first is unconfoundedness given only $T-1$ lags of the outcome:

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

and the second assumes stationarity and exchangeability: Then it follows that

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

which is testable.

6.I Assessing Overlap

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.

6.II Selecting a Subsample with Overlap

Define average effects for subsamples $\mathbb{A}$:

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

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 $q(\mathbb{A}) = \mathsf{Pr}(X \in \mathbb{A})$.

They derive the characterization for the set $\mathbb{A}$ that minimizes the asymptotic variance .

The optimal set 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

$$\frac{1}{\alpha \cdot (1-\alpha)} =$$

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

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. Applic. to Lalonde Data (Dehejia-Wahba Sample)

	Controls (N=260)		Trainees (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	diff / sd	mean	(s.d.)	diff / s
Age	25.1	7.06	25.8	7.16	0.1	33.2	11.1	-0.7
Black	0.83	0.38	0.84	0.36	0.0	0.07	0.26	2.8
Ed	10.1	1.61	10.4	2.01	0.1	12.0	2.87	-0.6
Hisp	0.11	0.31	0.06	0.24	-0.2	0.07	0.26	-0.1
Marr	0.15	0.36	0.19	0.39	0.1	0.71	0.45	-1.2
E '74	2.11	5.69	2.10	4.89	-0.0	14.0	9.57	-1.2
E '75	1.27	3.10	1.53	3.22	0.1	0.12	0.32	1.8
U '74	0.75	0.43	0.71	0.46	-0.1	13.7	9.27	-1.3
U '75	0.68	0.47	0.60	0.49	-0.2	0.11	0.31	1.5

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

	Experimental Controls			CPS Comparison Grou		
	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
P Score Weighting	0.15	0.30	0.5	-1.17	0.26	-4.5
P Score Blocking	0.10	0.17	0.6	-2.80	0.56	-5.0
P Score Regression	0.16	0.30	0.5	-1.68	0.79	-2.1
P 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 Regr	0.15	0.21	0.7	-1.23	0.24	-5.2
Blocking and Regr	0.09	0.15	0.6	-1.30	0.50	-2.6
Matching and Regr	0.06	0.28	0.2	-1.34	0.42	-3.2

Table 3: Sample Sizes for CPS Sample

	$\widehat{e}(X_i) < 0.1$	$0.1 \leq \widehat{e}(X_i) \leq 0.9$	$0.9 < \widehat{e}(X_i)$	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 may obtain more precise estimates for the remaining sample.

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

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
P Score Weighting	-0.16	0.15	-1.0	1.86	0.75	2.5
P Score Blocking	-0.25	0.25	-1.0	1.73	1.23	1.4
P Score Regression	-0.07	0.17	-0.4	2.09	0.73	2.9
P 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 Regr	-0.14	0.14	-1.1	1.96	0.77	2.5
Blocking and Regr	-0.25	0.25	-1.0	1.73	1.22	1.4
Matching and Regr	-0.11	0.19	-0.6	2.23	1.16	1.9

What's New in Econometrics?
Lecture 2
Linear Panel Data Models

Jeff Wooldridge
NBER Summer Institute, 2007

1. Overview of the Basic Model
2. New Insights Into Old Estimators
3. Behavior of Estimators without Strict Exogeneity
4. IV Estimation under Sequential Exogeneity
5. Pseudo Panels from Pooled Cross Sections

1. Overview of the Basic Model

- Unless stated otherwise, the methods discussed in these slides are for the case with a large cross section and small time series.
- For a generic i in the population,

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

where η_t is a separate time period intercept, $\mathbf{x}_{it}$ 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. We view the c_i as random draws along with the observed variables.

- An attractive assumption is *contemporaneous exogeneity conditional on* c_i :

$$E(u_{it}|\mathbf{x}_{it}, c_i) = 0, t = 1, \ldots, T. \tag{2}$$

This equation defines $\boldsymbol{\beta}$ in the sense that under (1) and (2),

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i, \tag{3}$$

so the β_j are partial effects holding c_i fixed.

- Unfortunately, $\boldsymbol{\beta}$ is not identified only under (2). If we add the strong assumption $Cov(\mathbf{x}_{it}, c_i) = \mathbf{0}$, then $\boldsymbol{\beta}$ is identified.
- Allow any correlation between $\mathbf{x}_{it}$ and c_i by assuming *strict exogeneity conditional on* c_i :

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = 0, t = 1, \ldots, T, \tag{4}$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_i, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \tag{5}$$

If $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ has 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.

- Make inference fully robust to heteroksedasticity and serial dependence, even if use GLS. With large N and small T, there is little excuse not to compute "cluster" standard errors.
- Violation of strict exogeneity: always if $\mathbf{x}_{it}$ contains lagged dependent variables, but also if changes in u_{it} cause changes in $\mathbf{x}_{i,t+1}$ ("feedback effect").
- *Sequential exogeneity condition on c_i*:

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it}, c_i) = 0, t = 1, \ldots, T \qquad (6)$$

or, maintaining the linear model,

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i). \qquad (7)$$

Allows for lagged dependent variables and other

feedback. Generally, $\boldsymbol{\beta}$ is identified under sequential exogeneity. (More later.)

- The key "random effects" assumption is

$$E(c_i|\mathbf{x}_i) = E(c_i). \tag{8}$$

Pooled OLS or any GLS procedure, including the RE estimator, are consistent. Fully robust inference is available for both.

- It is useful to define two *correlated random effects* assumptions. The first just defines a linear projection:

$$L(c_i|\mathbf{x}_i) = \psi + \mathbf{x}_i\boldsymbol{\xi}, \tag{9}$$

Called the *Chamberlain device*, after Chamberlain (1982). Mundlak (1978) used a restricted version

$$E(c_i|\mathbf{x}_i) = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \tag{10}$$

where $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$. Then

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it}, \qquad (11)$$

and we can apple pooled OLS or RE because $E(a_i + u_{it}|\mathbf{x}_i) = 0$. Both equal the FE estimator of $\boldsymbol{\beta}$.

- Equation (11) makes it easy to compute a fully robust Hausman test comparing RE and FE. Separate covariates into aggregate time effects, time-constant variables, and variables that change across i and t:

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it}. \qquad (12)$$

We cannot estimate $\boldsymbol{\gamma}$ by FE, so it is not part of the Hausman test comparing RE and FE. Less clear is that coefficients on the time dummies, $\boldsymbol{\eta}$, cannot be included, either. (RE and FE estimation only with aggregate time effects are identical.) We can only

compare $\hat{\boldsymbol{\delta}}_{FE}$ and $\hat{\boldsymbol{\delta}}_{RE}$ (M parameters).

- Convenient test:

$$y_{it} \text{ on } \mathbf{g}_t, \mathbf{z}_i, \mathbf{w}_{it}, \bar{\mathbf{w}}_i, \; t = 1, \ldots, T; \; i = 1, \ldots, N, \tag{13}$$

which makes it clear there are M restrictions to test. Pooled OLS or RE, fully robust!

- Must be cautious using canned procedures, as the df are often wrong and tests nonrobust.

2. New Insights Into Old Estimators

- Consider an extension of the usual model to allow for unit-specific slopes,

$$y_{it} = c_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} \tag{14}$$

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{b}_i) = 0, t = 1, \ldots, T, \tag{15}$$

where $\mathbf{b}_i$ is $K \times 1$. We act as if $\mathbf{b}_i$ is constant for all i but think c_i might be correlated with $\mathbf{x}_{it}$; we apply usual FE estimator. When does the usual FE

estimator consistently estimate the population average effect, $\boldsymbol{\beta} = \mathrm{E}(\mathbf{b}_i)$?

- A sufficient condition for consistency of the FE estimator, along with along with (15) and the usual rank condition, is

$$\mathrm{E}(\mathbf{b}_i | \ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1, \ldots, T \tag{16}$$

where $\ddot{\mathbf{x}}_{it}$ are the time-demeaned covariates. Allows the slopes, $\mathbf{b}_i$, to be correlated with the regressors $\mathbf{x}_{it}$ through permanent components. For example, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1, \ldots, T$. Then (16) holds if $\mathrm{E}(\mathbf{b}_i | \mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iT}) = \mathrm{E}(\mathbf{b}_i)$.

- Extends to a more general class of estimators. Write

$$y_{it} = \mathbf{w}_t \mathbf{a}_i + \mathbf{x}_{it} \mathbf{b}_i + u_{it}, \quad t = 1, \ldots, T \tag{17}$$

where $\mathbf{w}_t$ is a set of deterministic functions of time.

FE now sweeps away $\mathbf{a}_i$ by netting out $\mathbf{w}_t$ from $\mathbf{x}_{it}$.

- In the random trend model, $\mathbf{w}_t = (1, t)$. If $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, then $\mathbf{b}_i$ can be arbitrarily correlated with $(\mathbf{f}_i, \mathbf{h}_i)$.
- Generally, need $\dim(\mathbf{w}_t) < T$)
- Can apply to models with time-varying factor loads, η_t :

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t c_i + u_{it}, t = 1, \ldots, T. \tag{18}$$

Sufficient for consistency of FE estimator that ignores the η_t is

$$\text{Cov}(\ddot{\mathbf{x}}_{it}, c_i) = \mathbf{0}, t = 1, \ldots, T. \tag{19}$$

- Now let some elements of $\mathbf{x}_{it}$ be correlated with $\{u_{ir} : r = 1, \ldots, T\}$, but with strictly exogenous instruments (conditional on c_i). Assume

$$\text{E}(u_{it}|\mathbf{z}_i, \mathbf{a}_i, \mathbf{b}_i) = 0 \tag{20}$$

for all t. Also, replace (16) with

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1, \ldots, T. \tag{21}$$

Still not enough. A sufficient condition is

$$\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i|\ddot{\mathbf{z}}_{it}) = \mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i), t = 1, \ldots, T. \tag{22}$$

$\mathrm{Cov}(\ddot{\mathbf{x}}_{it}, \mathbf{b}_i)$, a $K \times K$ matrix, need not be zero, or even constant across time. The *conditional* covariance cannot depend on the time-demeaned instruments. Then, FEIV is consistent for $\boldsymbol{\beta} = E(\mathbf{b}_i)$ provided a full set of time dummies is included.

