Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Inference for Misspecified Models With Fixed Regressors

Alberto ABADIE, Guido W. IMBENS, and Fanyin ZHENG

Following the work by Eicker, Huber, and White it is common in empirical work to report standard errors that are robust against general misspecification. In a regression setting, these standard errors are valid for the parameter that minimizes the squared difference between the conditional expectation and a linear approximation, averaged over the population distribution of the covariates. Here, 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 the covariates. We argue that in some cases this may be a more interesting parameter. We derive the asymptotic variance for this parameter, which is generally smaller than the Eicker–Huber–White robust variance, and propose a consistent estimator for this asymptotic variance. Supplementary materials for this article are available online.

KEY WORDS: Bootstrap; Conditional inference; Confidence intervals; Robust standard errors.

1. INTRODUCTION

Following the seminal work by Eicker (1967), Huber (1967), and White (1980a, 1980b, 1982), researchers estimating regression functions routinely report standard errors that are robust to misspecification of the models that are being estimated. Müller (2013) gave the corresponding confidence intervals a Bayesian interpretation. A key feature of the approach developed by Eicker, Huber, and White (EHW from hereon), is that in regression settings it focuses on the best linear predictor that minimizes the distance between a linear function and the true conditional expectation, averaged over the joint distribution of all variables, with extensions to nonlinear settings. We argue that in some regression settings it may be more appropriate to focus on the conditional best linear predictor defined by minimizing this distance averaged over the empirical instead of the population distribution of the covariates. The first contribution of this article is to extend the EHW results to such settings. For a large class of estimators, including maximum likelihood and method of moment estimators, we 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, the robust variance for the conditional estimand is smaller than or equal to the EHW robust variance. Second, 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 and is related to the differencing methods used in Yatchew (1997, 1999). In correctly specified models, the new variance estimator is simply an alternative to the standard EHW robust variance estimator. In misspecified models, it is the only consistent estimator available for the asymptotic variance for the estimand conditional on covariates.

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 the appropriate one. We discuss some examples, first to clarify the distinctions between the two estimands and, second, to make an argument for our view that in some settings the conditional estimand, corresponding to the fixed regressor notion, is of interest. For example, we argue that in cases where the sample is the population there is a strong case for using the estimand conditional on at least some covariates, see also Abadie et al. (2014). Such cases are common in economic analyses, for example, when analyzing data where the units are all states of the United States, or all countries of the world. Most importantly, we argue that there is a 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.

Note that although we focus on estimands defined in terms of the finite sample distribution of the covariates, our inference relies on large sample approximations. To focus on the conceptual contribution of the current article and maintain comparability with the preceding literature, we focus on unconditional inference.

The rest of this article is organized as follows. Section 2 contains a heuristic discussion of the conceptual issues raised by this article 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 article to a dataset previously analyzed by Sachs and Warner (1997) to

Alberto Abadie, John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138, and NBER (E-mail: *alberto_abadie@harvard.edu*). Guido W. Imbens is Professor of Economics, and Graduate School of Business Trust Faculty Fellow, Graduate School of Business, Stanford University, Stanford, CA 94305 and NBER (E-mail: *imbens@stanford.edu*). Fanyin Zheng, Department of Economics, Harvard University 1805 Cambridge Street, Cambridge, MA 02138 (E-mail: *fzheng@fas.harvard.edu*). Financial support for this research was generously provided through NSF grants 0820361 and 0961707. We are grateful for comments by Hal White and participants in the econometrics lunch seminar at Harvard University, and in particular for discussions for Gary Chamberlain that greatly improved the article.

Journal of the American Statistical Association
December 2014, Vol. 109, No. 508, Theory and Methods
DOI: 10.1080/01621459.2014.928218

study the relation between country-level growth rates and government fiscal policies. In Section 6, we present two simulation studies, one in a linear and one in a nonlinear setting. Section 7 concludes. The Appendix contains proofs.

2. THE CONDITIONAL BEST LINEAR PREDICTOR

In this section, we lay out some of the conceptual issues in this article 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 being 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 $\boldsymbol{\varepsilon}$ 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. In this setting, researchers have often assumed homoscedasticity, independence of the errors terms, and Normality of the error terms,

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

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

$$\hat{\theta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right),$$

is Normal:

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

However, assumptions of linearity of the regression function, independence, homoscedasticity, and Normality of the error terms are often unrealistic. Eicker (1967), Huber (1967), and White (1980a, 1980b), considered the properties of the least-squares estimator $\hat{\theta}$ under substantially 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 restrictions. 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. 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*, formally defined as

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

Because

$$\mathbb{E}\left[\left(Y_i - X_i'\theta\right)^2\right] = \mathbb{E}\left[\left(\mu(X_i) - X_i'\theta\right)^2\right] + E\left[\sigma^2(X_i)\right],$$

with the last term free of dependence on θ, it follows that we can characterize θ_{pop} as

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

which in turn shows that θ_{pop} 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.

