NBER WORKING PAPER SERIES

ROBUST INFERENCE FOR MISSPECIFIED MODELS CONDITIONAL ON COVARIATES

Alberto Abadie
Guido W. Imbens
Fanyin Zheng

Working Paper 17442
http://www.nber.org/papers/w17442

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2011

Financial support for this research was generously provided through NSF grant 0820361. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Robust Inference for Misspecified Models Conditional on Covariates
Alberto Abadie, Guido W. Imbens, and Fanyin Zheng
NBER Working Paper No. 17442
September 2011
JEL No. C01

ABSTRACT

Following the work by White (1980ab; 1982) it is common in empirical work in economics to report standard errors that are robust against general misspecification. In a regression setting these standard errors are valid for the parameter that in the population minimizes the squared difference between the conditional expectation and the linear approximation, averaged over the population distribution of the covariates. In nonlinear settings a similar interpretation applies. In this note we discuss an alternative parameter that corresponds to the approximation to the conditional expectation based on minimization of the squared difference averaged over the sample, rather than the population, distribution of a subset of the variables. We argue that in some cases this may be a more interesting parameter. We derive the asymptotic variance for this parameter, generally smaller than the White robust variance, and we propose a consistent estimator for the asymptotic variance.

Alberto Abadie
John F. Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
alberto_abadie@harvard.edu

Guido W. Imbens
Department of Economics
Littauer Center
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
and NBER
imbens@fas.harvard.edu

Fanyin Zheng
Harvard University
fzheng@fas.harvard.edu

1 Introduction

Following the seminal work by White (1980ab, 1982), researchers in economics routinely report standard errors that are robust to misspecification of the models that are being estimated. Müller (2011) gives the corresponding confidence intervals a Bayesian interpretation. A key feature of the approach developed by White (see also Eicker (1967) and Huber (1967)) is that in regression settings it focusses on the best linear predictor (blp) that minimizes the distance between the linear predictor and the true conditional expectation, averaged over the joint distribution of all variables, with a similar interpretation in nonlinear settings. However, in some regression settings it may be more appropriate to focus on the conditional best linear predictor (cblp) defined by averaging over the conditional distribution given the sample values of the covariates. The conceptual contribution of this note is to extend the White results to such settings. For a large class of estimators, including maximum likelihood and method of moment estimators, we first formally characterize the generalization to nonlinear models of the conditional best linear predictor. We then derive a large sample approximation to the variance of the least squares and method of moments estimators relative to this conditional estimand. In general, in misspecified models, this robust variance for the conditional estimand is smaller than the White robust variance. Finally, in the main technical contribution we propose a consistent estimator for this variance so that asymptotically valid confidence intervals can be constructed. The proposed estimator generalizes the variance estimator proposed by Abadie and Imbens (2006) for matching estimators. In correctly specified models the new variance estimator is simply an alternative to the standard White robust variance estimator. In misspecified models the new variance estimator is the first estimator for the robust variance for the conditional estimand.

Whether conditional or unconditional estimand should be the primary focus is context specific and we do not take the position that either the conditional or unconditional estimand is always appropriate. This is related to discussions about "random" versus "fixed" regressors. We discuss some examples to clarify the distinctions between the two and to make an argument for our view that in at least some settings the conditional estimand, corresponding to the fixed regression notion, is of interest. Most importantly, we argue that there is a clear choice to be made by the researcher that has direct implications for inference. In making this choice the researcher should bear in mind that the variance for the conditional estimand is generally smaller than that for the population or unconditional estimand, and thus tests for the former will generally have better power than tests for the latter.

The rest of this note is organized as follows. In Section 2 we discuss the conceptual issues raised by this note heuristically in a linear regression model setting. In Section 3 we discuss the motivation for the conditional estimand. Next, in Section 4 we present formal results covering least squares, maximum likelihood, and method of moments estimators. In Section 5 we apply the methods developed in this note to a data set collected by Imbens, Rubin and Sacerdote (2001). In Section 6 we present a small simulation study.

Section 7 concludes. The appendix contains the proofs.

2 The Conditional Best Linear Predictor

In this section we lay out some of the conceptual issues in this note informally in the setting of a linear regression model. In Section 4 we provide formal results, covering both this linear model setting and more general cases including maximum likelihood and method of moments.

Consider the standard linear model

$$Y_i = X_i'\theta + \varepsilon_i, \tag{2.1}$$

with Y_i the outcome of interest, X_i a K-vector of observed covariates, possibly including an intercept, and ε_i an unobserved error. Let $\mathbf{X}$, $\mathbf{Y}$, and ε be the $N \times K$ matrix with ith row equal to X_i', the N-vector with ith element equal to Y_i, and the N-vector with ith element equal to ε_i, respectively. Traditionally in this setting researchers assumed homoskedasticity, independence of the errors terms, and Normality of the error terms,

$$\varepsilon|\mathbf{X} \sim \mathcal{N}(0, \sigma^2 \cdot I_N),$$

where I_N is the $N \times N$ identity matrix. Under those assumptions the exact (conditional) distribution of the least squares estimator for θ,

$$\hat{\theta}_{\text{ols}} = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{Y}),$$

is Normal:

$$\hat{\theta}_{\text{ols}}|\mathbf{X} \sim \mathcal{N}\left(\theta, \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}\right).$$

However, the set of assumptions, linearity of the regression function, independence, homoskedasticity, and Normality of the error terms is often unrealistic. White (1980ab), Eicker (1967), and Huber (1967) considered the properties of the least squares estimator $\hat{\theta}_{\text{ols}}$ under much weaker assumptions. For the most general case one needs to define the estimand if the regression function is not linear. Suppose the sample $(Y_i, X_i)_{i=1}^N$ is a random sample from a large population satisfying some moment conditions. Let $\mu(x) = \mathbb{E}[Y_i|X_i = x]$ be the conditional expectation of Y_i given $X_i = x$, and let $\sigma^2(x)$ be the conditional variance. Even if this conditional expectation $\mu(x)$ is not linear, one might still wish to approximate it by a linear function $x'\theta$, and be interested in the value of the slope coefficient of this linear approximation, θ. Traditionally the optimal approximation is defined as the value of θ that minimizes the expectation of the squared difference between the outcomes and the linear approximation to the regression function.

This is generally referred to as the *best linear predictor*,[1] formally defined as

$$\theta_{\text{blp}} = \arg\min_{\theta} \mathbb{E}\left[(Y_i - X_i'\theta)^2\right]. \tag{2.2}$$

Writing this as

$$\theta_{\text{blp}} = \arg\min_{\theta} \mathbb{E}\left[(\mu(X_i) - X_i'\theta)^2\right] = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}\left(\mathbb{E}\left[X_i \mu(X_i)\right]\right),$$

shows that this can be interpreted as the value of θ that minimizes the discrepancy between the true regression function $\mu(x)$ and the linear approximation, weighted by the population distribution of the covariates.

White (1980ab) shows that, under some regularity conditions,

$$\sqrt{N} \cdot \left(\hat{\theta}_{\text{ols}} - \theta_{\text{blp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{blp}}\right),$$

where the normalized large sample variance is

$$\mathbb{V}_{\text{blp}} = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}\left(\mathbb{E}\left[(Y_i - X_i'\theta_{\text{blp}})^2 X_i X_i'\right]\right)\left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}. \tag{2.3}$$

White also proposed a consistent estimator for $\mathbb{V}_{\text{blp}}$,

$$\hat{\mathbb{V}}_{\text{blp}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\hat{\theta}_{\text{ols}})^2 X_i X_i'\right)\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}. \tag{2.4}$$

Using the White variance estimator $\hat{\mathbb{V}}_{\text{blp}}$ is currently standard practice in empirical work in economics. The bootstrap (Efron, 1982; Efron and Tibshirani, 1993) can also be used to construct confidence intervals for θ_{blp}.