- Assumption (22) cannot be expected to hold when endogenous elements of $\mathbf{x}_{it}$ are discrete.

3. Behavior of Estimators without Strict Exogeneity

- Both the FE and FD estimators are inconsistent (with fixed T, $N \to \infty$) without the conditional strict

exogeneity assumption. Under certain assumptions, the FE estimator can be expected to have less "bias" (actually, inconsistency) for larger T.

- If we maintain $E(u_{it}|\mathbf{x}_{it}, c_i) = 0$ and assume $\{(\mathbf{x}_{it}, u_{it}) : t = 1, \ldots, T\}$ is "weakly dependent", can show

$$\text{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_{FE} = \boldsymbol{\beta} + O(T^{-1}) \tag{23}$$

$$\text{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_{FD} = \boldsymbol{\beta} + O(1). \tag{24}$$

- Interestingly, still holds if $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ has unit roots as long as $\{u_{it}\}$ is I(0) and contemporaneous exogeneity holds.
- Catch: if $\{u_{it}\}$ is I(1) – so that the time series "model" is a spurious regression (y_{it} and $\mathbf{x}_{it}$ are not *cointegrated*), then (23) is no longer true. FD eliminates any unit roots.
- Same conclusions hold for IV versions: FE has

bias of order T^{-1} if $\{u_{it}\}$ is weakly dependent.

- Simple test for lack of strict exogeneity in covariates:

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{i,t+1}\boldsymbol{\delta} + c_i + e_{it} \qquad (25)$$

Estimate the equation by fixed effects and test $H_0 : \boldsymbol{\delta} = \mathbf{0}$.

- Easy to test for contemporaneous endogeneity of certain regressors. Write the model now as

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

where, in an FE environment, we want to test $\mathrm{H}_0 : \mathrm{E}(\mathbf{y}'_{it3}u_{it1}) = \mathbf{0}$. Write a set of reduced forms for elements of $\mathbf{y}_{it3}$ as

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

and obtain the FE residuals, $\hat{\mathbf{v}}_{it3} = \mathbf{y}_{it3} - \mathbf{z}_{it}\hat{\boldsymbol{\Pi}}_3$, where the columns of $\hat{\boldsymbol{\Pi}}_3$ are the FE estimates.

Then, estimate

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

by FEIV, using instruments $(\mathbf{z}_{it}, \mathbf{y}_{it3}, \hat{\mathbf{v}}_{it3})$. The test that $\mathbf{y}_{it3}$ is exogenous is just the (robust) test that $\boldsymbol{\rho}_1 = \mathbf{0}$, and the test need not adjust for the first-step estimation.

4. IV Estimation under Sequential Exogeneity

We now consider IV estimation of the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \; t = 1, \ldots, T, \tag{26}$$

under sequential exogeneity assumptions; the weakest form is $Cov(\mathbf{x}_{is}, u_{it}) = 0$, all $s \leq t$.

This leads to simple moment conditions after first differencing:

$$\mathrm{E}(\mathbf{x}_{is}'\Delta u_{it}) = \mathbf{0}, \; s = 1, \ldots, t-1; \; t = 2, \ldots, T. \tag{27}$$

Therefore, at time t, the available instruments in the

FD equation are in the vector

$\mathbf{x}_{it}^{o} \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it})$. The matrix of instruments is

$$\mathbf{W}_i = \text{diag}(\mathbf{x}_{i1}^{o}, \mathbf{x}_{i2}^{o}, \ldots, \mathbf{x}_{i,T-1}^{o}), \tag{28}$$

which has $T - 1$ rows. Routine to apply GMM estimation.

- Simple strategy: 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, $\widehat{\Delta\mathbf{x}}_{it}$. Then, estimate the FD equation

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \; t = 2, \ldots, T \tag{29}$$

by pooled IV using instruments (not regressors) $\widehat{\Delta\mathbf{x}}_{it}$.

- Can suffer from a weak instrument problem when $\Delta\mathbf{x}_{it}$ has little correlation with $\mathbf{x}_{i,t-1}^{o}$.

- If we assume

$$\mathrm{E}(u_{it}|\mathbf{x}_{it}, y_{i,t-1}\mathbf{x}_{i,t-1}, \ldots, y_{i1}, \mathbf{x}_{i1}, c_i) = 0, \qquad (30)$$

many more moment conditions are available. Using linear functions only, for $t = 3, \ldots, T$,

$$E[(\Delta y_{i,t-1} - \Delta \mathbf{x}_{i,t-1}\boldsymbol{\beta})'(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}. \qquad (31)$$

- Drawback: we often do not want to assume (30). Plus, the conditions in (31) are nonlinear in parameters.
- Arellano and Bover (1995) suggested instead the restrictions

$$Cov(\Delta \mathbf{x}'_{it}, c_i) = 0, \ t = 2, \ldots, T, \qquad (32)$$

which imply linear moment conditions in the levels equation,

$$E[\Delta \mathbf{x}'_{it}(y_{it} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}, t = 2, \ldots, T. \qquad (33)$$

- Simple AR(1) model:

$$y_{it} = \rho y_{i,t-1} + c_i + u_{it}, t = 1, \ldots, T. \quad (34)$$

Typically, the minimal assumptions imposed are

$$E(y_{is}u_{it}) = 0,\ s = 0, \ldots, t-1,\ t = 1, \ldots, T, \quad (35)$$

so for $t = 2, \ldots, T$,

$$E[y_{is}(\Delta y_{it} - \rho \Delta y_{i,t-1}) = 0, s \leq t-2. \quad (36)$$

Again, can suffer from weak instruments when ρ is close to unity. Blundell and Bond (1998) showed that if the condition

$$Cov(\Delta y_{i1}, c_i) = Cov(y_{i1} - y_{i0}, c_i) = 0 \quad (37)$$

is added to $E(u_{it}|y_{i,t-1}, \ldots, y_{i0}, c_i) = 0$ then

$$E[\Delta y_{i,t-1}(y_{it} - \alpha - \rho y_{i,t-1})] = 0 \quad (38)$$

which can be added to the usual moment conditions (35). We have two sets of moments linear in the parameters.

- Condition (37) can be intepreted as a restriction on the initial condition, y_{i0}. Write y_{i0} as a deviation from its steady state, $c_i/(1-\rho)$ (obtained for $|\rho|<1$ by recursive subsitution and then taking the limit), as $y_{i0} = c_i/(1-\rho) + r_{i0}$. Then $(1-\rho)y_{i0} + c_i = (1-\rho)r_{i0}$, and so (37) reduces to

$$Cov(r_{i0}, c_i) = 0. \tag{39}$$

The deviation of y_{i0} from its SS is uncorrelated with the SS.

- Extensions of the AR(1) model, such as

$$y_{it} = \rho y_{i,t-1} + \mathbf{z}_{it}\boldsymbol{\gamma} + c_i + u_{it}, \quad t = 1, \ldots, T. \tag{40}$$

and use FD:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta \mathbf{z}_{it}\boldsymbol{\gamma} + \Delta u_{it}, \quad t = 1, \ldots, T. \tag{41}$$

- Airfare example in notes: $\hat{\rho}_{POLS} = -.126$ $(.027)$, $\hat{\rho}_{IV} = .219$ $(.062)$, $\hat{\rho}_{GMM} = .333$ $(.055)$.

- Arellano and Alvarez (1998) show that the GMM estimator that accounts for the MA(1) serial correlation in the FD errors has better properties when T and N are both large.

5. Pseudo Panels from Pooled Cross Sections

- It is important to distinguish between the population model and the sampling scheme. We are interested in estimating the parameters of

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f + u_t,\ t = 1, \ldots, T, \tag{42}$$

which represents a population defined over T time periods.

- Normalize $E(f) = 0$. Assume all elements of $\mathbf{x}_t$ have some time variation. To interpret $\boldsymbol{\beta}$, contemporaneous exogeneity conditional on f:

$$E(u_t|\mathbf{x}_t, f) = 0, t = 1, \ldots, T. \tag{43}$$

But, the current literature does not even use this assumption. We will use an implication of (43):

$$E(u_t|f) = 0, t = 1, \ldots, T. \tag{44}$$

Because f aggregates all time-constant unobservables, we should think of (44) as implying that $E(u_t|g) = 0$ for any time-constant variable g, whether unobserved or observed.

- Deaton (1985) considered the case of independently sampled cross sections. Assume that the population for which (42) holds is divided into G groups (or cohorts). Common is birth year. For a random draw i at time t, let g_i be the group indicator, taking on a value in $\{1, 2, \ldots, G\}$. Then , by our earlier discussion,

$$E(u_{it}|g_i) = 0. \tag{45}$$

Taking the expected value of (42) conditional on group membership and using only (45), we have

$$E(y_t|g) = \eta_t + E(\mathbf{x}_t|g)\boldsymbol{\beta} + E(f|g), t = 1, \ldots, T. \qquad (46)$$

This is Deaton's starting point, and Moffitt (1993). If we start with (42) under (44), there is no "randomness" in (46). Later authors have left $u_{gt}^* = E(u_t|g)$ in the error term.

- Define the population means

$$\alpha_g = E(f|g), \mu_{gt}^y = E(y_t|g), \boldsymbol{\mu}_{gt}^{\mathbf{x}} = E(\mathbf{x}_t|g) \qquad (47)$$

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

$$\mu_{gt}^y = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g. \qquad (48)$$

- Equation (48) holds without any assumptions restricting the dependence between $\mathbf{x}_t$ and u_r across t and r. In fact, $\mathbf{x}_t$ can contain lagged dependent

variables or contemporaneously endogenous variables. Should we be suspicious?