The results in EHW imply that, under some regularity conditions,

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

where the asymptotic variance is

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

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

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

Using the EHW variance estimator $\hat{\mathbb{V}}_{\text{pop}}$ is currently the standard practice in empirical work in economics, see, for example, Angrist and Pischke (2009). See Imbens and Kolesár (2012) for a discussion of finite sample improvements. Resampling methods such as the jackknife and the bootstrap (Efron 1982; Efron and Tibshirani 1993) can also be used to construct confidence intervals for θ_{pop}.

In this article, 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* $\theta_{\text{cond}}(\mathbf{X})$ as

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

The difference with the best linear predictor defined in (2.2) is that in (2.5) the expectation is taken over the empirical distribution of the covariates, whereas in (2.2) the expectation is taken over the population distribution of the covariates. To be explicit about the dependence of the conditional best linear predictor on the sample values of the covariates we write $\theta_{\text{cond}}(\mathbf{X})$ as a function of the matrix of covariate values $\mathbf{X}$. Denoting the N-vector with ith element equal to $\mu(X_i)$ by $\boldsymbol{\mu}(\mathbf{X})$, we can write $\theta_{\text{cond}}(\mathbf{X})$ as

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

to stress the interpretation of $\theta_{\text{cond}}(\mathbf{X})$ as the best approximation to the true regression function, now with the weights based on the empirical distribution of the covariates. Both θ_{pop} and $\theta_{\text{cond}}(\mathbf{X})$ base the linear approximation to $\mu(x)$ on a minimizing of the squared difference between the true regression function $\mu(x)$ and the linear approximation $x'\theta$. The difference between the two approximations is solely in 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_{pop}, uses the population distribution of the covariates. The second approximation, leading to $\theta_{\text{cond}}(\mathbf{X})$, uses the empirical distribution of the covariates.

We defer to Section 3 the important question whether, and why, in a specific application, $\theta_{\text{cond}}(\mathbf{X})$ rather than θ_{pop} might be the object of interest. In some applications we argue that θ_{pop} is the estimand of interest. However, as discussed in detail in Section 3, we also think that in other applications $\theta_{\text{cond}}(\mathbf{X})$ is of more interest than θ_{pop}. Given that the main focus of the previous literature is on population parameters like θ_{pop}, we view the question of inference for $\theta_{\text{cond}}(\mathbf{X})$ as of general interest.

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

$$\mathbb{V}\left(\hat{\theta}\middle|\mathbf{X}\right) = \mathbb{E}\left[\left(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})\right)\left(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})\right)'\middle|\mathbf{X}\right] \tag{2.6}$$
$$= \frac{1}{N}\left(\mathbf{X}'\mathbf{X}/N\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sigma^2(X_i)X_iX_i'\right)\left(\mathbf{X}'\mathbf{X}/N\right)^{-1}.$$

Directly comparing the normalized variance $N \cdot \mathbb{V}(\hat{\theta}|\mathbf{X})$ to the EHW variance $\mathbb{V}_{\text{pop}}$ is complicated by the fact that $N \cdot \mathbb{V}(\hat{\theta}|\mathbf{X})$ is a conditional variance, rather than an asymptotic variance like $\mathbb{V}_{\text{pop}}$. We therefore look at the unconditional variance of the ols estimator $\hat{\theta}$ as an estimator of $\theta_{\text{cond}}(\mathbf{X})$. Because $\hat{\theta}$ is unbiased for $\theta_{\text{cond}}(\mathbf{X})$, it follows that the marginal variance is the expected value of the conditional variance. Under random sampling the asymptotic variance is

$$\mathbb{V}_{\text{cond}} = \text{plim}\left(N \cdot \mathbb{V}(\hat{\theta}|\mathbf{X})\right) = \left(\mathbb{E}\left[X_iX_i'\right]\right)^{-1}$$
$$\times\left(\mathbb{E}\left[\sigma^2(X_i)X_iX_i'\right]\right)\left(\mathbb{E}\left[X_iX_i'\right]\right)^{-1}, \tag{2.7}$$

and we have, under regularity conditions, a large sample approximation to the distribution of $\sqrt{N} \cdot \left(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})\right)$:

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

The key difference between the robust variance $\mathbb{V}_{\text{pop}}$ proposed by White and the robust variance $\mathbb{V}_{\text{cond}}$ arises from the difference between the conditional variance $\sigma^2(X_i)$ in (2.7) and the expectation of the squared residual $\mathbb{E}[(Y_i - X_i'\theta_{\text{pop}})^2|X_i]$ in (2.3). The latter is in general larger:

$$\mathbb{E}[(Y_i - X_i'\theta_{\text{pop}})^2|X_i] = \sigma^2(X_i) + (\mu(X_i) - X_i'\theta_{\text{pop}})^2,$$

where $\mu(X_i) - X_i'\theta_{\text{pop}}$ captures the difference between the linear approximation and the conditional expectation. For the asymptotic variances of $\hat{\theta}$ we have

$$\mathbb{V}_{\text{pop}} = \mathbb{V}_{\text{cond}} + \mathbb{V}(\theta_{\text{cond}}(\mathbf{X})), \tag{2.8}$$

where

$$\mathbb{V}(\theta_{\text{cond}}(\mathbf{X})) = \text{plim}\; N \cdot \mathbb{E}\left[\left(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}}\right)\right.$$
$$\left.\times\left(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}}\right)'\right] \tag{2.9}$$