In this note we explore an alternative linear approximation to the possibly nonlinear regression function $\mu(x)$. Instead of minimizing the marginal expectation of the squared difference between the outcomes and the regression function, we minimize this expectation conditional on the observed covariates. Define the *conditional best linear predictor* θ_{cblp} as

$$\theta_{\text{cblp}} = \arg\min_{\theta} \sum_{i=1}^{N} \mathbb{E}\left[\left.(Y_i - X_i'\theta)^2\right| \mathbf{X}\right]. \tag{2.5}$$

Denoting the N-vector with i-th element equal to $\mu(X_i)$ by $\mu(\mathbf{X})$, we can write θ_{cblp} as

$$\theta_{\text{cblp}} = \arg\min_{\theta} \sum_{i=1}^{N}\left(\mu(X_i) - X_i'\theta\right)^2 = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mu(\mathbf{X})\right),$$

[1]As far as we can tell, this term originates in the department of economics at Wisconsin, perhaps due to Art Goldberger (e.g., Goldberger 1991). The term is also used in Manski (1988). Earlier, Chamberlain (1982) used the terms "minimum mean square error linear predictor," and in the vector case, "multivariate linear predictor" for the same concept.

to stress the interpretation of θ_{cblp} as the best approximation to the true regression function, now with the weights based on the empirical distribution of the covariates. Both θ_{blp} and θ_{cblp} choose the linear approximation by minimizing the squared difference between the true regression function $\mu(x)$ and the linear approximation $x'\theta$. The difference between the two approximations is how they weight, as a function of the covariates, the squared difference between the regression function and the linear approximation for each x. The first approximation, leading to θ_{blp}, uses the population distribution of the covariates. The second approximation, leading to θ_{cblp}, uses the empirical distribution of the covariates.

We defer to Section 3 the question whether and why in a specific application θ_{blp} or θ_{cblp} might be the object of interest. In some applications we argue that θ_{blp} is unambiguously the estimand of interest. However, as discussed in detail in Section 3, we also think that in at least some applications θ_{cblp} may be of more interest than θ_{blp}. Therefore, given that the econometric literature has focused exclusively on inference estimands like θ_{blp}, we view the question of inference for θ_{cblp} as potentially of interest.

Next we point out the implications of the difference between θ_{blp} and θ_{cblp}. The first issue to note is that for point estimation it is irrelevant whether we are interested in θ_{blp} or θ_{cblp}. In both cases $\hat{\theta}_{\text{ols}}$ is the natural estimator. However, for inference it does matter whether we are interested in estimating θ_{blp} or θ_{cblp}, unless $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$ and the conditional expectation is linear. Consider the variance of the least squares estimator $\hat{\theta}_{\text{ols}}$, viewed as an estimator of θ_{cblp}. The exact (conditional) variance of $\hat{\theta}_{\text{ols}}$ is

$$\mathbb{V}\left(\hat{\theta}_{\text{ols}}\middle|\,\mathbf{X}\right) = \mathbb{E}\left[\left(\hat{\theta}_{\text{ols}} - \theta_{\text{cblp}}\right)\left(\hat{\theta}_{\text{ols}} - \theta_{\text{cblp}}\right)'\middle|\,\mathbf{X}\right] \tag{2.6}$$

$$= \frac{1}{N}\left(\mathbf{X}'\mathbf{X}/N\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sigma^2(X_i)X_iX_i'\right)\left(\mathbf{X}'\mathbf{X}/N\right)^{-1}.$$

Because $\hat{\theta}_{\text{ols}}$ is unbiased for θ_{cblp}, it follows that the marginal variance is the expected value of the conditional variance. Under random sampling this variance, normalized by the sample size, converges to

$$\mathbb{V}_{\text{cblp}} = \left(\mathbb{E}\left[X_iX_i'\right]\right)^{-1}\left(\mathbb{E}\left[\sigma^2(X_i)X_iX_i'\right]\right)\left(\mathbb{E}\left[X_iX_i'\right]\right)^{-1}, \tag{2.7}$$

and we have

$$\sqrt{N}\cdot\left(\hat{\theta}_{\text{ols}} - \theta_{\text{cblp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{cblp}}\right).$$

The key difference between the robust variance $\mathbb{V}_{\text{blp}}$ proposed by White and the robust variance $\mathbb{V}_{\text{cblp}}$ is the difference between the conditional variance $\sigma^2(X_i)$ in (2.9) and the expectation of the squared residual $\mathbb{E}[(Y_i - X_i'\theta_{\text{blp}})^2|X_i]$ in (2.3). For the overall variances we have

$$\mathbb{V}_{\text{blp}} = \mathbb{V}_{\text{cblp}} + N\cdot\mathbb{E}\left[\left(\theta_{\text{cblp}}(\mathbf{X}) - \theta_{\text{blp}}\right)\left(\theta_{\text{cblp}}(\mathbf{X}) - \theta_{\text{blp}}\right)'\right],$$

where the last expectation is over the distribution of θ_{cblp} as a function of $\mathbf{X}$. Note that in general $\mathbb{V}_{\text{blp}}$ exceeds $\mathbb{V}_{\text{cblp}}$. The difference arises from the misspecification in the regression function, that is, the difference between the conditional expectation and the best linear predictor, $\mu(x) - x\theta_{\text{blp}}$.

The final question we address in this section is how to estimate $\mathbb{V}_{\text{cblp}}$. The challenge is that the conditional variance function $\sigma^2(x)$ is generally unknown. Estimating this is straightforward in the case with discrete covariates. One can simply calculate the sample variance of Y_i at each distinct value of the covariates. Often that is not feasible, however, because some of the covariates are (close to) continuous. In such cases estimating $\sigma^2(x)$ consistently for all x would require nonparametric estimation involving bandwidth choices. Such an estimator would be more complicated than the White robust variance estimator which simply uses squared residuals to estimate the expectation of the squared errors. Here we build on work by Abadie and Imbens (2006) in the context of matching estimators to develop a general estimator for $\mathbb{V}_{\text{cblp}}$ that does not require consistent estimation of $\sigma^2(x)$, much like the White variance estimator does not consistently estimate $\mathbb{E}[(Y_i - X_i'\theta_{\text{blp}})^2|X_i = x]$ for all x. First define or the $M \times N$ matrix A with (i,j)th element equal to a_{ij} the norm $\|A\| = \max_{i,j} |a_{ij}|$. Next, define $\ell_X(i)$ to be the index of the unit closest to i in terms of X:

$$\ell_X(i) = \arg \min_{j \in \{1,\ldots,N\}, j \neq i} \|X_i - X_j\|,$$

Then our proposed variance estimator is

$$\widehat{\mathbb{V}}_{\text{cblp}} = \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \tag{2.8}$$

$$\cdot \left(\frac{1}{2N} \sum_{i=1}^{N} \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_X(i)} X_{\ell_X(i)} \right) \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_X(i)} X_{\ell_X(i)} \right)' \right) \cdot \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1}.$$

In Section 4 we show in a more general setting that this variance estimator is consistent for $\mathbb{V}_{\text{cblp}}$. An alternative estimator for $\mathbb{V}_{\text{cblp}}$ exploits the fact that the conditional variance of $\varepsilon_i X_i$ conditional on X_i is the same as X_i times the conditional variance of ε_i given X_i,

$$\widetilde{\mathbb{V}}_{\text{cblp}} = \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \cdot \left(\frac{1}{2N} \sum_{i=1}^{N} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{\ell_X(i)} \right)^2 X_i X_i' \right) \cdot \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1}.$$