- Equation (48) looks like a linear regression model in the population means, μ_{gt}^{y} and $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$. One can use a "fixed effects" regression to estimate η_t, α_g, and $\boldsymbol{\beta}$.
- With large cell sizes, N_{gt} (number of observations in each group/time period cell), better to treat as a minimum distance problem. One inefficient MD estimator is fixed effects applied to the sample means, based on the same relationship in the population:

$$\boldsymbol{\beta} = \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} \right)^{-1} \left(\sum_{g=1}^{G} \sum_{t=1}^{T} \ddot{\boldsymbol{\mu}}_{gt}^{\mathbf{x}\prime} \mu_{gt}^{y} \right) \qquad (49)$$

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

regression

$$\boldsymbol{\mu}_{gt}^{\mathbf{x}} \text{ on } 1, d2, \ldots, dT, c2, \ ..., cG, \tag{50}$$

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

- Equation (49) makes it clear that the underlying model in the population cannot contain a full set of group/time interactions. We *could* allow this feature with individual-level data. The absense of full cohort/time effects in the population model is the key identifying restriction.
- $\boldsymbol{\beta}$ is not identified if we can write $\boldsymbol{\mu}_{gt}^{\mathbf{x}} = \boldsymbol{\lambda}_t + \boldsymbol{\omega}_g$ for vectors $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}_g$, $t = 1, \ldots, T$, $g = 1, \ldots, G$. So, we must exclude a full set of group/time effects in the structural model but we need some interaction between them in the distribution of the covariates.

Even then, identification might be weak if the variation in $\{\ddot{\boldsymbol{\mu}}^{\mathbf{x}}_{gt} : t = 1,..,T, g = 1,\ldots,G\}$ is small: a small change in the estimates of $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$ can lead to large changes in $\hat{\boldsymbol{\beta}}$.

- Estimation by nonseparable MD because $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0}$ are the restrictions on the structural parameters $\boldsymbol{\theta}$ given cell means $\boldsymbol{\pi}$ (Chamberlain, lecture notes). But given $\boldsymbol{\pi}$, conditions are linear in $\boldsymbol{\theta}$. After working it through, the optimal estimator is intuitive and easy to obtain. After "FE" estimation, obtain the residual variances within each cell, $\hat{\tau}^2_{gt}$, based on $y_{itg} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - \hat{\alpha}_g - \check{\eta}_t$, where $\check{\boldsymbol{\beta}}$ is the "FE" estimate, and so on.
- Define "regressors" $\hat{\boldsymbol{\omega}}_{gt} = (\hat{\boldsymbol{\mu}}^{\mathbf{x}\prime}_{gt}, \mathbf{d}_t, \mathbf{c}_g)$, and let $\hat{\mathbf{W}}$ be the $GT \times (K + T + G - 1)$ stacked matrix (where we drop, say, the time dummy for the first period.).

Let $\hat{\mathbf{C}}$ be the $GT \times GT$ diagonal matrix with $\hat{\tau}_{gt}^2/(N_{gt}/N)$ down the diagonal. The optimal MD estimator, which is $\sqrt{N}$-asymptotically normal, is

$$\hat{\boldsymbol{\theta}} = (\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}^{y}. \qquad (51)$$

As in separable cases, the efficient MD estimator looks like a "weighted least squares" estimator and its asymptotic variance is estimated as $(\hat{\mathbf{W}}'\hat{\mathbf{C}}^{-1}\hat{\mathbf{W}})^{-1}/N$. (Might be better to use resampling method here.)

- Inoue (2007) obtains a different limiting distribution, which is stochastic, because he treats estimation of $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ and μ_{gt}^{y} asymmetrically.
- 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.

- Allows for models with lagged dependent variables, but now the vectors of means contain redundancies. If

$$y_t = \eta_t + \rho y_{t-1} + \mathbf{z}_t\boldsymbol{\gamma} + f + u_t,\ E(u_t|g) = 0, \tag{52}$$

then the same moments are 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. We now have fewer moment conditions to estimate the parameters.

- The MD approach applies to extensions of the basic model. Random trend model (Heckman and Hotz (1989)):

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + f_1 + f_2 t + u_t. \tag{53}$$

$$\mu_{gt}^y = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \alpha_g + \varphi_g t, \tag{54}$$

We can even estimate models with time-varying factor loads on the heterogeneity:

$$y_t = \eta_t + \mathbf{x}_t\boldsymbol{\beta} + \lambda_t f + u_t, \tag{55}$$

$$\mu_{gt}^{y} = \eta_t + \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \lambda_t \alpha_g. \tag{56}$$

- How can we use a stronger assumption, such as $E(u_t|\mathbf{z}_t, f) = \mathbf{0}$, $t = 1, \ldots, T$, for instruments $\mathbf{z}_t$, to more precisely estimate $\boldsymbol{\beta}$? Gives lots of potentially useful moment conditions:

$$E(\mathbf{z}_t' y_t|g) = \eta_t E(\mathbf{z}_t'|g) + E(\mathbf{z}_t'\mathbf{x}_t|g)\boldsymbol{\beta} + E(\mathbf{z}_t' f|g), \tag{57}$$

using $E(\mathbf{z}_t' u_t|g) = \mathbf{0}$.

"What's New in Econometrics"

Lecture 3

Regression Discontinuity Designs

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction
2. Basics
3. Graphical Analyses
4. Local Linear Regression
5. Choosing the Bandwidth
6. Variance Estimation
7. Specification Checks

1. Introduction

A Regression Discontinuity (RD) Design is a powerful and widely applicable identification strategy.

Often access to, or incentives for participation in, a service or program is assigned based on transparent rules with criteria based on clear cutoff values, rather than on discretion of administrators.

Comparisons of individuals that are similar but on different sides of the cutoff point can be credible estimates of causal effects for a specific subpopulation.

Good for internal validity, not much external validity.

2. Basics

Two potential outcomes $Y_i(0)$ and $Y_i(1)$,
causal effect $Y_i(1) - Y_i(0)$,
binary treatment indicator W_i, covariate X_i,
and the observed outcome equal to:

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

At $X_i = c$ incentives to participate change.

Two cases, **Sharp Regression Discontinuity**:

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

and **Fuzzy Regression Discontinuity Design**:

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

Sharp Regression Discontinuity

Example (Lee, 2007)

What is effect of incumbency on election outcomes? (More specifically, what is the probability of a Democrat winning the next election given that the last election was won by a Democrat?)

Compare election outcomes in cases where previous election was very close.

SRD

Key assumption:
$\mathbb{E}[Y(0)|X = x]$ and $\mathbb{E}[Y(1)|X = x]$ are continuous in x.

Under this assumption,

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

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

Extrapolation is unavoidable.

Fuzzy Regression Discontinuity

Examples (VanderKlaauw, 2002)

What is effect of financial aid offer on acceptance of college admission.

College admissions office puts applicants in a few categories based on numerical score.

Financial aid offer is highly correlated with category.

Compare individuals close to cutoff score.

FRD

What do we look at in the FRD case: ratio of discontinuities in regression function of outcome and treatment:

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

(FRD Estimand)

Interpretation of FRD (Hahn, Todd, VanderKlaauw)

Let $W_i(x)$ be potential treatment status given cutoff point x, for x in some small neigborhood around c (which requires that the cutoff point is at least in principle manipulable)

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

A complier is a unit such that

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

Then

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

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

External Validity

The estimatand has little external validity. It is at best valid for a population defined by the cutoff value c, and by the sub-population that is affected at that value.

FRD versus Unconfoundedness

Unconfoundedness:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \,\Big|\, X_i. \qquad \text{(unconfoundedness)}$$

Under this assumption:

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

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

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 $X_i = c$ have a causal interpretation.

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.

3. Graphical Analyses

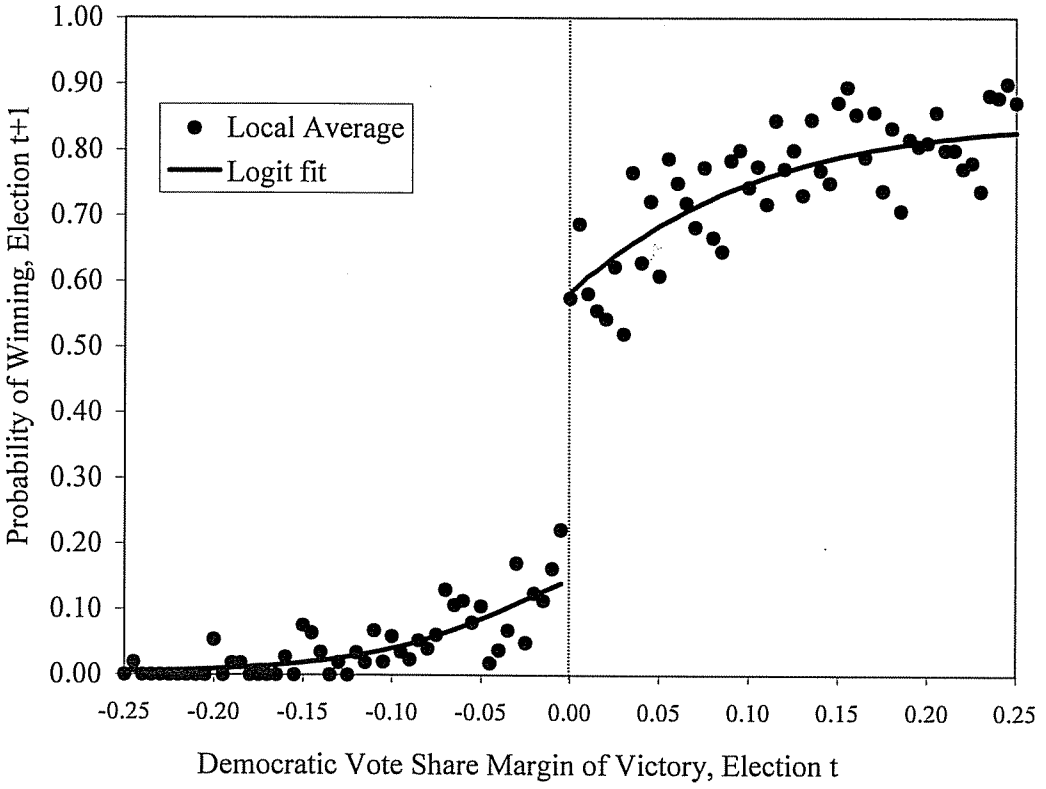
A. Plot regression function $\mathbb{E}[Y_i|X_i = x]$

B. Plot regression functions $\mathbb{E}[Z_i|X_i = x]$ for covariates that do not enter the assignment rule Z_i

C. Plot density $f_X(x)$.

In all cases use estimators that do not smooth around the cutoff value. For example, for binwidth h define bins $[b_{k-1}, b_k]$, where $b_k = c - (K_0 - k + 1) \cdot h$, and average outcomes within bins.