The last expectation is over the distribution of $\theta_{\text{cond}}(\mathbf{X})$ as a function of $\mathbf{X}$. Thus in general $\mathbb{V}_{\text{pop}}$ exceeds $\mathbb{V}_{\text{cond}}$, and as a result inference based on $\mathbb{V}_{\text{pop}}$ is conservative for $\theta_{\text{cond}}(\mathbf{X})$. The difference between the two variances is the result of the misspecification in the regression function, that is, the difference between the conditional expectation and the best linear predictor, $\mu(x) - x'\theta_{\text{pop}}$.

The final question we address in this section is how to estimate $\mathbb{V}_{\text{cond}}$. Simple bootstrapping methods do not work, see Tibshirani (1986) and Wu (1986). 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 consistently estimate the conditional variance $\sigma^2(X_i)$ at each distinct value of the covariates and plug that in (2.7), followed by replacing the expectations by averages over the sample. If the covariates are continuous, however, this is not feasible. In the remainder of this discussion, we focus on the continuous covariate case. Dealing with the setting where some of the covariates are discrete is conceptually straightforward, but would require carrying along additional notation and come at the expense of clarity. In the continuous covariate case estimating $\sigma^2(x)$ consistently for all x would require nonparametric estimation involving bandwidth choices. Such an estimator would be more complicated than the EHW robust variance estimator which simply uses squared residuals to estimate the expectation of the squared errors. Here we build on work by Yatchew (1997, 1999) and Abadie and Imbens (2006, 2008, 2010) to develop a general estimator for $\mathbb{V}_{\text{cond}}$ that does not require consistent estimation of $\sigma^2(x)$, much like the EHW variance estimator does not consistently estimate $\mathbb{E}[(Y_i - X_i'\theta_{\text{pop}})^2|X_i = x]$ for all x. Let V_X be the covariance matrix of X, $V_X = \sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})'/N$, where $\overline{X} = \sum_{i=1}^{N} X_i/N$. Next define $\ell_X(i)$ to be the index of the unit closest to i in terms of X:

$$\ell_X(i) = \arg\min_{j\in\{1,\ldots,N\},\, j\neq i}\left\|X_i - X_j\right\|, \tag{2.10}$$

where the norm we use is the Mahalanobis distance, $\|x\| = x'V_X^{-1}x$, although others could be used. Then, our proposed variance estimator is

$$\hat{\mathbb{V}}_{\text{cond}} = \left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i'\right)^{-1} \tag{2.11}$$
$$\cdot\left(\frac{1}{2N}\sum_{i=1}^{N}\left(\hat{\varepsilon}_iX_i - \hat{\varepsilon}_{\ell_X(i)}X_{\ell_X(i)}\right)\left(\hat{\varepsilon}_iX_i - \hat{\varepsilon}_{\ell_X(i)}X_{\ell_X(i)}\right)'\right)$$
$$\cdot\left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i'\right)^{-1}.$$

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

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

Although in this linear regression case with conditioning on all covariates both $\widehat{\mathbb{V}}_{\text{cond}}$ and $\tilde{\mathbb{V}}_{\text{cond}}$ are consistent for $\mathbb{V}_{\text{cond}}$, for nonlinear settings, or with conditioning on a subset of the covariates, only the first estimator $\widehat{\mathbb{V}}_{\text{cond}}$ 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

$$\mathbb{V}_{\text{cond}} = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1} \left(\mathbb{E}\left[\mathbb{V}\left(\varepsilon_i X_i \,|\, X_{1i}\right)\right]\right) \times \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}. \tag{2.12}$$

Our proposed estimator for this conditional variance is

$$\widehat{\mathbb{V}}_{\text{cond}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \cdot \left(\frac{1}{2N}\sum_{i=1}^{N}\left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{X_1}(i)} X_{\ell_{X_1}(i)}\right) \times \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{X_1}(i)} X_{\ell_{X_1}(i)}\right)'\right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}. \tag{2.13}$$

This estimator is consistent for the conditional variance $\mathbb{V}_{\text{cond}}$. In contrast, replacing $\hat{\varepsilon}_{\ell_X(i)}$ by $\hat{\varepsilon}_{\ell_{X_1}(i)}$ in the expression for $\widehat{\mathbb{V}}_{\text{cond}}$ would not lead to a consistent estimator for the variance. Although the asymptotic variance $\mathbb{V}_{\text{cond}}$ is less than or equal to the EHW variance $\mathbb{V}_{\text{pop}}$, this need not hold for the estimators. In finite samples, it may well be the case that $\widehat{\mathbb{V}}_{\text{cond}}$ is larger than $\widehat{\mathbb{V}}_{\text{pop}}$. We study the finite sample behavior of the variance estimator in a simulation study in Section 6.