Although in this linear regression case with the conditioning on all covariates both $\widehat{\mathbb{V}}_{\text{cblp}}$ and $\widetilde{\mathbb{V}}_{\text{cblp}}$ are consistent for $\mathbb{V}_{\text{cblp}}$, for nonlinear settings, or with conditioning on a subset of the covariates, only the first estimator $\widehat{\mathbb{V}}_{\text{cblp}}$ generalizes. To be specific, suppose that the covariate vector X_i can be partitioned as $X_i = (X_{1i}', X_{2i})'$ and correspondingly $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, and suppose we wish to estimate the variance conditional on $\mathbf{X}_1$ only. In this case the probability limit of the normalized variance for the least squares estimator is

$$\mathbb{V}_{\text{cblp}} = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1} \left(\mathbb{E}\left[\mathbb{V}\left(\varepsilon_i X_i \middle| X_{1i}\right)\right]\right) \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}. \tag{2.9}$$

Our proposed estimator for this conditional variance is

$$\widehat{\mathbb{V}}_{\text{cblp}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \tag{2.10}$$

$$\cdot \left(\frac{1}{2N}\sum_{i=1}^{N} \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{X_1}(i)} X_{\ell_{X_1}(i)}\right)\left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{X_1}(i)} X_{\ell_{X_1}(i)}\right)'\right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}.$$

This estimator is consistent for the conditional variance $\mathbb{V}_{\text{cblp}}$. In contrast, replacing $\hat{\varepsilon}_{\ell_X(i)}$ by $\hat{\varepsilon}_{\ell_{X_1}(i)}$ in the expression for $\widetilde{\mathbb{V}}_{\text{cblp}}$ would not lead to a consistent estimator for the variance.

In the remainder of this paper we will generalize the results in this section to maximum likelihood and method of moments settings, and state formal results concerning the large sample properties of the varaince estimators. In the general settings the estimators are no longer least squares estimators, and we will modify the terminology to reflect this. We will use θ_{pop} for population estimands that generalize the best linear predictor θ_{blp} in the regression case, and θ_{cond} for the conditional version that generalizes the conditional best linear predictor θ_{cblp} in the regression case.

3 Motivation for Conditional Estimands

In this section we address the question whether, when, and why the estimand conditional on the covariates may be of interest. We emphatically do not wish to argue that in all cases it is the conditional estimand is the appropriate object of interest. Rather, we wish to make the case, through four examples, that it depends on the context what the appropriate object is, and that at least in some settings, the conditional best linear predictor may be more appropriate or at least a reasonable alternative, to the standard, unconditional estimand.

One way to frame the question is in terms of different repeated sampling perspectives one can take. We can consider the distribution of the least squares estimator over repeated samples where we redraw the pairs X_i and Y_i (the random regressor case), or we can consider the distribution over repeated samples where we keep the values of X_i fixed and only redraw the Y_i (the fixed regressor case). Under general misspecification both the mean and variance of these two distributions will differ. The population estimand θ_{pop} is the approximate (in a large sample sense) average over the repeated samples when we redraw both X_i and Y_i, and θ_{cond} is the approximate average over the repeated samples where X_i is held fixed. Many introductory treatments of regression analyses briefly introduce the fixed and random regressor concepts, with a variety of opinions on what the most relevant perspective is. Wooldridge writes that "reliance on fixed regressors ... can have unintended consequences. ... Because our focus is on asymptotic analysis, we have the luxury of allowing for random explanatory variables throughout the book" (Wooldridge, 2002, p10-11). Cameron and Trivedi write "The fixed regressors assumption is rarely appropriate for microeconometrics data" (Cameron and Trivedi, 2005, p. 77). Stock and Watson (2003) focus on the random regressor case, arguing that "the i.i.d. assumption is a reasonable one for many data collection schemes" but acknowledging that "Not all sampling schemes produce i.i.d. observations on (X_i, Y_i)" (Stock and Watson, 2003, p. 105). Goldberger (1991) takes a different position, assuming "$\mathbf{X}$ nonstochastic, which says that the elements of $\mathbf{X}$ are constants, that is, degenerate random variables. Their values are fixed in repeated samples ..." (Goldberger, p. 164). These discussions are in the context of correctly specified regression models, however, where the averages of the distributions under the two repeated sampling perspectives coincide, and their variances agree in large samples. A point that has not received attention in the literature is that under general misspecifiaction, the random versus fixed regressor distinction has implications for inference that do not vanish with the sample size.

Another point is that the sole difference between the population and conditional estimands is the weight function used to measure the difference between the model and the true data generating process. For the population estimand the weight function depends on the population distribution of the potential conditioning variables, and for the conditional estimand it is the sample distribution of these variables. Because the population distribution of these variables, unlike the sample distribution, is unknown, in general there is more uncertainty about the population estimand. Thus, in practial terms, focusing on the conditional estimand θ_{cond} leads to smaller standard errors than focusing on the population estimand θ_{pop}.

Example I (Finite versus Infinite Population)

In the first example we want to argue that if the sample is a random sample from a large population θ_{pop} is of more interest than θ_{cond}, whereas in the case where the sample is equal to the population, the conditional estimand θ_{cond} is of more interest.

Consider estimation of the average effect of a binary treatment. Each unit in the population is characterized by two potential outcomes $Y_i(\text{c})$ and $Y_i(\text{t})$, and a binary covariate $X_i \in \{\text{f}, \text{m}\}$. Let $W_i \in \{\text{c}, \text{t}\}$ denote the treatment received and $Y_i = Y_i(W_i)$ the realized outcome. Assume assignment to treatment is random given X_i. Define, for $w = \text{c}, \text{t}$ and $x = \text{f}, \text{m}$,

$$\mu(x, w) = \mathbb{E}[Y_i(w)|X_i = x], \qquad \text{and} \quad \tau(x) = \mu(x, \text{t}) - \mu(x, \text{c}).$$

Let N_{xw} be the number of units in the sample with $X_i = x$ and $W_i = w$, let q be the population share of the $X_i = f$ types and $\hat{q}$ the sample share:

$$q = \mathbb{E}[W_i], \qquad \text{and} \quad \hat{q} = \frac{1}{N}\sum_{i=1}^{N} W_i = \frac{N_{\text{fc}} + N_{\text{ft}}}{N_{\text{fc}} + N_{\text{ft}} + N_{\text{mc}} + N_{\text{mt}}}.$$

Consider two estimands, first the population average treatment effect,

$$\theta_{\text{pop}} = \mathbb{E}\left[Y_i(\text{t}) - Y_i(\text{c})\right] = q \cdot \tau(\text{f}) + (1-q) \cdot \tau(\text{m}),$$

where q is the population fraction of $X_i = \text{f}$ types, and second, the conditional average effect,

$$\theta_{\text{cond}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i(\text{t}) - Y_i(\text{c}) \middle| X_i\right] = \hat{q} \cdot \tau(\text{f}) + (1-\hat{q}) \cdot \tau(\text{m}).$$

What is the rationale for focusing on θ_{pop} versus θ_{cond}? If the sample is a random sample from a large population, it seems natural to focus on the population average treatment effect θ_{pop} as the object of interest. On the other hand, suppose the sample is the entire population. For example, the population could be the 50 states of the United States in which case we might have observations on all 50 states. The covariate could be an indicator for a state being on the coast versus inland. In that case it would appear reasonable to keep fixed the number of coastal versus inland states, rather than view the share of coastal states as random. That perspective suggests focusing on θ_{cond} rather than θ_{pop} in cases where the sample is the entire population. We may still wish to use large sample approximations, but focus on estimation of θ_{cond} rather than θ_{pop}.

Note that if we observe the entire population in this case, we cannot interpret the uncertainty in the estimator as due to sampling variation in the units. Instead we can interpret the uncertainty in the estimator as due to random variation in the treatment assignment W_i (see, for example, Neyman, 1923, 1990). To justify large sample approximation, however, we will resort to a random sampling argument.