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



4. Local Linear Regression

We are interested in the value of a regression function at the boundary of the support. Standard kernel regression

$$\widehat{\mu_l(c)} = \sum_{i|c-h<X_i<c}^{N} Y_i \Big/ \sum_{i|c-h<X_i<c}^{N} 1 \tag{2}$$

does not work well for that case (slower convergence rates)

Better rates are obtained by using local linear regression. First

$$\min_{\alpha_l,\beta_l} \sum_{i|c-h<X_i<c}^{N} \left(Y_i - \alpha_l - \beta_l \cdot (X_i - c)\right)^2, \tag{3}$$

The value of lefthand limit $\mu_l(c)$ is then estimated as

$$\widehat{\mu_l(c)} = \widehat{\alpha}_l + \widehat{\beta}_l \cdot (c - c) = \widehat{\alpha}_l. \tag{4}$$

Similarly for righthand side. Not much gained by using a non-uniform kernel.

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

$$Y_i = \alpha + \beta \cdot (X_i - c) + \tau \cdot W_i + \gamma \cdot (X_i - c) \cdot W_i + \varepsilon_i$$

thus solving

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{N} 1\{c - h \leq X_i \leq c + h\}$$

$$\times \left(Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i\right)^2,$$

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

This interpretation extends easily to the inclusion of covariates.

Estimation for the FRD Case

Do local linear regression for both the outcome and the treatment indicator, on both sides,

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

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

and similarly $(\widehat{\alpha}_{yr}, \widehat{\beta}_{yr})$ and $(\widehat{\alpha}_{wr}, \widehat{\beta}_{wr})$. Then the FRD estimator is

$$\widehat{\tau}_{\text{FRD}} = \frac{\widehat{\tau}_y}{\widehat{\tau}_w} = \frac{\widehat{\alpha}_{yr} - \widehat{\alpha}_{yl}}{\widehat{\alpha}_{wr} - \widehat{\alpha}_{wl}}.$$

Alternatively, define the vector of covariates

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

Then we can write

$$Y_i = \delta' V_i + \tau \cdot W_i + \varepsilon_i. \qquad \text{(TSLS)}$$

Then estimating τ based on the regression function (TSLS) by Two-Stage-Least-Squares methods, using

W_i as the endogenous regressor,
the indicator $1\{X_i \geq c\}$ as the excluded instrument
V_i as the set of exogenous variables

This is is numerically identical to $\widehat{\tau}_{\text{FRD}}$ before (because of uniform kernel)

Can add other covariates in straightfoward manner.

5. Choosing the Bandwidth (Ludwig & Miller)

We wish to take into account that (i) we are interested in the regression function at the boundary of the support, and (ii) that we are interested in the regression function at $x = c$.

Define $\widehat{\alpha}_l(x)$, $\widehat{\beta}_l(x)$, $\widehat{\alpha}_r(x)$ and $\widehat{\beta}_r(x)$ as the solutions to

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

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

Define

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

Define $q_{X,\delta,l}$ to 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 $X_i \geq c$.

Now we use the cross-validation criterion

$$\mathsf{CV}_Y(h) = \sum_{i: q_{X,\delta,l} \leq X_i \leq q_{X,1-\delta,r}} (Y_i - \widehat{\mu}(X_i))^2 ,$$

for, say $\delta = 1/2$, with the corresponding cross-validation choice for the binwidth

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

Bandwidth for FRD Design

1. Calculate optimal bandwidth separately for both regression functions and choose smallest.

2. Calculate optimal bandwidth only for outcome and use that for both regression functions.

Typically the regression function for the treatment indicator is flatter than the regression function for the outcome away from the discontinuity point (completely flat in the SRD case). So using same criterion would lead to larger bandwidth for estimation of regression function for treatment indicator. In practice it is easier to use the same bandwidth, and so to avoid bias, use the bandwidth from criterion for SRD design or smallest.

6. Variance Estimation

$$\sigma^2_{Yl} = \lim_{x \uparrow c} \mathsf{Var}(Y_i | X_i = x), \quad C_{YWl} = \lim_{x \uparrow c} \mathsf{Cov}(Y_i, W_i | X_i = x),$$

$$V_{\tau_y} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Yr} + \sigma^2_{Yl} \right), \quad V_{\tau_w} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Wr} + \sigma^2_{Wl} \right)$$

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

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

Finally, the asymptotic distribution has the form

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

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

Can use plug in estimators for components of variance.

TSLS Variance for FRD Design

The second estimator for the asymptotic variance of $\hat{\tau}$ exploits the interpretation of the $\hat{\tau}$ as a TSLS estimator.

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 as the exogenous covariates.

7. Concerns about Validity

Two main conceptual concerns in the application of RD designs, sharp or fuzzy.

Other Changes

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.

Manipulation of Forcing Variable

The second concern is that of manipulation of the covariate value.

Specification Checks

A. Discontinuities in Average Covariates

B. A Discontinuity in the Distribution of the Forcing Variable

C. Discontinuities in Avareage Outcomes at Other Values

D. Sensitivity to Bandwidth Choice

E. RD Designs with Misspecification

7.A Discontinuities in Average Covariates

Test 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.

Although not required for the validity of the design, 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.

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

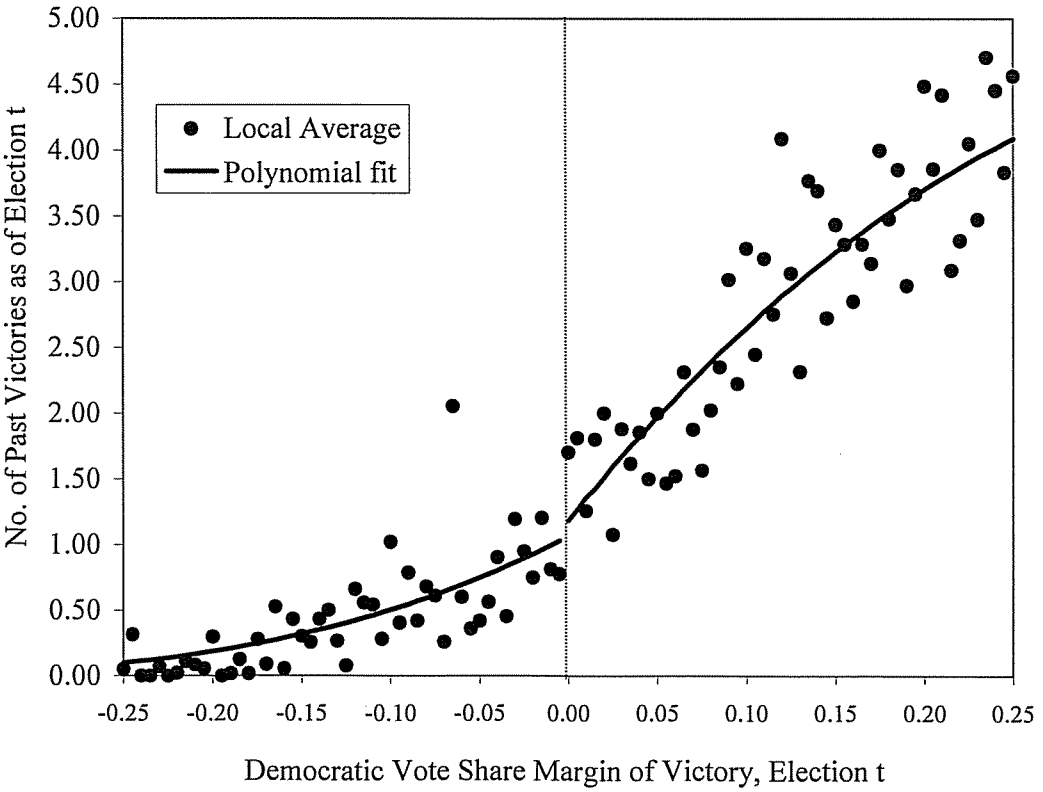
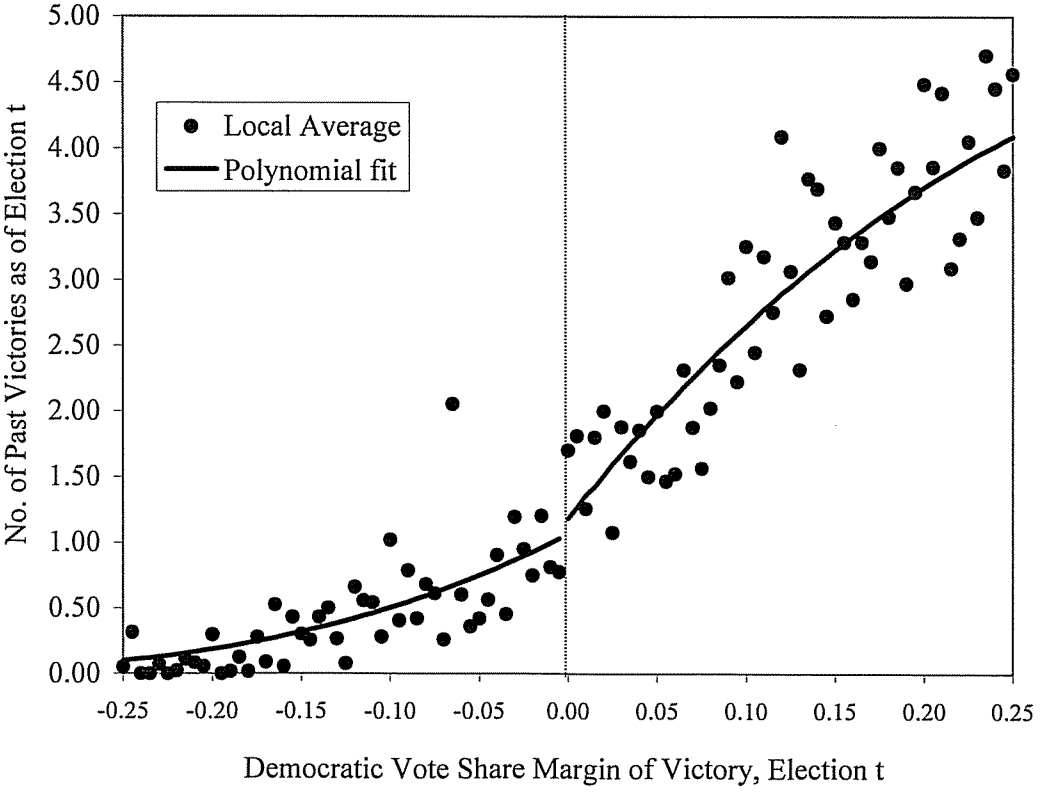