In the remainder of this article, 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 variance estimators.

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 the conditional estimand is the appropriate object of interest in all cases. Rather, we wish to make the case, through two examples, that it depends on the context what the appropriate object is, and that in some settings the conditional best linear predictor is more appropriate than 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 of the estimator 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 $\theta_{\text{cond}}(\mathbf{X})$ is the approximate average over the repeated samples where X_i is held fixed. Many introductory treatments of regression analysis 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, pp. 10–11). Goldberger (1991) takes a different position, assuming "**X** nonstochastic, which says that the elements of **X** are constants, that is, degenerate random variables. Their values are fixed in repeated samples . . ." (Goldberger, p. 164). VanderVaart (2000) wrote "We assume that the independent variables are a random sample to fit the example in our iid notation, but the analysis could be carried out conditionally as well." (VanderVaart, p. 57), and Gelman and Hill (2007) focus on the fixed regressor perspective, writing "This book follows the usual approach of setting up regression models in the measurement-error framework ($y = a + bx + \epsilon$), with the sampling interpretation implicit in that the errors $\epsilon_1, \ldots, \epsilon_n$, can be considered as a random sample from a distribution" (Gelman and Hill, p. 17). 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 misspecification, 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, focusing on the conditional estimand θ_{cond} generally leads to smaller standard errors than focusing on the population estimand θ_{pop}.

3.1 Example I (Convenience Sample)

In the first 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). Imbens, Rubin, and Sacerdote surveyed individuals who won large prizes in the lottery. Using a standard life-cycle model of labor supply, they focused 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.

3.2 Example II (Experimental Design)

Karlan and List (2007) 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 $\mathbb{V}_{\text{pop}}$.

4. INFERENCE FOR CONDITIONAL ESTIMANDS

In this section, we present the main formal results of the article, 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. We focus on the just-identified case, although the results can be extended to over-identified generalized method of moment (GMM) settings, for example, using empirical likelihood approaches (e.g., Qin and Lawless 1994; Imbens 1997; Imbens, Johnson and Spady 1998; Newey and Smith 2004).

Suppose we have a random sample of size N of a pair of random vectors, (X_i, Y_i), $i = 1, \ldots, N$. Let $\mathbf{X}$ and $\mathbf{Y}$ be the $N \times K_X$ and $N \times K_Y$ matrices with ith rows equal to X_i' and Y_i' respectively. The distinction between $\mathbf{X}$ and $\mathbf{Y}$ is that we may wish to condition on the $\mathbf{X}_i$ in defining the estimand. We are interested in a finite-dimensional parameter θ, defined in general as some function of the joint distribution of (X_i, Y_i). Under some statistical model, it follows that

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

with the dimension of θ equal to that of ψ. The model may have additional implications beyond this moment restriction, but these are not used for estimation. For example, it may be the case that the conditional moment has expectation zero,

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

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 minus 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 $\hat{\theta}$, which satisfies

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

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

The standard approach to GMM and empirical likelihood estimation (Hansen 1984; Qin and Lawless 1993; Newey and McFadden 1994; Wooldridge 2002; Imbens, Johnson and Spady 1997) 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 $i = 1, \ldots, N$ are independent and identically distributed, then under regularity conditions,

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

with

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

Now we focus on the conditional estimand, where we condition on $\mathbf{X}$. Define $\theta_{\text{cond}}(\mathbf{X})$ as the solution to

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

If the original model implied that the conditional expectation of $\psi(Y_i, X_i, \theta)$ given X_i is equal to zero, then $\theta(\mathbf{X}) = \theta_{\text{pop}}$ for all $\mathbf{X}$, 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 restriction, 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. (Y_i, X_i), for $i = 1, \ldots, N$, are independent and identically distributed.

Assumption 2. (i) For some compact $\Theta \subset \mathbb{R}^K$, there is a unique value, $\theta_{\text{pop}} \in \Theta$, such that $\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})] = 0$; (ii) $\psi(Y, X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; (iii) $\mathbb{E}[\sup_{\theta\in\Theta} \|\psi(Y_i, X_i, \theta)\|] < \infty$.

Theorem 1. If Assumptions 1 and 2 hold, then:

$$\hat{\theta} - \theta_{\text{pop}} \xrightarrow{p} 0,$$

and

$$\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}) \xrightarrow{p} 0.$$

All proofs are given in the appendix.