Example II (Convenience Sample)

In the second example we want to make the case that sometimes there is intrinsically no more interest in θ_{pop} than θ_{cond} because neither the weighting scheme corresponding to the population distribution, nor the weighting scheme corresponding to the empirical distribution function, is obviously of primary interest.

Consider the study of lottery winners by Imbens, Rubin and Sacerdote (2001). We use data from this study in Section 5. Imbens, Rubin and Sacerdote surveyed individuals who won large prizes in the lottery. Using a standard life-cycle model of labor supply they focus on linear regressions of subsequent labor earnings on the annual prize and some additional covariates including prior earnings. The coefficient on the prize in this linear regression can be interpreted as the marginal propensity to consume out of unearned income, an economically meaningful parameter (e.g., Pencavel, 1986). Even if the conditional expectation as a function of the prize is nonlinear, it may still be interesting to focus on the coefficient in the linear regression, partly because it facilitates comparison across studies. The question is whether the linear approximation should be based on weighting the squared difference between the true regression function and the linear predictor by the population or empirical distribution of lottery prizes. There does not appear to be a strong substantive argument for preferring one weighting function (and thus the corresponding estimand) over the other.

Example III (Experimental Design)
Karlan and List (2009) carried out an experimental evaluation of incentives for charitable giving. Among the results Karlan and List report are probit regression estimates where the object of interest is the regression coefficient on the indicator for being offered a matching incentive for charitable giving. The specification of the probit regression function also includes characteristics of the matching incentives.

In this case the difference between $\mathbb{V}_{\text{pop}}$ and $\mathbb{V}_{\text{cond}}$ is that $\mathbb{V}_{\text{pop}}$ takes into account sampling variation in $\hat{\theta}$ due to variation in the sample values of the matching incentives over the repeated samples, whereas $\mathbb{V}_{\text{cond}}$ conditions on these values. Given that the distribution of these incentives in this experiment is fixed by the researchers there appears to be no reason to take this uncertainty into account, and we submit that the appropriate measure of uncertainty is $\mathbb{V}_{\text{cond}}$ rather than $\mathbb{V}_{\text{pop}}$.

Example IV (Average Derivative in Sample versus Population)
In the last example we again want to make the case that there is no compelling reason to prefer one estimand to the other.

Suppose one estimates a parametric binary response model, say a probit model with $\Pr(Y_i = 1|X_i) = \Phi(X_i'\theta)$, where $\Phi(a) = \int_{-\infty}^{a}(1/\sqrt{2\pi})\exp(-z^2/2)dz$. (The same argument would apply to other nonlinear parametric models.) Parameter estimates are difficult to interpret for such models, and often researchers report derivatives of the conditional expectation of Y_i given X_i with respect to X_i to facilitate comparisons with other models. For the probit model the derivative of the conditional expectation is $\phi(x'\theta)\cdot\theta$, where $\phi(a) = \partial\Phi(a)/\partial a = \exp(-a^2/2)/\sqrt{2\pi}$. In nonlinear models the value of the derivative depends on the value of the covariates, so often researchers report the average derivative evaluated at the estimated parameters:

$$\hat{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\phi(X_i'\hat{\theta})\cdot\hat{\theta}.$$

The variance of this estimator $\hat{\gamma}$ for the average derivative differs depending on whether we condition on the covariates or not. The two estimands are

$$\gamma_{\text{cond}} = \frac{1}{N}\sum_{i=1}^{N}\phi(X_i'\theta)\cdot\theta, \qquad \text{and} \quad \gamma_{\text{pop}} = \mathbb{E}\left[\phi(X_i'\theta)\cdot\theta\right].$$

Because the average derivative is presented primarily as a more interpretable parameter than θ itself, taking into account the uncertainty in the distribution of the covariates that is averaged over may not serve any useful purpose, suggesting that γ_{cond} may be just as relevant as γ_{pop}.

4 Inference for Conditional Estimands

In this section we present the main formal results of the paper, covering linear regression, maximum likelihood, and method of moments estimators. We cover settings where we condition on the full set of regressors as well as cases where we condition on a subset of the regressors.

Suppose we have a random sample of size N of a pair of random vectors, (X_i, Y_i), $i = 1, \dots, N$. Let K_X and K_Y be the dimensions of X_i and Y_i, and let $\mathbf{X}$ and $\mathbf{Y}$ be the $N \times K_X$ and $N \times K_Y$ matrices with i-th rows equal to X_i' and Y_i' respectively. We are interested in a finite dimensional parameter θ, defined as some function of the joint distribution of (X_i, Y_i). Under some economic model it follows that

$$\mathbb{E}\left[\psi(Y_i, X_i, \theta)\right] = 0. \tag{4.1}$$

The model may have additional implications beyond this moment condition, but these are not used for estimation. For example, it may be the case that the conditional moment has expectation zero,

$$\mathbb{E}\left[\psi(Y_i, X_i, \theta) \middle| X_i\right] = 0.$$

Alternatively, we may have specified the joint distribution of Y_i and X_i, in which case $\psi(y, x, \theta)$ could equal to the score function. In that case the model has the additional implication that the expected value of the derivatives of $\psi(y, x, \theta)$ with respect to θ is equal to the expected value of the second moments of $\psi(y, x, \theta)$. Based only on (4.1), and not on any other implications of the motivating model, we may wish to estimate θ by solving

$$\sum_{i=1}^{N} \psi(Y_i, X_i, \hat{\theta}) = 0.$$

We are interested in the properties of the estimator $\hat{\theta}$ under general misspecification of the model that motivated the moment condition.

The standard approach (Hansen, 1984; Newey and McFadden, 1994; Wooldridge, 2002) focuses on the value θ_{pop} that solves

$$\mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}})\right] = 0.$$

If the pairs (X_i, Y_i), for $i = 1, \dots, N$ are independent and identically distributed, then under regularity conditions,

$$\sqrt{N}\left(\hat{\theta} - \theta_{\text{pop}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{gmm,pop}}\right), \qquad \text{where } \mathbb{V}_{\text{gmm,pop}} = \left(\Gamma' \Delta^{-1} \Gamma\right)^{-1},$$

with

$$\Gamma = \mathbb{E}\left[\frac{\partial}{\partial \theta'} \psi(Y_i, X_i, \theta_{\text{pop}})\right], \qquad \text{and } \Delta = \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}}) \psi(Y_i, X_i, \theta_{\text{pop}})'\right].$$

Now we focus on the conditional estimand. Define θ_{cond} as the solution to

$$\mathbb{E}\left[\sum_{i=1}^{N} \psi(Y_i, X_i, \theta) \middle| \mathbf{X}\right] = 0. \tag{4.2}$$

Note that implicitly θ_{cond} is a function of $\mathbf{X}$. If the original model implied that the conditional expectation of $\psi(Y_i, X_i, \theta)$ given X_i is equal to zero, then $\theta_{\text{cond}} = \theta_{\text{pop}}$, but this need not hold in general. The motivation for the estimand is the same as in the best-linear-predictor case. In cases where the model implies a conditional moment condition, but we are concerned about misspecification, we may wish to focus on the value for θ that minimizes the discrepancy between $\mathbb{E}[\psi(Y_i, X_i, \theta)|X_i]$ and zero. We can weight the discrepancy by the population distribution of the X_i's, or by the empirical distribution. The conditional estimand corresponds to the case where the weights are based on the empirical distribution function.

We make the following assumptions. These are closely related to standard assumptions used for establishing asymptotic properties for moment-based estimators. See for example Newey and McFadden (1994).