Figure IIb: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and parametric fit



7.B A Discontinuity in the Distribution of the Forcing Variable

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, the design does not require 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.

7.C Discontinuities in Avareage Outcomes at Other Values

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.

7.D Sensitivity to Bandwidth Choice

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.

7.E 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 $\mu(x)$.

They consider a parametric specification for the regression function that does not fully saturate the model and interpret the deviation between the true conditional expectation 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 and compare them to the conventional ones.

What's New in Econometrics?
Lecture 4
Nonlinear Panel Data Models

Jeff Wooldridge
NBER Summer Institute, 2007

1. Basic Issues and Quantities of Interest

2. Exogeneity Assumptions

3. Conditional Independence

4. Assumptions about the Unobserved Heterogeneity

5. Nonparametric Identification of Average Partial Effects

6. Dynamic Models

7. Applications to Specific Models

8. Estimating the Fixed Effects

1. Basic Issues and Quantities of Interest

- Let $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$ be a random draw from the cross section. Typically interested in

$$D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \tag{1}$$

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? If x_{tj} is continuous, then

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}, \tag{2}$$

or discrete changes. How do we account for unobserved $\mathbf{c}_i$? If we know enough about the distribution of $\mathbf{c}_i$ 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)*,

$$PEA_j(\mathbf{x}_t) = \theta_j(\mathbf{x}_t, \boldsymbol{\mu}_\mathbf{c}). \quad (3)$$

Of course, we need to estimate the function m_t and $\boldsymbol{\mu}_\mathbf{c}$. We might be able to insert different quantiles, or a certain number of standard deviations from the mean.

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

$$APE(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)]. \quad (4)$$

The difference between (3) and (4) can be nontrivial. In some leading cases, (4) is identified while (3) is not. (4) is closely related to the notion of the average structural function (ASF) (Blundell and Powell (2003)). The ASF is defined as

$$ASF(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \quad (5)$$

- Passing the derivative through the expectation in

(5) gives the APE.

- How do APEs relate to parameters? Suppose

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{6}$$

where, say, $G(\cdot)$ is strictly increasing and continuously differentiable. Then

$$\theta_j(\mathbf{x}_t, c) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \tag{7}$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Then estimating β_j means we can sign of the partial effect, and the relative effects of any two continuous variables. Even if $G(\cdot)$ is specified, the magnitude of effects cannot be estimated without making assumptions about the distribution of c_i

- Altonji and Matzkin (2005) define 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

$$LAR_j(\mathbf{x}_t) = \int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \mathrm{d}H_t(\mathbf{c}|\mathbf{x}_t), \tag{8}$$

where $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. "Local" because it averages out the heterogeneity for the slice of the population described by the vector $\mathbf{x}_t$. The APE is a "global" average response."

- Definitions of partial effects do not depend on whether $\mathbf{x}_t$ is correlated with $\mathbf{c}$. Of course, whether and how we estimate them certainly does.

2. Exogeneity Assumptions

- As in linear case, cannot get by with just specifying a model for $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$.
- The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \tag{9}$$

Chamberlain (1984) labeled (9) *strict exogeneity*

conditional on the unobserved effects $\mathbf{c}_i$.

Conditional mean version:

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \quad (10)$$

- The sequential exogeneity assumption is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \quad (11)$$

Unfortunately, it is much more difficult to allow sequential exogeneity in in nonlinear models.

- Neither (9) nor (10) 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}. (Later in control function estimation.)

3. Conditional Independence

- In linear models, serial dependence of idiosyncratic shocks is easily dealt with, either by

robust inference or GLS extensions of FE and FD. With strictly exogenous covariates, never results in biased estimation, even if it is ignored or improperly model. The situation is different with nonlinear models estimated by MLE.

- The conditional independence assumption is

$$D(y_{i1},\ldots,y_{iT}|\mathbf{x}_i,\mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i) \qquad (12)$$

(where we also impose strict exogeneity). In a parametric context, the CI 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 correlated random effects frameworks, CI plays a critical role in being able to estimate the “structural” parameters and the

parameters in distribution the of $\mathbf{c}_i$ (and therefore, PAEs). In a broad class of models, CI plays no role in estimating APEs.

4. Assumptions about the Unobserved Heterogeneity

Random Effects

$$D(\mathbf{c}_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}) = D(\mathbf{c}_i). \tag{13}$$

Under (13), the APEs are nonparametrically identified from

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t). \tag{14}$$

- In some leading cases (RE probit and RE Tobit with heterogeneity normally distributed), if we want PEs for different values of $\mathbf{c}$, we must assume more: strict exogeneity, conditional independence,

and (13) with a parametric distribution for $D(\mathbf{c}_i)$.

Correlated Random Effects

A CRE framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but 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 Chamberlain (1980,1982), and much work since. Can allow $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ to depend in a "nonexchangeable" manner. (Chamberlain's CRE probit and Tobit models.) Distributional assumptions that lead to simple estimation – homoskedastic normal with a linear conditional mean — are restrictive.

- Possible to drop parametric assumptions with

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i), \tag{15}$$

without restricting $\mathrm{D}(c_i|\bar{\mathbf{x}}_i)$.

- As T gets larger, can allow $\mathbf{c}_i$ to be correlated

with features of the covariates other than just the time average. Altonji and Matzkin (2005) allow for $\bar{\mathbf{x}}_i$ in equation (15) to be replaced by other functions of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$, such as sample variances and covariance. Non-exchangeable functions, such as unit-specific trends, can be used, too. Generally, assume

$$\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\mathbf{w}_i). \tag{16}$$

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 $\mathbf{w}_i$.

- Altonji and Matzkin use exchangeability and other restrictions, such as monotonicity

Fixed Effects

The label “fixed effects” is used in different ways by different researchers. One view: $\mathbf{c}_i,\ i = 1, \ldots, N$

are parameters to be estimated. Usually leads to an "incidental parameters problem" (which attentuates with large T).

- A second meaning of "fixed effects" is that $D(\mathbf{c}_i|\mathbf{x}_i)$ is unrestricted and we look for objective functions that do not depend on $\mathbf{c}_i$ but still identify the population parameters. Leads to "conditional maximum likelihood" if we can find a "sufficient statistic" such that

$$D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{c}_i,\mathbf{s}_i) = D(y_{i1},\ldots,y_{it}|\mathbf{x}_i,\mathbf{s}_i). \qquad (17)$$

- The CI assumption is usually maintained.

5. Nonparametric Identification of Average Partial Effects

- Identification of PAEs can fail even under a strong set of parametric assumptions. In the probit model

$$P(y = 1|\mathbf{x}, c) = \Phi(\mathbf{x}\boldsymbol{\beta} + c), \tag{18}$$

the PE for a continuous variable x_j is $\beta_j \phi(\mathbf{x}\boldsymbol{\beta} + c)$. The PAE at $\mu_c = E(c) = 0$ is $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$. Suppose $c|\mathbf{x}$ ~Normal$(0, \sigma_c^2)$. Then

$$P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}), \tag{19}$$

so only the scaled parameter vector $\boldsymbol{\beta}_c \equiv \boldsymbol{\beta}/(1 + \sigma_c^2)^{1/2}$ is identified; $\boldsymbol{\beta}$ and $\beta_j \phi(\mathbf{x}\boldsymbol{\beta})$ are not identified.

- The APE is identified from $P(y = 1|\mathbf{x})$, and is given by $\beta_{cj} \phi(\mathbf{x}\boldsymbol{\beta}_c)$. (Attenuation bias?)
- Panel data example due to Hahn (2001): x_{it} is a binary indicator and

$$P(y_{it} = 1|\mathbf{x}_i, c_i) = \Phi(\beta x_{it} + c_i), t = 1, 2. \tag{20}$$

β is not known to be identified in this model, even under conditional independence *and* the random

effects assumption $D(c_i|\mathbf{x}_i) = D(c_i)$. But the APE is $\tau \equiv E[\Phi(\beta + c_i)] - E[\Phi(c_i)]$ and is identified by a difference of means for the treated and untreated groups, for either time period.