Assumption 3. (i) θ_{pop} is an interior point of Θ; (ii) $\psi(y, x, \theta)$ is continuously differentiable with respect to θ in an open neighborhood $\mathcal{B}$ of θ_{pop}; (iii) $\mathbb{E}[\|\psi(Y_i, X_i, \theta_{\text{pop}})\|^2] < \infty$; (iv) $\mathbb{E}[\sup_{\theta\in\mathcal{B}} \|\partial\psi(Y_i, X_i, \theta)/\partial\theta'\|] < \infty$; (v) $\Gamma = \mathbb{E}[\partial\psi(Y_i, X_i, \theta_{\text{pop}})/\partial\theta']$ is nonsingular.

Theorem 2. Under Assumptions 1–3,

$$\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{d} N(0, \Gamma^{-1}\Delta_{\text{pop}}(\Gamma^{-1})'),$$

and

$$\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})) \xrightarrow{d} N(0, \Gamma^{-1}\Delta_{\text{cond}}(\Gamma^{-1})'),$$

where $\Delta_{\text{cond}} = \mathbb{E}[\mathbb{V}(\psi(Y_i, X_i, \theta_{\text{pop}})|X_i)]$.

Corollary 1. Under the conditions of Theorem 2, if $\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})|X_i = x] = 0$ for almost all x in the support of X_i, then $\theta_{\text{cond}}(\mathbf{X}) = \theta_{\text{pop}}$ for all $\mathbf{X}$ and $(\hat{\theta} - \theta_{\text{pop}})$ and $\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))$ have the same asymptotic distribution.

Assumption 3(ii) requires differentiability of $\psi(y, x, \theta)$. This assumption can, however, be replaced by asymptotic equicontinuity conditions as in Huber (1967), Pakes and Pollard (1989), Andrews (1994), or Newey and McFadden (1994). In a supplementary Web Appendix we show that the results of Theorem 2 and Corollary 1 hold under an asymptotic equicontinuity condition, with the only change that for the nondifferentiable case we have $\Gamma = \partial\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})]/\partial\theta'$. Example VI below discusses the case of L_1 (quantile) regression. Notice that the consistency result in Theorem 1 does not require everywhere differentiability of $\psi(y, x, \theta)$.

We now discuss two additional examples that illustrate the differences between the large sample variances of $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}})$ and $\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))$. The first example is related to the discussion in Chow (1984).

4.1 Example III (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

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

Huber (1967) and White (1982) analyzed the properties of the maximum likelihood estimator under general misspecification of the conditional density. Let

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

They showed that under general misspecification,

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

with

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

$$\Delta_{\text{pop}} = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}}) \cdot \frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}})'\right].$$

The conditional version of the estimand under general misspecification is

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

where the expectation is taken only over the conditional distribution of Y_i given X_i. Theorem 2 implies that

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

where

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

If the model is correctly specified, then $\Delta_{\text{pop}} = \Delta_{\text{cond}}$. If the model is misspecified, then

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

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

for x in a set of positive probability. For such x,

$$\begin{aligned}&\mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}}) \cdot \frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}})'\middle| X_i = x\right] \\ &\geq \mathbb{V}\left(\frac{\partial}{\partial\theta} \ln f(Y_i|X_i, \theta_{\text{pop}})\middle| X_i = x\right),\end{aligned}$$

implying that in general $\Delta_{\text{pop}} - \Delta_{\text{cond}}$ is positive semidefinite.

4.2 Example IV (Quantile Regression)

Suppose that the τth conditional quantile of Y_i given X_i is a linear function, so $\mathbb{E}[I_{[Y_i\leq X_i'\theta_{\text{pop}}]}|X_i = x] = \tau$, where I_A is the indicator function for the event A. Therefore, $\mathbb{E}[X_i(I_{[Y_i\leq X_i'\theta_{\text{pop}}]} - \tau)] = 0$. The quantile regression estimator $\hat{\theta}$ (Koenker and Bassett 1978) solves the analogous sample moment restrictions:

$$\left\|\frac{1}{N}\sum_{i=1}^{N} X_i(I_{[Y_i\leq X_i'\hat{\theta}]} - \tau)\right\| = o_p(1/\sqrt{N}) \qquad (4.3)$$

(see Powell 1984). If the quantile regression model is misspecified, so $\mathbb{E}[I_{[Y_i\leq X_i'\theta_{\text{pop}}]}|X_i = x] \neq \tau$ for some x in a set of positive probability, there will generally still be a value θ_{pop} that solves (4.3). Under regularity conditions the quantile regression estimator estimates that parameter, and its distribution is

$$\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{p} \mathcal{N}(0, \Gamma^{-1}\Delta_{\text{pop}}\Gamma^{-1}),$$

where

$$\Gamma = \mathbb{E}[f_{Y|X=X_i}(X_i'\theta_{\text{pop}})X_i X_i']$$

and

$$\Delta_{\text{pop}} = \mathbb{E}[X_i(I_{[Y_i - X_i'\theta_{\text{pop}}\leq 0]} - \tau)^2 X_i']$$