Assumption 1 *(X_i, Y_i), for $i = 1, \ldots, N$, are independent and identically distributed. The support of X_i is a compact subset of $\mathbb{R}^L$.*

Assumption 2 *(i) The K-component vector of moment conditions $\psi(y, x, \theta)$ is continuously differentiable in θ for $\theta \in \Theta$ with Θ a compact subset of $\mathbb{R}^K$, with both $\psi(y, x, \theta)$ and its derivative with respect to θ, $\frac{\partial}{\partial \theta'}\psi(y, x, \theta)$, continuous in x and y for all $\theta \in \Theta$, (ii) there is a unique value $\theta_{\text{pop}} \in \text{int}(\Theta)$ such that $\mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}})\right] = 0$, (iii) Δ and Γ are finite and full rank, and (iv) $\mathbb{E}\left[\sup_{\theta\in\Theta} \left\|\frac{\partial}{\partial \theta'}\psi(Y_i, X_i, \theta)\right\|\right]$, and for some positive δ, $\mathbb{E}\left[\sup_{\theta\in\Theta} \left\|\psi(Y_i, X_i, \theta)\right\|^{2+\delta}\right]$ are finite.*

Theorem 1 *Suppose Assumptions 1 and 2 hold. Then (i), $\hat{\theta} - \theta_{\text{cond}} = o_p(1)$, and (ii)*

$$\sqrt{N} \cdot \left(\hat{\theta} - \theta_{\text{cond}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{gmm,cond}}\right),$$

where

$$\mathbb{V}_{\text{gmm,cond}} = \left(\Gamma' \Delta_{\text{cond}}^{-1} \Gamma\right)^{-1}, \qquad \text{and} \quad \Delta_{\text{cond}} = \mathbb{E}\left[\mathbb{V}\left(\psi(Y_i, X_i, \theta_{\text{pop}})\right) \middle| X_i\right].$$

If also $\mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}}) \middle| X_i = x\right] = 0$ for all x, then (iii),

$$\theta_{\text{cond}} = \theta_{\text{pop}}, \qquad \text{and} \quad \sqrt{N}\left(\hat{\theta} - \theta_{\text{pop}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{gmm,cond}}\right).$$

Proof: See Appendix.

Let us consider an additional example to illustrate the differences between the two variances. This example is related to the discussion in Chow (1984).

Example V (Maximum Likelihood Estimation)
Suppose we specify the conditional distribution of Y_i given X_i as $f(y|x;\theta)$. We estimate the model by maximum likelihood:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln f(Y_i|X_i;\theta).$$

The normalized asymptotic variance under correct specification, and under some regularity conditions, is equal to the inverse of the information matrix $\mathcal{I}_\theta^{-1}$, where

$$\mathcal{I}_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f(Y_i|X_i;\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta) \cdot \frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta)'\right].$$

White (1982) analyzed the properties of the estimator under general misspecification of the conditional density. Let

$$\theta_{\text{pop}} = \arg\max_{\theta} \mathbb{E}\left[\ln f(Y_i|X_i;\theta)\right].$$

Then White (1982) showed that under general misspecification,

$$\hat{\theta} \xrightarrow{p} \theta_{\text{pop}}, \qquad \text{and} \quad \sqrt{N}\cdot\left(\hat{\theta} - \theta_{\text{pop}}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\Gamma'\Delta^{-1}\Gamma\right)^{-1}\right),$$

with

$$\Gamma = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f(Y_i|X_i;\theta_{\text{pop}})\right], \quad \text{and } \Delta = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}}) \cdot \frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}})'\right].$$

The conditional version of the estimand under general misspecification is

$$\theta_{\text{cond}} = \arg\max_{\theta} \sum_{i=1}^{N} \mathbb{E}\left[\ln f(Y_i|X_i;\theta)|\, X_i\right],$$

where the expectation is taken only over Y_i. Theorem 1 implies that

$$\sqrt{N}\cdot\left(\hat{\theta} - \theta_{\text{cond}}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\Gamma'\Delta_{\text{cond}}^{-1}\Gamma\right)^{-1}\right),$$

where

$$\Delta_{\text{cond}} = \mathbb{E}\left[\mathbb{V}\left(\frac{\partial}{\partial\theta} \ln f(Y_i|X_i,\theta_{\text{pop}})\right)\middle| X_i\right].$$

If the model is correctly specified, then $\Delta = \Delta_{\text{cond}}$, but if the model is misspecified then

$$\mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i,\theta_{\text{pop}})\right] = 0,$$

but it is not true that for all x

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(Y_i|X_i,\theta_{\text{pop}})\middle| X_i = x\right] = 0,$$

implying that in general $\Delta - \Delta_{\text{cond}}$ is positive semi-definite. $\square$

Next, we consider estimation of the variance in the general case. Estimation of Γ is the same as for the population estimand:

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}\psi(Y_i, X_i, \hat{\theta}).$$

The key question concerns estimation of Δ_{cond}. Our proposed estimator matches each unit to the closest unit in terms of X_i, and then differences the values of the moment function:

$$\hat{\Delta}_{\text{cond}} = \frac{1}{2N}\sum_{i=1}^{N}\left(\psi(Y_i, X_i, \hat{\theta}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \hat{\theta})\right)\left(\psi(Y_i, X_i, \hat{\theta}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \hat{\theta})\right)'.$$

We then combine these estimates to get an estimator for the variance for the conditional estimand:

$$\widehat{\mathbb{V}}_{\text{gmm,cond}} = \left(\hat{\Gamma}'\hat{\Delta}_{\text{cond}}^{-1}\hat{\Gamma}\right)^{-1}.$$

Theorem 2 (Conditional Variance for Method of Moments Estimators) *Suppose Assumptions 1 and 2 hold. Then*

$$\widehat{\mathbb{V}}_{\text{gmm,cond}} \xrightarrow{p} \mathbb{V}_{\text{gmm,cond}}.$$

Proof: See Appendix.

5 An Application to the Imbens-Rubin-Sacerdote Lottery Data

To illustrate the issues raised in this note we look at some data previously analyzed by Imbens, Rubing and Sacerdote (2001). Imbens, Rubin and Sacerdote collected data on individuals who played the lottery in the mid-eighties. Here we focus on a subset of their data for 194 individuals who won large prizes. We use three variables, the yearly prize won by each individual, the average of yearly earnings over six years prior to winning the lottery and the average of yearly earnings over the six years after winning the lottery. Table 1 reports some summary statistics.

Using a standard life-cycle model for consumption and savings Imbens, Rubin and Sacerdote estimate a linear model relating subsequent labor earnings to prior earnings and the yearly prize. The coefficient on the yearly prize can be interpreted as the propensity to earn out of unearned income, an economically meaningful parameter (e.g., Pencavel, 1986). Following the Imbens-Rubin-Sacerdote specification we focus on the regression function

$$Y_i = \theta_0 + \theta_1 \cdot P_i + \theta_2 \cdot X_i + \varepsilon_i,$$

where Y_i is the average of post-lottery earnings, X_i is the average of pre-lottery earnings, and P_i is the yearly prize. As we discussed in Example II in Section 3, we may wish to estimate a linear regression function even if one does not believe the conditional expectation is exactly linear. The question then arises how to approximate the conditional expectation by a linear function: averaging the squared difference between the conditional expectation and the linear approximation over the population or over the sample distribution of the covariates. Arguably one is interested in estimating a representative value for the marginal propensity to earn out of unearned income, acknowledging that this parameter may vary between individuals, and, for a given individual, may vary by income levels. There is in our view no compelling argument that the population distribution in the lottery sample, or the sample distribution is closer to being representative of the population of interest.

In Table 2 we report estimates for this regression function, with both the conventional robust standard errors and the standard error for the conditional estimand.