- As shown in Wooldridge (2005a), identification of the APE holds if we replace Φ with an unknown function G and allow $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$.
- Are we focusing too much on parameters? In many cases, yes, but not always so clear cut. From Wooldridge (2005c): $y = 1[\mathbf{x}\boldsymbol{\beta} + u > 0]$ where $u|\mathbf{x}$ ~Normal$(0, \exp(2\mathbf{x}\boldsymbol{\delta}))$ ("heteroskedastic probit"). $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ estimable by MLE. The APE for x_j is *not* obtained by differentiating $P(y = 1|\mathbf{x}) = \Phi[\exp(-\mathbf{x}\boldsymbol{\delta})\mathbf{x}\boldsymbol{\beta}]$ with respect to x_j, which can have a different sign from β_j. Instead, for given $\mathbf{x}$, it is consistently estimated as

$$\widehat{APE}_j(\mathbf{x}) = \hat{\beta}_j \left\{ N^{-1} \sum_{i=1}^{N} \phi[\exp(-\mathbf{x}_i\hat{\boldsymbol{\delta}})\mathbf{x}\hat{\boldsymbol{\beta}}] \right\},$$

which always has the same sign as $\hat{\beta}_j$.

- We can establish identification of APEs in panel data applications very under strict exogeneity along with $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. These two assumptions identify the APEs. Write the average structural function at time t as

$$\begin{aligned} \mathrm{ASF}_t(\mathbf{x}_t) &= \mathrm{E}_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] \\ &= \mathrm{E}_{\bar{\mathbf{x}}_i}\{\mathrm{E}[m_t(\mathbf{x}_t, \mathbf{c}_i)|\bar{\mathbf{x}}_i]\} \\ &\equiv \mathrm{E}_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)], \end{aligned} \tag{21}$$

Given a consistent estimator of $\hat{r}_t(\cdot, \cdot)$, the ASF can be estimated as

$$\widehat{\mathrm{ASF}}_t(\mathbf{x}_t) \equiv N^{-1} \sum_{i=1}^{N} \hat{r}_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)., \tag{22}$$

- Equation (21) holds without strict exogeneity $\mathrm{D}(c_i|\mathbf{x}_i) = \mathrm{D}(c_i|\bar{\mathbf{x}}_i)$. But these assumptions allow us to estimate estimate $r_t(\cdot,\cdot)$:

$$\begin{aligned}\mathrm{E}(y_{it}|\mathbf{x}_i) &= \mathrm{E}[\mathrm{E}(y_{it}|\mathbf{x}_i,\mathbf{c}_i)|\mathbf{x}_i] = \mathrm{E}[m_t(\mathbf{x}_{it},\mathbf{c}_i)|\mathbf{x}_i] \\ &= \int m_t(\mathbf{x}_{it},\mathbf{c})\mathrm{d}F(\mathbf{c}|\mathbf{x}_i) \\ &= \int m_t(\mathbf{x}_{it},\mathbf{c})\mathrm{d}F(\mathbf{c}|\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it},\bar{\mathbf{x}}_i), \qquad (23)\end{aligned}$$

where $F(\mathbf{c}|\mathbf{x}_i)$ denotes the cdf of $\mathrm{D}(\mathbf{c}_i|\mathbf{x}_i)$ Because $\mathrm{E}(y_{it}|\mathbf{x}_i)$ depends only on $(\mathbf{x}_{it},\bar{\mathbf{x}}_i)$, we must have

$$\mathrm{E}(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i) = r_t(\mathbf{x}_{it},\bar{\mathbf{x}}_i), \qquad (24)$$

and $r_t(\cdot,\cdot)$ is identified with sufficient time variation in $\mathbf{x}_{it}$.

6. Dynamic Models

- Nonlinear models with only sequentially exogenous variables are difficult to deal with. More is known about models with lagged dependent

variables and otherwise strictly exogenous variables:

$$D(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \ldots, \mathbf{z}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i), \; t = 1, \ldots, T, \tag{25}$$

which we assume also is $D(\mathbf{y}_{it}|\mathbf{z}_i, \mathbf{y}_{i,t-1}, \ldots, \mathbf{y}_{i1}, \mathbf{y}_{i0}, \mathbf{c}_i)$. Suppose this distribution depends only on $(\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i)$ with density $f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta})$. The joint density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $(\mathbf{y}_{i0}, \mathbf{z}_i, \mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}). \tag{26}$$

- How do we deal with $\mathbf{c}_i$ along with the initial condition, $\mathbf{y}_{i0}$? Approaches: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (incidental parameters problem). (ii) Try to estimate the parameters without specifying conditional or unconditional

distributions for c_i (available in some special cases). Generally, cannot estimate partial effects.). (iii) Approximate $D(\mathbf{y}_{i0}|\mathbf{c}_i, z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. Leads to $D(\mathbf{y}_{i0}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{z}_i)$ and MLE conditional on $\mathbf{z}_i$. (iv) Model $D(\mathbf{c}_i|\mathbf{y}_{i0}, \mathbf{z}_i)$. Leads to $D(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}|\mathbf{y}_{i0}, \mathbf{z}_i)$ and MLE conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$. Wooldridge (2005b) shows this can be computationally simple for popular models.

- 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, the APEs are easy to obtain.

7. Applications to Specific Models

Binary and Fractional Response

- Unobserved effects (UE) probit model:

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \; t = 1, \ldots, T. \qquad (27)$$

Assume strict exogeneity (as always, conditional on c_i) and use Chamberlain-Mundlak device under

conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, a_i|\mathbf{x}_i \sim \text{Normal}(0, \sigma_a^2). \quad (28)$$

If we still assume conditional serial independence then all parameters are identified and MLE (RE probit) can be used. $\hat{\mu}_c = \hat{\psi} + \left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}$ and $\hat{\sigma}_c^2 \equiv \hat{\boldsymbol{\xi}}'\left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i'\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. c_i is not generally normally distributed unless $\bar{\mathbf{x}}_i\boldsymbol{\xi}$ is. But can evaluate PEs at, say, $\hat{\mu}_c \pm k\hat{\sigma}_c$.

- The APEs are identified from the ASF, which is consistently estimated as

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \quad (29)$$

where, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1+\hat{\sigma}_a^2)^{1/2}$.

- APEs are identified without the conditional serial independence assumption. Use the marginal

probabilities to estimate scaled coefficients:

$$P(y_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a). \quad (30)$$

(Time dummies have been supressed for simplicity.)

- Can used pooled probit or minimum distance or "generalized estimating equations."
- 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:
$E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$. Full MLE difficult.
- A more radical suggestion, but in the spirit of Altonji and Matzkin (2005), is to just use a flexible model for $\mathrm{E}(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ directly, say,

$$E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}].$$

Just average out over $\bar{\mathbf{x}}_i$ to get APEs.

- Can use same idea with logit. But, if we have a binary response, start with

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{31}$$

and assume conditional independence assumption, we can estimate $\boldsymbol{\beta}$ without restricting $D(c_i|\mathbf{x}_i)$.

- Because we have not restricted $D(c_i|\mathbf{x}_i)$ in any way, it appears that we cannot estimate average partial effects. See table in notes for the tradeoffs in using CRE models and conditional MLE.

- Example from notes. Estimated APEs for number of small children on women's labor force participation: linear, –.0389 (.0092); probit

(pooled), –.0660 (.0048); CRE probit (pooled) –.0389 (.0085); CRE probit (MLE), –.0403 (.104), FE logit, coefficient = –.644 (.125).

- What would CMLE logit estimate in the model

$$P(y_{it} = 1|\mathbf{x}_{it}, \mathbf{c}_i) = \Lambda(a_i + \mathbf{x}_{it}\mathbf{b}_i), \tag{32}$$

where $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$?

- There are methods that allow estimation, up to scale, of the coefficients without even specifying the distribution of u_{it} in

$$y_{it} = 1[\mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0]. \tag{33}$$

under strict exogeneity.conditional on c_i. Arellano and Honoré (2001).

- Simple dynamic model:

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i). \tag{34}$$

A simple analysis is available if we specify

$$c_i|\mathbf{z}_i, y_{i0} \sim Normal(\psi + \xi_0 y_{i0} + \mathbf{z}_i\xi, \sigma_a^2) \tag{35}$$

Then

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i), \tag{36}$$

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:

$$\widehat{ASF}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N} \Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a), \tag{37}$$

Example in notes: dynamic labor force partication. The APE estimated from this method is about .259. If we ignore the heterogeneity, APE is .837.

- For estimating parameters, Honoré and Kyriazidou (2000) extend an idea of Chamberlain. With four 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. By design, cannot estimate magnitudes of effects.

Count and Other Multiplicative Models

- Several options are available for models with conditional means multiplicative in the heterogeneity. The most common is

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{38}$$

where $c_i \geq 0$. If we assume strict exogeneity,

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i), \tag{39}$$

a particular quasi-MLE is attractive as it does not restrict $D(y_{it}|\mathbf{x}_i, c_i)$, $D(c_i|\mathbf{x}_i)$, or serial dependence: the "fixed effects" Poisson estimator. It is the conditional MLE derived under a Poisson distributional assumption and the conditional independence assumption. But it is fully robust, even if y_{it} is not a count variable! It turns out that there is no incidental parameters problem in this case. Fully robust inference is easy to obtain (Wooldridge (1999)).

- Estimation under sequential exogeneity has been studied by Chamberlain (1992) and Wooldridge

(1997). In particular, they obtain moment conditions for models such as

$$E(y_{it}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}). \quad (40)$$

Under this assumption, it can be shown that

$$E\{[y_{it} - y_{i,t+1}\exp((\mathbf{x}_{it} - \mathbf{x}_{i,t+1})\boldsymbol{\beta})|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}] = 0, \quad (41)$$

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.

- Wooldridge (2005b) shows how a dynamic Poisson model with conditional Gamma heterogeneity can be easily estimated.

8. Estimating the Fixed Effects

- Except in special cases (linear and Poisson), treating the c_i as parameters to estimate leads to

inconsistent estimates of the population parameters $\boldsymbol{\theta}$. But 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 APEs, based on

$$N^{-1}\sum_{i=1}^{N}\frac{\partial m_t(\mathbf{x}_t,\hat{\boldsymbol{\theta}},\hat{\mathbf{c}}_i)}{\partial x_{tj}}, \tag{42}$$

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?