(see, e.g., Angrist, Chernozhukov, and Fernández-Val (2006), or the online supplementary materials). Angrist, Chernozhukov, and Fernández-Val (2006) provided an interpretation of quantile regression under misspecification. In the online supplementary meterials, we show that, in addition:

$$\sqrt{N}(\hat{\theta} - \theta(\mathbf{X})) \xrightarrow{p} \mathcal{N}(0, \Gamma^{-1}\Delta_{\text{cond}}\Gamma^{-1}),$$

where

$$\Delta_{\text{cond}} = \mathbb{E}[X_i \mathbb{V}(I_{[Y_i - X_i'\theta \leq 0]}|X_i)X_i'].$$

Because $\mathbb{E}[(I_{[Y_i - X_i'\theta_{\text{pop}} \leq 0]} - \tau)^2|X_i] \geq V(I_{[Y_i - X_i'\theta \leq 0]}|X_i)$, it follows that $\Delta_{\text{pop}} - \Delta_{\text{cond}}$ is positive semidefinite. Under correct specification, $\mathbb{E}[I_{[Y_i - X_i'\theta \leq 0]}|X_i] = \tau$, so $\mathbb{E}[(I_{[Y_i - X_i'\theta \leq 0]} - \tau)^2|X_i] = \mathbb{V}(I_{[Y_i - X_i'\theta \leq 0]}|X_i) = \tau(1-\tau)$ and $\Delta_{\text{cond}} = \Delta_{\text{pop}}$.

4.3 Variance Estimation

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:

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

where $\ell_X(i)$ is as defined in (2.10). We then combine these estimates to get an estimator for the variance of the conditional estimand:

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

Assumption 4. The support of X_i is compact. The conditional expectation $\mathbb{E}[\psi^k(Y_i, X_i, \theta)|X_i = x]$ is Lipschitz in x with constant C_k for $k \leq 4$, for all θ in an open neighborhood of θ_{pop}, where C_k does not depend on θ.

Theorem 3. (Conditional Variance for Method of Moments Estimators) Suppose Assumptions 1–4 hold. Then,

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

5. AN APPLICATION TO CROSS-COUNTRY GROWTH REGRESSIONS

For an illustration of the methods discussed in this article, we turn to an analysis in Sachs and Warner (1997) of the determinants of country-level growth rates. Sachs and Warner have data for 83 countries on the country's per capita growth rate between 1965 and 1990, and wish to relate this outcome to country-level fiscal policies. These policies include the degree of openness of the country ("open") and the central government budget balance ("cgb"). Sachs and Warner estimated a linear regression of the per capita growth rate on these variables, also including a number of characteristics of the country such as its location relative to the tropics and the sea (landlocked or not), and some measures of the economic conditions at the beginning of this period, including gross domestic product in 1965 ("gdp65").

Table 1. Cross-country growth regression, dependent variable: Per capita GDP growth between 1965 and 1990

	$\hat{\beta}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
constant	1.66	3.08	3.03
gdp65	−1.50	0.18	0.17
open	10.91	2.76	2.56
open65	−1.08	0.35	0.33
dpop	0.69	0.40	0.45
cgb	0.115	0.025	0.023
inst	0.315	0.071	0.068
tropics	−0.83	0.25	0.24
land	−0.58	0.21	0.26
sxp	−3.92	1.22	1.21
life	0.35	0.12	0.12
life2	−0.003	0.001	0.001
$N = 83$			
$R^2 = 0.862$			

Description of variables: Dependent variable: Average annual growth in real GDP per economically active population between 1970 and 1989. gdp65: Log of real GDP per economically active population in 1965. open: Fraction of years during the period 1965–1990 in which the country is rated as an open economy according to the criteria in Sachs and Warner (1995). open65: open*gdp65. dpop: Difference between the growth rate of the economically active population (between ages 15 and 65) and growth of total population. cgb: Current revenues minus current expenditures of the central government, expressed as a fraction of GDP. inst: Institutional quality index. tropics: Approximate proportion of land area subject to a tropical climate. land: Dummy variable that equals one if a country is landlocked. sxp: Share of exports of primary products in GNP in 1970. life: Life expectancy at birth, ca. 1965–1970. life2: life squared.

The estimates are reported in Table 1, with the variables described at the bottom of the table. We calculate the EHW standard errors, as well as our proposed conditional standard errors where the variables we condition on include all characteristics of the countries other than the economic policy variables open, open×gdp65, and cgb which are directly under the control of the government. It would appear reasonable that at least some of these variables should be conditioned on, including whether a country is landlocked and what share of its landmass is in the tropics.

We find that the standard errors for the key variables, the indicator for being open and its interaction with gdp in 1965 go down by about 7%.