6 A Small Simulation Study

In this section we assess the small sample properties of the variance estimators. We center our simulation study around the lottery data set. We focus on estimating a linear regression function

$$Y_i = \theta_0 + \theta_1 \cdot P_i + \theta_2 \cdot X_i + \varepsilon_i.$$

The joint distribution of the two covariates in the population is

$$\begin{pmatrix} P_i \\ X_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 32.0 \\ 12.1 \end{pmatrix}, \begin{pmatrix} 443.1 & 54.8 \\ 54.8 & 124.9 \end{pmatrix} \right).$$

The means and variances of this joint distribution were estimated on the lottery data. The conditional distribution of Y_i given P_i and X_i is normal:

$$Y_i | X_i, P_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right),$$

where

$$\mu_i = 6.46 - 0.13 \times P_i + 0.75 \times X_i + \frac{\delta}{1000} \times \left(P_i^2 + 1420 - 87 \times P_i - 5 \times X_i\right),$$

and

$$\ln \sigma_i^2 = 2.611 - 0.012 \cdot P_i + 0.070 \cdot X_i.$$

Again the parameter values are motivated by the lottery data. A non-zero value for δ makes the model nonlinear. We use two values for δ. In the first design we fix $\delta = 1.43$ corresponding most closely to the lottery data. In the second design we use a larger value, $\delta = 14.3$.

Table 3 presents the results. We focus on the coefficient on the prize, θ_1. For both designs we report the the average of the population and conditional standard error for $\hat{\theta}_1$, and four coverage rates. First the coverage frequency of the conventional (White standard error based) 95% confidence interval for θ_{pop}. This coverage should be 0.95. Next, the fequency with which the same confidence interval covers θ_{cond}. This should be more than 0.95. In the next row we report the coverage rates for confidence intervals based on the conditional standard errors. Now the coverage for θ_{pop} could be less than 0.95, but the coverage for θ_{cond} should be 0.95. In the first design the model is too close to being linear to detect these effects, and all coverage rates are close to 0.95. In the second design the average conditional standard error is about 10% less than the average unconditional (White) standard error, and this shows up in the coverage rates of the confidence intervals. The confidence interval based on White standard errors covers θ_{pop} with probability 0.95, and θ_{cond} with probability 0.97, and the confidence interval based on the conditional standard error covers θ_{cond} with probability 0.94, and θ_{pop} with probability 0.91.

7 Conclusion

In this note we discuss inference for conditional estimands in misspecified models. Following the work by White (1980ab, 1982) it is common in empirical work to report robust standard errors. These robust standard errors are valid for the population value of the estimator given random sampling. We show that if one is interested in the conditional estimand, conditional on all or a subset of the variables, robust standard errors are generally smaller than the White robust standard errors. We derive a general characterization of the variance for the conditional estimand and propose a consistent estimator for this variance. We argue that in some settings the conditional estimand may be of more interest than the unconditional one.

Appendix: Proofs of Theorems

Proof of Theorem 1: Assumptions 1 and 2 imply the assumptions in Theorems 2.6 and 3.4 in Newey and McFadden (1994). Their results imply that $\hat{\theta}$ is consistent for θ_{pop}, and that $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1})$. To prove part (i), that $\hat{\theta} - \theta_{\text{cond}} = o_p(1)$, we first prove that $\theta_{\text{cond}} - \theta_{\text{pop}} = o_p(1)$. Then, by the triangle inequality, because $\hat{\theta} - \theta_{\text{pop}} = o_p(1)$ by Theorem 2.6 in Newey McFadden (1994), it follows that $\hat{\theta} - \theta_{\text{cond}} = o_p(1)$. Define $\rho(x, \theta) = \mathbb{E}[\psi(Y_i, X_i, \theta)|X_i = x]$, so that $\mathbb{E}[\rho(X_i, \theta_{\text{pop}})] = 0$, and θ_{cond} solves

$$\frac{1}{N}\sum_{i=1}^{N} \rho(X_i, \theta) = 0.$$

Hence θ_{cond} can be thought of as a method of moments estimator for θ_{pop} with moment condition $\rho(X_i, \theta)$. Because of Assumption 2 it follows that $\rho(x, \theta)$ satisfies the conditions for consistency of the method of moments estimator in Theorem 2.6 in Newey and McFadden (1994), and thus $\theta_{\text{cond}} - \theta_{\text{pop}} = o_p(1)$.

Next we prove part (ii) of the theorem. Theorem 3.4 in Newey and McFadden also implies that $\theta_{\text{cond}} - \theta_{\text{pop}} = O_p(N^{-1/2})$. Because $\hat{\theta} - \theta_{\text{pop}} = O_p(N^{-1/2})$ it follows by the triangle inequality that $\theta_{\text{cond}} - \theta_{\text{pop}} = O_p(N^{-1/2})$. By a mean value theorem it follows that

$$0 = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi(Y_i, X_i, \hat{\theta}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}}) + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta'}\psi(Y_i, X_i, \tilde{\theta})\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}),$$

for some intermediate value $\tilde{\theta}$. Because $\theta_{\text{cond}} - \theta_{\text{pop}} = o_p(1)$ and $\hat{\theta} - \theta_{\text{pop}} = o_p(1)$, it follows that the intermediate value $\tilde{\theta}$ satisfies $\tilde{\theta} - \theta_{\text{pop}} = o_p(1)$, and thus $\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta'}\psi(Y_i, X_i, \tilde{\theta}) = \Gamma + o_p(1)$. Thus

$$0 = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}}) + \Gamma\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}) + o_p(1),$$

and therefore

$$\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}) = \Gamma^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}}) + o_p(1). \tag{A.1}$$

Next we show that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}}) \sim N(0, \Delta_{\text{cond}}). \tag{A.2}$$

Because θ_{cond} is the solution to

$$\mathbb{E}\left[\left.\sum_{i=1}^{N} \psi(Y_i, X_i, \theta)\right| \mathbf{X}\right] = 0,$$

θ_{cond} is a function of $\mathbf{X}$, i.e. $\theta_{\text{cond}} = \theta_{\text{cond}}(\mathbf{X})$. Therefore conditional on $\mathbf{X}$ the $\psi(Y_i, X_i, \theta_{\text{cond}})$ are independent, but not identically distributed. For ease of exposition we focus on the case where $K = 1$. We first apply the Lyapunov Central Limit Theorem to show that

$$\frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \psi(Y_i, X_i, \theta_{\text{cond}}) - \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| \mathbf{X} \right] \right\}}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \mathbb{V}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| \mathbf{X} \right]}} \xrightarrow{d} \mathcal{N}\left(0, 1\right). \tag{A.3}$$

The Lyapounov condition $\mathbb{E}[|\psi(Y_i, X_i, \theta_{\text{cond}}) - \mathbb{E}[\psi(Y_i, X_i, \theta_{\text{cond}})|\mathbf{X}]|^{2+\delta}] < \infty$ for some positive δ follows from Assumption 2(iv).
Because $\sum_{i=1}^{N} \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| \mathbf{X} \right] = 0$ by the definition of θ_{cond}, it follows that the numerator in (A.3) simplifies to

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \psi(Y_i, X_i, \theta_{\text{cond}}) - \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| \mathbf{X} \right] \right\}$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}}) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| \mathbf{X} \right] = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}}).$$

The denominator in (A.3) converges in probability

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{V}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| \mathbf{X} \right] \xrightarrow{p} \mathbb{E}\left\{ \mathbb{V}\left[\psi(Y_i, X_i, \theta_{\text{cond}}) \middle| X_i \right] \right\} = \mathbb{V}_{\text{cond}}.$$

In combination with (A.3) this implies

$$\frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Y_i, X_i, \theta_{\text{cond}})}{\sqrt{\mathbb{V}_{\text{cond}}}} \xrightarrow{d} \mathcal{N}\left(0, 1\right),$$

and thus (A.2) follows.
Combining (A.1) and (A.2) implies

$$\sqrt{N}\left(\hat{\theta} - \theta_{\text{cond}}\right) \sim N\left(0, \left(\Gamma' \Delta_{\text{cond}}^{-1} \Gamma\right)^{-1}\right)\right).$$

finishing the proof of part (ii).
If also $\mathbb{E}\left[\psi(Y_i, X_i, \theta) \middle| X_i \right] = 0$, θ_{cond} is the solution to

$$\mathbb{E}\left[\psi(Y_i, X_i, \theta)\right] = \mathbb{E}\left[\sum_{i=1}^{N} \psi(Y_i, X_i, \theta) \middle| \mathbf{X} \right] = 0.$$

Then $\theta_{\text{cond}} = \theta_{\text{pop}}$, and part ($iii$) of the theorem follows. $\square$

Next we state a useful lemma from Abadie and Imbens (2010).