- Hahn and Newey (2004) propose both jackknife and analytical bias corrections and show that they work well for the probit case. The jackknife FE estimator is

$$\tilde{\boldsymbol{\theta}} = T\hat{\boldsymbol{\theta}} - (T-1)T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\theta}}_{(t)}, \qquad (43)$$

where $\hat{\boldsymbol{\theta}}$ is the FE estimate using all time periods and $\hat{\boldsymbol{\theta}}_{(t)}$ is the estimate that drops time period t. The asymptotic bias of $\tilde{\boldsymbol{\theta}}$ is on the order of T^{-2}.

- Practical limitations of the jackknife. First, aggregate time effects are not allowed, and they would be difficult to include because the analysis is with $T \to \infty$. Also, heterogeneity in the distributions across t changes the bias terms and so (43) does not remove the bias. Hahn and Newey assume independence across t conditional on c_i. Even relaxing this, the "leave-one-out" method does not apply to dynamic models.
- Fernández-Val (2007) shows that in a model with time series dependence in strictly exogenous

regressors, the APEs based on the fixed effects estimator have bias of order T^{-2} in the case that there is no heterogeneity.

WNE
7/30/07
#5

"What's New in Econometrics"

Lecture 5

Instrumental Variables with Treatment Effect Heterogeneity: Local Average Treatment Effects

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction
2. Basics
3. Local Average Treatment Effects
4. Extrapolation to the Population
5. Covariates
6. Multivalued Instruments
7. Multivalued Endogenous Regressors

1. Introduction

1. Instrumental variables estimate average treatment effects, with the average depending on the instruments.

2. Population averages are only estimable under unrealistically strong assumptions ("identification at infinity", or under the constant effect).

3. Compliers (for whom we can identify effects) are not necessarily the subpopulations that are *ex ante* the most interesting subpopulations, but need extrapolation for others.

4. 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.

2. Basics

Linear IV with Constant Coefficients. Standard set up:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \varepsilon_i.$$

There is concern that the regressor W_i is endogenous, correlated with ε_i. Suppose that we have an instrument Z_i that is both uncorrelated with ε_i and correlated with W_i.

In the single instrument / single endogenous regressor, we end up with the ratio of covariances

$$\hat{\beta}_1^{\text{IV}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}) \cdot (Z_i - \bar{Z})}{\frac{1}{N}\sum_{i=1}^{N}(W_i - \bar{W}) \cdot (Z_i - \bar{Z})}.$$

Using a central limit theorem for all the moments and the delta method we can infer the large sample distribution without additional assumptions.

Potential Outcome Set Up

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. Let W_i be the realized value of the endogenous regressor, equal to zero or one. We observe W_i and

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

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 Z_i. The actual or realized value of the endogenous variable is

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

So we observe the triple Z_i, $W_i = W_i(Z_i)$ and $Y_i = Y_i(W_i(Z_i))$.

3. Local Average Treatment Effects

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.

Assumptions (ctd)

Alternatively, we separate the assumption by postulating 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 $W_i = w$.

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)

$$Y_i(z, w) = Y_i(z', w), \quad \text{for all } z, z', w.$$

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.

Compliance Types

It is useful for our approach to think about the compliance behavior of the different units

		$W_i(0)$	
		0	1
$W_i(1)$	0	never-taker	defier
	1	complier	always-taker

We cannot directly establish the type of a unit based 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.

		Z_i	
		0	1
W_i	0	complier/never-taker	never-taker/defier
	1	always-taker/defier	complier/always-taker

Monotonicity

Assumption 4 (Monotonicity/No-Defiers)

$$W_i(1) \geq W_i(0).$$

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.

Implications for Compliance types:

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

For individuals with $(Z_i = 0, W_i = 1)$ and for $(Z_i = 1, W_i = 0)$ we can now infer the compliance type.

Distribution of Compliance Types

Under random assignment and monotonicity we can estimate the distribution of compliance types:

$$\pi_a = \Pr(W_i(0) = W_i(1) = 1) = \mathbb{E}[W_i|Z_i = 0]$$

$$\pi_c = \Pr(W_i(0) = 0, W_i(1) = 1) = \mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]$$

$$\pi_n = \Pr(W_i(0) = W_i(1) = 0) = 1 - \mathbb{E}[W_i|Z_i = 1]$$

Now consider average outcomes by instrument and treatment:

$$\mathbb{E}[Y_i|W_i = 0, Z_i = 0] =$$

$$\frac{\pi_c}{\pi_c + \pi_n} \cdot \mathbb{E}[Y_i(0)|\text{complier}] + \frac{\pi_n}{\pi_c + \pi_n} \cdot \mathbb{E}[Y_i(0)|\text{never} - \text{taker}],$$

$$\mathbb{E}[Y_i|W_i = 0, Z_i = 1] = \mathbb{E}[Y_i(0)|\text{never} - \text{taker}],$$

$$\mathbb{E}[Y_i|W_i = 1, Z_i = 0] = \mathbb{E}[Y_i(1)|\text{always} - \text{taker}],$$

$$\mathbb{E}[Y_i|W_i = 1, Z_i = 1] =$$

$$\frac{\pi_c}{\pi_c + \pi_a} \cdot \mathbb{E}[Y_i(1)|\text{complier}] + \frac{\pi_a}{\pi_c + \pi_a} \cdot \mathbb{E}[Y_i(1)|\text{always} - \text{taker}].$$

From this we can infer the average outcome for compliers,

$$\mathbb{E}[Y_i(0)|\text{complier}], \quad \text{and} \quad \mathbb{E}[Y_i(1)|\text{complier}],$$

Local Average Treatment Effect Hence the instrumental variables estimand, the ratio of these two reduced form estimands, is equal to the local average treatment effect

$$\beta^{\text{IV}} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]}$$

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

4. Extrapolating to the Full Population

We can estimate

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \quad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right]$$

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 $Y_i(0)$;

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \quad \text{and} \quad \mathbb{E}\left[Y_i(0)|\text{complier}\right],$$

and the pair of average outcomes of $Y_i(1)$:

$$\mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right], \quad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{complier}\right].$$

If compliers, never-takers and always-takers are found to be substantially different in levels, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types.

5. Covariates

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 between treatment effects and 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.

Heckman Selection Model

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 impose restrictions on the relation between compliance types, covariates and outcomes:

$$i \text{ is a } \begin{cases} \text{never}-\text{taker} & \text{if} \quad \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{complier} & \text{if} \quad -\pi_0 - \pi_1 - \pi_2' X_i \leq \eta_i < -\pi_0 - \pi_1 - \pi_2' X \\ \text{always}-\text{taker} & \text{if} \quad -\pi_0 - \pi_2' X_i \leq \eta_i, \end{cases}$$

which imposes strong restrictions, e.g., if $\mathbb{E}[Y_i(0)|n, X_i] < \mathbb{E}[Y_i(0)|c, X_i]$, then $\mathbb{E}[Y_i(1)|c, X_i] < \mathbb{E}[Y_i(1)|a, X_i]$

Flexible Alternative Model

Specify

$$f_{Y(w)|X,T}(y|x,t) = f(y|x;\theta_{wt}),$$

for $(w,t) = (0,n),(0,c),(1,c),(1,a)$. A natural model for the distribution of type is a trinomial logit model:

$$\Pr(T_i = \text{complier}|X_i) = \frac{1}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)},$$

$$\Pr(T_i = \text{never}-\text{taker}|X_i) = \frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)},$$

$$\Pr(T_i = \text{always}-\text{taker}|X_i) =$$

$$1 - \Pr(T_i = \text{complier}|X_i) - \Pr(T_i = \text{never}-\text{taker}|X_i).$$

The log likelihood function is then, factored in terms of the contribution by observed (W_i, Z_i) values:

$$\mathcal{L}(\pi_n, \pi_a, \theta_{0n}, \theta_{0c}, \theta_{1c}, \theta_{1a}) =$$

$$\times \prod_{i|W_i=0,Z_i=1} \frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{0n})$$

$$\times \prod_{i|W_i=0,Z_i=0} \left(\frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)} \cdot f(Y_i|X_i;\theta_{0n}) + \frac{1}{1+\exp(\pi_n' X_i)} \cdot f(Y\right.$$

$$\times \prod_{i|W_i=1,Z_i=1} \left(\frac{\exp(\pi_a' X_i)}{1+\exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{1a}) + \frac{1}{1+\exp(\pi_a' X_i)} \cdot f(Y\right.$$

$$\times \prod_{i|W_i=1,Z_i=0} \frac{\exp(\pi_a' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)} \cdot f(Y_i|X_i;\theta_{1a}).$$

Application: Angrist (1990) effect of military service

The simple ols regression leads to:

$$\widehat{\log(\text{earnings})}_i = \underset{(0079)}{5.4364} - \underset{(0.0167)}{0.0205} \cdot \widehat{\text{veteran}}_i$$

In Table 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.

		Z_i = 0	Z_i = 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 , under the non-defiers assumption. For example, the proportion of nevertakers is estimated as the conditional probability of $W_i = 0$ given $Z_i = 1$:

$$\Pr(\text{nevertaker}) = \frac{1915}{1915 + 865}.$$

		$W_i(0)$	
		0	1
$W_i(1)$	0	never-taker (0.6888)	defier (0)
	1	complier (0.1237)	always-taker (0.3112)

Estimated Average Outcomes by Treatment and Instrument

		Z_i	
		0	1
W_i	0	$\widehat{\mathbb{E}[Y]} = 5.4472$	$\widehat{\mathbb{E}[Y]} = 5.4028$
	1	$\widehat{\mathbb{E}[Y]} = 5.4076,$	$\widehat{\mathbb{E}[Y]} = 5.4289$

Not much variation by treatment status given instrument, but these comparisons are not causal under IV assumptions.