6. TWO SIMULATION STUDIES

In this section, we assess the small sample properties of the variance estimators. We focus on two models, first a linear regression and second a logistic regression model.

6.1 A Simulation Study of a Linear Model

We consider estimating a regression function with K regressors. the first regressor, X_{1i}, has a mixture of a normal distribution with mean zero and unit variance, and a log normal distribution with parameters $\mu = 0$ and $\sigma^2 = 0.5$. The mixture probability for the log normal component is p. We use two values for p in the simulations, $p = 0$ and $p = 0.1$ with the latter corresponding to a design with high leverage covariates.

Table 2. Coverage rate 95% confidence interval and median estimated standard error (linear model, 50,000 replications)

Estimand	$\longrightarrow$					θ_{pop}		θ_{cond}		Median	
Variance	$\longrightarrow$					$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Mis-spec	Homo	Samp Size	High Lev	K						
I	No	Yes	50	No	1	0.926	0.921	0.926	0.921	0.367	0.368
II	No	Yes	50	No	5	0.912	0.897	0.912	0.897	0.366	0.354
III	No	Yes	50	Yes	1	0.923	0.916	0.923	0.916	0.345	0.344
IV	No	Yes	50	Yes	5	0.907	0.892	0.907	0.892	0.344	0.331
V	No	Yes	200	No	1	0.945	0.941	0.945	0.941	0.190	0.188
VI	No	Yes	200	No	5	0.940	0.927	0.940	0.927	0.190	0.182
VII	No	Yes	200	Yes	1	0.942	0.936	0.942	0.936	0.177	0.175
VIII	No	Yes	200	Yes	5	0.939	0.925	0.939	0.925	0.177	0.169
IX	No	No	50	No	1	0.914	0.877	0.914	0.877	0.561	0.548
X	No	No	50	No	5	0.893	0.853	0.893	0.853	0.542	0.508
XI	No	No	50	Yes	1	0.921	0.879	0.921	0.879	0.508	0.493
XII	No	No	50	Yes	5	0.901	0.861	0.901	0.861	0.492	0.460
XIII	No	No	200	No	1	0.938	0.915	0.938	0.915	0.318	0.310
XIV	No	No	200	No	5	0.937	0.903	0.937	0.903	0.316	0.291
XV	No	No	200	Yes	1	0.943	0.917	0.943	0.917	0.284	0.276
XVI	No	No	200	Yes	5	0.940	0.904	0.940	0.904	0.282	0.260
XVII	Yes	Yes	50	No	1	0.904	0.811	0.978	0.938	0.503	0.397
XVIII	Yes	Yes	50	No	5	0.885	0.826	0.967	0.941	0.489	0.422
XIX	Yes	Yes	50	Yes	1	0.816	0.673	0.984	0.948	0.535	0.404
XX	Yes	Yes	50	Yes	5	0.789	0.695	0.976	0.954	0.516	0.434
XXI	Yes	Yes	200	No	1	0.938	0.806	0.993	0.948	0.278	0.195
XXII	Yes	Yes	200	No	5	0.934	0.845	0.991	0.964	0.276	0.215
XXIII	Yes	Yes	200	Yes	1	0.796	0.569	0.998	0.965	0.333	0.204
XXIV	Yes	Yes	200	Yes	5	0.791	0.627	0.997	0.980	0.329	0.233
XXV	Yes	No	50	No	1	0.892	0.827	0.937	0.887	0.655	0.567
XXVI	Yes	No	50	No	5	0.870	0.819	0.922	0.884	0.628	0.556
XXVII	Yes	No	50	Yes	1	0.871	0.763	0.950	0.903	0.675	0.561
XXVIII	Yes	No	50	Yes	5	0.841	0.757	0.939	0.904	0.644	0.555
XXIX	Yes	No	200	No	1	0.931	0.865	0.966	0.919	0.376	0.314
XXX	Yes	No	200	No	5	0.928	0.868	0.964	0.924	0.373	0.312
XXXI	Yes	No	200	Yes	1	0.878	0.727	0.982	0.941	0.423	0.318
XXXII	Yes	No	200	Yes	5	0.873	0.739	0.981	0.950	0.418	0.324

The remaining $K-1$ covariates have normal distributions with mean zero and unit variance. All covariates are independent. We use two values for the number of covariates: $K=1$ where only X_{1i} is present in the regression function, and $K=5$ where there are four additional regressors. We use two sample sizes, $N=50$ and $N=200$. The conditional distribution of Y_i given $(X_{1i},\dots,X_{Ki})$ is Normal:

$$Y_i|X_{1i},\dots,X_{Ki}\sim\mathcal{N}\left(\mu_i,\sigma_i^2\right),$$

where

$$\mu_i=X_{1i}+\delta\cdot\left(X_{1i}^2-1\right),\text{ and }\ \ln\sigma_i^2=1-\gamma\cdot X_{1i}.$$

A nonzero value for δ makes the model nonlinear and implies that the linear regression model is misspecified. We use two values for δ. In the first design, we fix $\delta=0$ (correct specification), and in the second design we use a larger value, $\delta=1$ (misspecification). A nonzero value for γ implies heteroscedasticity. We use two values for γ, $\gamma=0$ (homoscedasticity) and $\gamma=0.5$ (heteroscedasticity). With two values for each of five parameters of the design, $p\in\{0,0.1\}$, $K\in\{1,5\}$, $N\in\{50,200\}$, $\delta\in\{0,0.1\}$, and $\gamma\in\{0.0.5\}$, we consider a total of 32 designs.