Lemma A.1 (Lemma 1, Abadie and Imbens (2010, page 180)) *Suppose that $W_1, W_2, \ldots$ is a sequence with $W_i \in \mathbb{W}$ where $\mathbb{W}$ a compact subset of $\mathbb{R}^K$. Then*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \left\| W_i - W_{\ell_W(i)} \right\|^2 = 0.$$

Lemma A.2 (Average Conditional Moments) *Let (V_i, W_i), $i = 1, \ldots, N$, be a sequence of independent, identically distributed random vectors, with dimension K_V and K_W respectively, and compact support for W_i. For some positive integer n, and for $j = 1, 2, \ldots, n$, let $\mu_j(w) = \mathbb{E}[V_i^j | W_i = w]$ be Lipschitz in w with constant C_j, and suppose all moments of V_i up to the $2n$-th moment exist. Then for all nonnegative k, m such that* $\min(k, m) \leq n$,

$$\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m \xrightarrow{p} \mathbb{E}\left[\mathbb{E}\Big(V_i^k\Big| W_i\Big) \cdot \mathbb{E}\Big(V_i^m | W_i\Big)\right].$$

Proof of Lemma A.2: We focus on the scalar case. The vector case can be shown by the same argument. First we show

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\Big(V_i^k\Big| W_i\Big) \cdot \mathbb{E}\Big(V_i^m\Big|W_i\Big)\right]\right] = o(1). \tag{A.4}$$

Because V_i and $V_{\ell_W(i)}$ are independent conditional on $\mathbf{W} = (W_1, ..., W_N)'$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\mathbb{E}\left[V_i^k \cdot V_{\ell_W(i)}^m\Big| \mathbf{W}\right]\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\mathbb{E}(V_i^k\Big| \mathbf{W}) \cdot \mathbb{E}\left(V_{\ell_W(i)}^m\Big| \mathbf{W}\right)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\mathbb{E}(V_i^k\Big| W_i) \cdot \mathbb{E}\left(V_{\ell_W(i)}^m\Big| W_{\ell_W(i)}\right)\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mu_k(W_i) \cdot \mu_m\left(W_{\ell_W(i)}\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\mu_k(W_i) \cdot \left[\mu_m(W_i) + \mu_m\left(W_{\ell_W(i)}\right) - \mu_m(W_i)\right]\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mu_k(W_i) \cdot \mu_m(W_i)\right] + \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\mu_m(W_i)\left[\mu_m\left(W_{\ell_W(i)}\right) - \mu_m(W_i)\right]\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right) \cdot \mathbb{E}\left(V_i^m|W_i\right)\right] + \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\mu_m(W_i)\left[\mu_m\left(W_{\ell_W(i)}\right) - \mu_m(W_i)\right]\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right) \cdot \mathbb{E}\left(V_i^m|W_i\right)\right]\right]\right| \\
&\qquad \leq \left|\mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\mu_m(W_i)\left[\mu_m\left(W_{\ell_W(i)}\right) - \mu_m(W_i)\right]\right\}\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}|\mu_m(W_i)|\cdot\left|\mu_m\left(W_{\ell_W(i)}\right)-\mu_m(W_i)\right|\right\}\\
&\leq \sup_w|\mu_m(w)|\cdot\mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}C_m\left\|W_i-W_{\ell_W(i)}\right\|\right\}\\
&= o(1),
\end{aligned}
$$

by Lemma A.1. This finishes the proof of (A.4).
Next, we will show that

$$
\mathbb{E}\left\{\left[\frac{1}{N}\sum_{i=1}^{N}V_i^k\cdot V_{\ell_W(i)}^m-\mathbb{E}\left[\mathbb{E}\Big(V_i^k\Big|\,W_i\Big)\cdot\mathbb{E}\Big(V_i^m|\,W_i\Big)\right]\right]^2\right\}=o(1), \tag{A.5}
$$

which, together with (A.4), proves the claim in the Lemma. First we expand the square:

$$
\begin{aligned}
&\mathbb{E}\left\{\left[\frac{1}{N}\sum_{i=1}^{N}V_i^k\cdot V_{\ell_W(i)}^m-\mathbb{E}\left[\mathbb{E}\Big(V_i^k\Big|\,W_i\Big)\cdot\mathbb{E}\Big(V_i^m|\,W_i\Big)\right]\right]^2\right\}\\
&\qquad=\mathbb{E}\left\{\left[\frac{1}{N}\sum_{i=1}^{N}V_i^k\cdot V_{\ell_W(i)}^m\right]^2\right\}+\left\{\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}^2\\
&\qquad\qquad-2\mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}V_i^k\cdot V_{\ell_W(i)}^m\cdot\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}
\end{aligned}
$$

By (A.4), this is equal to

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}V_i^k\cdot V_{\ell_W(i)}^m\right)^2\right]-\left\{\mathbb{E}\left[\mathbb{E}\Big(V_i^k|W_i\Big)\cdot\mathbb{E}\Big(V_i^m|W_i\Big)\right]\right\}^2+o(1)\\
&\qquad=\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[V_i^{2k}\cdot V_{\ell_W(i)}^{2m}\right]+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^kV_{\ell_W(i)}^mV_j^kV_{\ell_W(j)}^m\right]\\
&\qquad\qquad-\left\{\mathbb{E}\left[\mathbb{E}\Big(V_i^k|W_i\Big)\cdot\mathbb{E}\Big(V_i^m|W_i\Big)\right]\right\}^2+o(1).
\end{aligned}
$$

Because the moments of V_i up to at least the $2m$-th and $2k$-th moments exist, it follows that the first term is $o_p(1)$, and the entire expression is

$$
\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^kV_{\ell_W(i)}^mV_j^kV_{\ell_W(j)}^m\right]-\left\{\mathbb{E}\left[\mathbb{E}\Big(V_i^k|W_i\Big)\cdot\mathbb{E}\Big(V_i^m|W_i\Big)\right]\right\}^2+o(1).
$$

Because $\mathrm{pr}\left\{\ell_W(i)=\ell_W(j)\right\}\longrightarrow 0$, $i\neq j$, when $N\longrightarrow\infty$, this is equal to

$$
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}V_i^kV_{\ell_W(i)}^m\right]\cdot\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i}V_j^kV_{\ell_W(j)}^m\right]-\left\{\mathbb{E}\left[\mathbb{E}\Big(V_i^k|W_i\Big)\cdot\mathbb{E}\Big(V_i^m|W_i\Big)\right]\right\}^2+o(1)=o(1),
$$

by (A.4). This finishes the proof of (A.5), and thus the claim in the lemma. □

Lemma A.3 (Average Conditional Variances) *Let (V_i, W_i), $i = 1, \ldots, N$, be a random sample from the distribution of (V, W) where (V, W) are a pair of random vectors, with dimension K_V and K_W respectively, with compact support for W_i. Suppose that $\mu_k(w) = \mathbb{E}[V_i^k | W_i = w]$ is Lipschitz in w with constant C_k for $k \leq 2$, and that the fourth moment of V_i is finite. Define*

$$\widehat{\mathbb{V}}_{\text{cond}} = \frac{1}{2N} \sum_{i=1}^{N} \left(V_i - V_{\ell_W(i)}\right) \left(V_i - V_{\ell_W(i)}\right)'.$$