		$W_i(0)$	
		0	1
$W_i(1)$	0	$\widehat{\mathbb{E}[Y_i(0)]} = 5.4028$	defier (NA)
	1	$\widehat{\mathbb{E}[Y_i(0)]} = 5.6948$, $\widehat{\mathbb{E}[Y_i(1)]} = 5.4612$	$\widehat{\mathbb{E}[Y_i(1)]} = 5.4076$

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:

$$\begin{array}{rcccc} \widehat{\log(\text{earnings})}_i & = & 5.4836 & - & 0.2336 \cdot \widehat{\text{veteran}}_i \\ & & (0.0289) & & (0.1266) \end{array}$$

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.

Note that
$\mathbb{E}[Y_i(0)|n, X_i] < \mathbb{E}[Y_i(0)|c, X_i]$,
but also $\mathbb{E}[Y_i(1)|c, X_i] > \mathbb{E}[Y_i(1)|a, X_i]$

Compliers earn more than nevertakers when not serving, and more than always-takers when serving. Does not fit standard gaussian selection model.

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 (z_0, z_1).

Comparisons of estimates based on different instruments underly conventional tests of overidentifying restrictions in TSLS settings. An alternative interpretation of rejections in such testing procedures is therefore treatment effect heterogeneity.

Interpretation of IV Estimand

Suppose that monotonicity holds for all (z, z'), and suppose that the instruments are ordered in such a way that $p(z_{k-1}) \leq p(z_k)$, where $p(z) = \mathbb{E}[W_i|Z_i = z]$. 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:

$$\tau_{g(\cdot)} = \frac{\mathsf{Cov}(Y_i, g(Z_i))}{\mathsf{Cov}(W_i, g(Z_i))} = \sum_{k=1}^{K} \lambda_k \cdot \tau_{z_k, z_{k-1}},$$

$$\lambda_k = \frac{(p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]}{\sum_{k=1}^{K} p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]},$$

$$\pi_k = \mathsf{Pr}(Z_i = z_k).$$

These weights are nonnegative and sum up to one.

Marginal Treatment Effect

If the instrument is continuous, and $p(z)$ is continuous in z, we can define the limit of the local average treatment effects

$$\tau_z = \lim_{z' \downarrow z, z'' \uparrow z} \tau_{z',z''}.$$

Suppose we have a latent index model for the receipt of treatment:

$$W_i(z) = 1\{h(z) + \eta_i \geq 0\},$$

with the scalar unobserved component η_i independent of the instrument Z_i. Then we can define the marginal treatment effect $\tau(\eta)$ (Heckman and Vytlacil, 2005) 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

$$\tau(\eta) = \tau_z, \quad \text{with} \quad \eta = -h(z)).$$

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 $[0, 1]$, this restricts η to be in the interval $[\inf_z p(z), \sup_z p(z)]$, where $p(z) = \Pr(W_i = 1 | Z_i = z)$.

Now we can characterize various average treatment effects in terms of this limit. E.g.:

$$\tau = \int_\eta \tau(\eta) dF_\eta(\eta).$$

7. Multivalued Endogenous Variables

$$\tau = \frac{\mathsf{Cov}(Y_i, Z_i)}{\mathsf{Cov}(W_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]}.$$

Exclusion restriction and monotonicity:

$$Y_i(w)\ W_i(z)\ \perp\!\!\!\perp\ Z_i, \quad W_i(1) \geq W_i(0),$$

Then

$$\tau = \sum_{j=1}^{J} \lambda_j \cdot \mathbb{E}[Y_i(j) - Y_i(j-1)|W_i(1) \geq j > W_i(0)],$$

$$\lambda_j = \frac{\mathsf{Pr}(W_i(1) \geq j > W_i(0)}{\sum_{i=1}^{J} \mathsf{Pr}(W_i(1) \geq i > W_i(0)}.$$

with the weights λ_j estimable.

Illustration: Angrist-Krueger (1991) Returns to Educ.

$$\begin{array}{rcrcl} \widehat{\text{educ}_i} & = & 12.797 & - & 0.109 \cdot \text{qob}_i \\ & & (0.006) & & (0.013) \end{array}$$

$$\begin{array}{rcrcl} \widehat{\log(\text{earnings})_i} & = & 5.903 & - & 0.011 \cdot \text{qob}_i \\ & & (0.001) & & (0.003) \end{array}$$

The instrumental variables estimate is the ratio

$$\widehat{\beta}^{\text{IV}} = \frac{-0.1019}{-0.011} = 0.1020.$$

Weights $\gamma_j = \Pr(W_i(1) \geq j > W_i(0)$ can be estimated as

$$\widehat{\gamma}_j = \frac{1}{N_1} \sum_{i|Z_i=1} 1\{W_i \geq j\} - \frac{1}{N_0} \sum_{i|Z_i=0} 1\{W_i \geq j\}.$$

Figure 1: histogram estimate of density of years of education

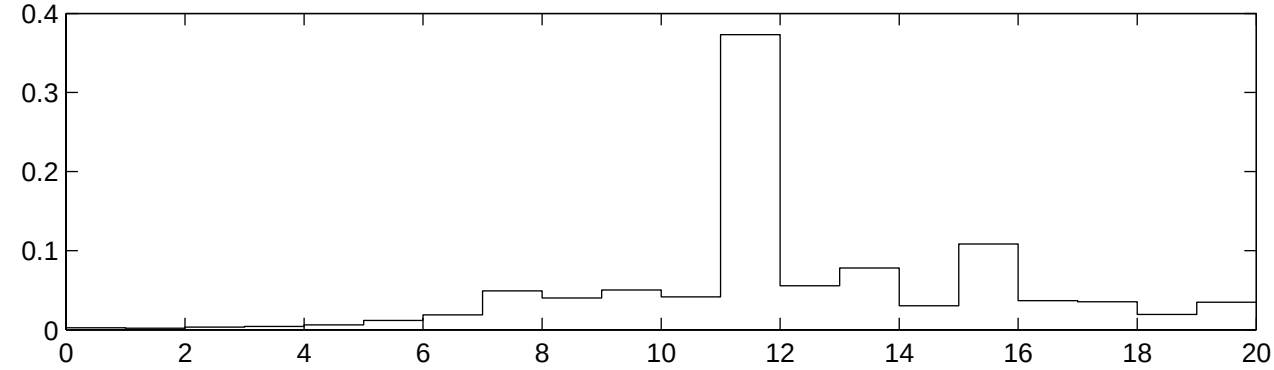


Figure 2: Normalized Weight Function for Instrumental Variables Estimand

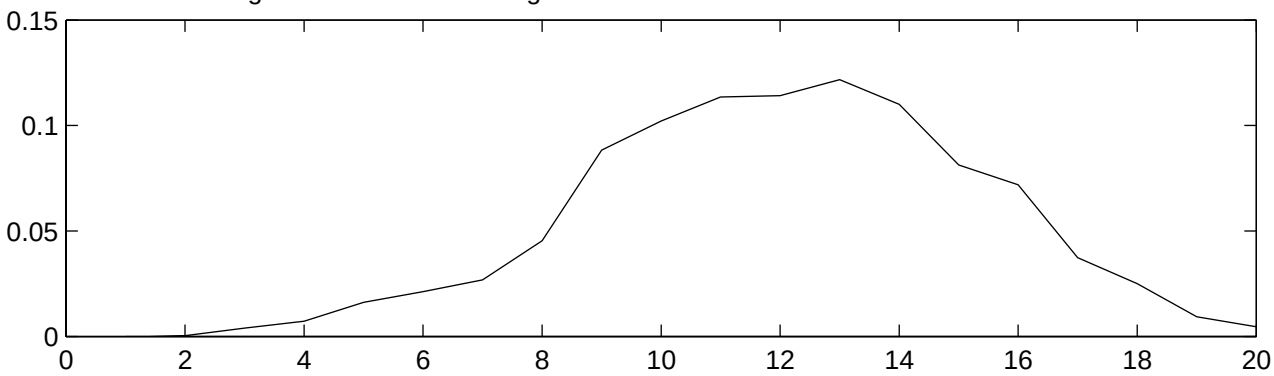


Figure 3: Unnormalized Weight Function for Instrumental Variables Estimand

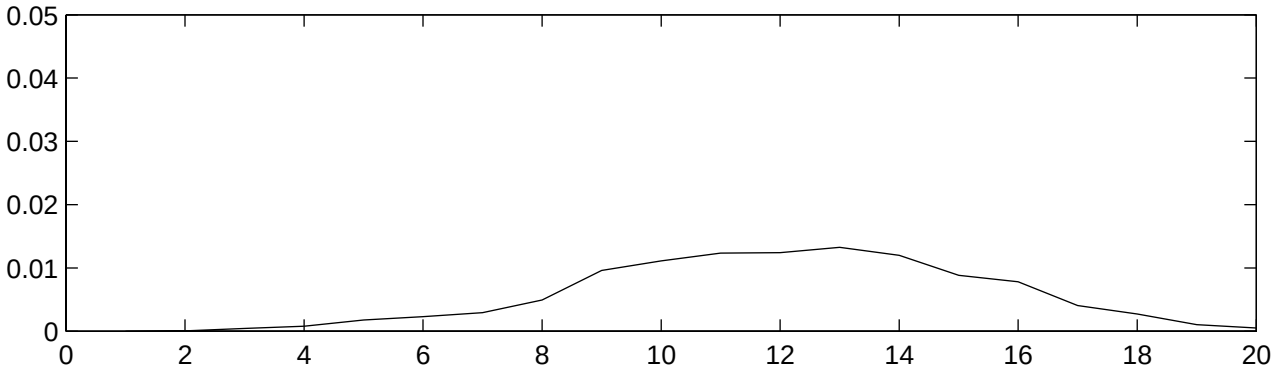


Figure 3: Education Distribution Function by Quarter