Then:

$$\widehat{\mathbb{V}}_{\text{cond}} \xrightarrow{p} \mathbb{E}\left[\mathbb{V}(V_i | W_i)\right]. \tag{A.6}$$

Proof of Lemma A.3: To prove $\widehat{\mathbb{V}}_{\text{cond}} \xrightarrow{p} \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]$, we show

$$\mathbb{E}\left\{\widehat{\mathbb{V}}_{\text{cond}} - \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]\right\}^2 = o(1).$$

Without loss of generality we focus on the case with $K_V = 1$:

$$\widehat{\mathbb{V}}_{\text{cond}} = \frac{1}{2N} \sum_{i=1}^{N} \left(V_i - V_{\ell_W(i)}\right)^2 = \frac{1}{2N} \sum_{i=1}^{N} V_i^2 + \frac{1}{2N} \sum_{i=1}^{N} V_{\ell_W(i)}^2 - \frac{1}{N} \sum_{i=1}^{N} V_i V_{\ell_W(i)},$$

and

$$\mathbb{E}\left[\mathbb{V}(V_i|W_i)\right] = \mathbb{E}\left\{\mathbb{E}\left(V_i^2 \middle| W_i\right) - \left[\mathbb{E}\left(V_i \middle| W_i\right)\right]^2\right\} = \mathbb{E}\left[V_i^2\right] - \mathbb{E}\left[\mathbb{E}\left(V_i \middle| W_i\right)^2\right].$$

Because $\sum_{i=1}^{N} V_i^2 / N \xrightarrow{p} \mathbb{E}[V_i^2]$ by a law of large numbers, it is sufficient to show

$$\frac{1}{N} \sum_{i=1}^{N} V_{\ell_W(i)}^2 \xrightarrow{p} \mathbb{E}\left[V_i^2\right], \qquad \text{and} \quad \frac{1}{N} \sum_{i=1}^{N} V_i \cdot V_{\ell_W(i)} \xrightarrow{p} \mathbb{E}\left[\mathbb{E}\left(V_i \middle| W_i\right)^2\right]. \tag{A.7}$$

The first part of (A.7) follows from applying Lemma A.2 with $k = 0$ and $m = 2$, and the second part follows from applying Lemma A.2 with $k = m = 1$. $\square$

Proof of Theorem 2: Since $\hat{\theta} \xrightarrow{p} \theta_{\text{pop}}$ and $\psi(Y_i, X_i, \theta)$ is differentiable in θ, $\hat{\Gamma} \xrightarrow{p} \Gamma$ by the law of large numbers. Then it is sufficient to show $\hat{\Delta}_{\text{cond}} \xrightarrow{p} \Delta_{\text{cond}}$. Define

$$\tilde{\Delta}_{\text{cond}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\psi(Y_i, X_i, \theta_{\text{cond}}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \theta_{\text{cond}})\right) \left(\psi(Y_i, X_i, \theta_{\text{cond}}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \theta_{\text{cond}})\right)'.$$

Let $V_i = \psi(Y_i, X_i, \theta_{\text{cond}})$, and $W_i = X_i$. By Lemma A.3, $\tilde{\Delta}_{\text{cond}} \xrightarrow{p} \mathbb{V}\left(\psi(Y_i, X_i, \theta_{\text{pop}})\right)$. Because $\hat{\theta} \xrightarrow{p} \theta_{\text{cond}}$ and $\psi(Y_i, X_i, \theta)$ is differentiable in θ, it follows that $\hat{\Delta}_{\text{cond}} \xrightarrow{p} \tilde{\Delta}_{\text{cond}}$. Therefore, $\widehat{\mathbb{V}}_{\text{gmm,cond}} = \hat{\Gamma}^{-1} \hat{\Delta}_{\text{cond}} (\hat{\Gamma}')^{-1} \xrightarrow{p} \Gamma^{-1} \Delta_{\text{cond}} (\Gamma')^{-1} = \mathbb{V}_{\text{gmm,cond}}$. $\square$

References

Abadie, A., and G. Imbens (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, Vol. 74(1) 235-267.

Abadie, A., and G. Imbens (2010), "Estimation of the Conditional Variance in Paired Experiments," *Annales dEconomie et de Statistique*, No 91, 175-187.

Angrist, J., and S. Pischke, (2009), *Mostly Harmless Econometrics*, Princeton University Press, Princeton, NJ.

Cameron, C., and P. Trivedi, (2005), *Microeconometrics, Methods and Applications*, Cambridge University Press, Cambridge.

Chamberlain, G., (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, Vol. 18, 5-46.

Chow, G., (1984), "Maximum-likelihood estimation of misspecified models," *Economic Modelling*, Vol. 1(2): 134-138.

Efron, B., (1982), *The Bootstrap and other Resampling Plans*, Philadelphia, Society for Industrial and Applied Mathematics.

Efron, B., and Tibshirani, (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Eicker, F., (1967), "Limit Theorems for Regression with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 59-82, University of California Press, Berkeley.

Goldberger, A., (1991), *A Course in Econometrics*, Harvard University Press.

Hansen, L–P., (1982), "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, vol. 50, 1029–1054.

Huber, P., (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 221-233, University of California Press, Berkeley.

Imbens, G., and M. Kolesár, (2010), "The Behrens-Fisher Problem and Robust Standard Errors in Small Samples With and Without Clustering," Unpublished Manuscript.

Imbens, G., D. Rubin, and B. Sacerdote, (2001), "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91, 778-794.

Karlan, D., and J. List, (2001), "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review* 97(5): 1774-1793.

MacKinnon, J., and H. White, (1985), "Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, Vol. 29, 305-325.

MANSKI, C., (1988), *Analogue Estimation Methods*, Chapman and Hall.

MÜLLER, U., (2011) "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix", Unpublished Manuscript, Princeton University.

NEWEY, W., AND D. MCFADDEN, (1994) "Estimation in Large Samples", in: McFadden and Engle (Eds.), *The Handbook of Econometrics*, Vol. 4.

NEYMAN, J., (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science*, (with discussion), Vol 5, No 4, 465–480, 1990.

PENCAVEL, J., (1986) "Labor Supply of Men: A Survey", in O. Ashenfelter and R. Layard eds., *Handbook of Labor Economics*, North Holland: Elsevier, pp. 3-102.

STOCK, J., AND M. WATSON, (2003), *Introduction to Econometrics*, Addison Wesley.

WHITE, H., (1980a), "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, Vol. 21(1):149-170.

WHITE, H. (1980b), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

WHITE, H., (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*, Vol 50(1): l-25.

WOOLDRIDGE, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Table 1: Summary Statistics for Lottery Data (N=194)

	Average	Standard Deviation
Earnings Post Lottery	11.6	12.3
Earnings Pre Lottery	12.1	11.2
Yearly Prize	32.0	21.1

Table 2: Estimates for Lottery Data

	est.	$\sqrt{\hat{\mathbb{V}}_{\text{blp}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cblp}}}$
intercept	6.497	1.429	1.396
yearly prize	-0.127	0.032	0.028
average lagged earnings	0.755	0.077	0.079

Table 3: Coverage Rate 95% Confidence Interval

Design I: ($\delta = 1.43$)		average s.e.	θ_{blp}	θ_{cblp}
	$\sqrt{\hat{\mathbb{V}}_{\text{blp}}}$	0.0466	0.951	0.951
	$\sqrt{\hat{\mathbb{V}}_{\text{cblp}}}$	0.0450	0.938	0.940
Design II: ($\delta = 14.3$)		average s.e.	θ_{blp}	θ_{cblp}
	$\sqrt{\hat{\mathbb{V}}_{\text{blp}}}$	0.0512	0.953	0.969
	$\sqrt{\hat{\mathbb{V}}_{\text{cblp}}}$	0.0451	0.914	0.941