For each of the 32 designs, we focus on estimating a linear regression function

$$Y_i=\theta_0+\sum_{k=1}^{K}\theta_k\cdot X_{ki}+\varepsilon_i.$$

Table 2 presents the results, based on 50,000 replications for each design. We focus on the coefficient on X_{1i}, denoted by θ (dropping the subscript 1 for ease of notation). For all designs, we report four coverage rates. First, the coverage frequency of the conventional (EHW standard error based) 95% confidence interval for θ_{pop}. This coverage frequency is calculated as the frequency with which $(\hat{\theta}-\theta_{\text{pop}})/\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$ is less than 1.96 in absolute value. Note that both θ_{pop} and θ_{cond} need to be numerically evaluated for these data-generating processes. The nominal coverage rate of the confidence intervals is 0.95. Next, the frequency

Table 3. Coverage rate 95% confidence interval and median estimated standard error (logistic model, 50,000 replications)

Estimand	⟶					θ_{pop}		θ_{cond}		Median	
Variance	⟶					$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Mis-spec	Homo	Samp Size	High Lev	K						
I	No	Yes	50	No	1	0.946	0.941	0.946	0.941	0.378	0.387
II	No	Yes	50	No	5	0.934	0.929	0.934	0.929	0.419	0.428
III	No	Yes	50	Yes	1	0.946	0.940	0.946	0.940	0.370	0.379
IV	No	Yes	50	Yes	5	0.933	0.927	0.933	0.927	0.412	0.420
V	No	Yes	200	No	1	0.941	0.934	0.941	0.934	0.286	0.291
VI	No	Yes	200	No	5	0.947	0.945	0.947	0.945	0.191	0.191
VII	No	Yes	200	Yes	1	0.948	0.945	0.948	0.945	0.183	0.183
VIII	No	Yes	200	Yes	5	0.947	0.945	0.947	0.945	0.188	0.187
IX	No	No	50	No	1	0.947	0.941	0.948	0.941	0.350	0.358
X	No	No	50	No	5	0.935	0.930	0.936	0.931	0.383	0.389
XI	No	No	50	Yes	1	0.943	0.935	0.945	0.937	0.340	0.346
XII	No	No	50	Yes	5	0.930	0.924	0.933	0.927	0.371	0.377
XIII	No	No	200	No	1	0.950	0.946	0.950	0.946	0.173	0.173
XIV	No	No	200	No	5	0.945	0.941	0.945	0.941	0.177	0.176
XV	No	No	200	Yes	1	0.946	0.942	0.947	0.942	0.170	0.169
XVI	No	No	200	Yes	5	0.944	0.940	0.945	0.940	0.173	0.172
XVII	Yes	Yes	50	No	1	0.965	0.649	0.999	0.926	0.114	0.055
XVIII	Yes	Yes	50	No	5	0.957	0.802	0.999	0.978	0.136	0.089
XIX	Yes	Yes	50	Yes	1	0.962	0.658	0.999	0.926	0.120	0.059
XX	Yes	Yes	50	Yes	5	0.930	0.872	0.978	0.948	0.399	0.333
XXI	Yes	Yes	200	No	1	0.955	0.636	1.000	0.942	0.056	0.026
XXII	Yes	Yes	200	No	5	0.954	0.715	1.000	0.970	0.058	0.032
XXIII	Yes	Yes	200	Yes	1	0.955	0.648	1.000	0.941	0.058	0.028
XXIV	Yes	Yes	200	Yes	5	0.953	0.724	1.000	0.972	0.061	0.034
XXV	Yes	No	50	No	1	0.963	0.635	1.000	0.924	0.120	0.056
XXVI	Yes	No	50	No	5	0.956	0.784	0.999	0.980	0.138	0.089
XXVII	Yes	No	50	Yes	1	0.960	0.764	1.000	0.924	0.127	0.061
XXVIII	Yes	No	50	Yes	5	0.956	0.795	0.999	0.978	0.145	0.096
XXIX	Yes	No	200	No	1	0.956	0.627	1.000	0.942	0.058	0.027
XXX	Yes	No	200	No	5	0.954	0.702	1.000	0.971	0.061	0.033
XXXI	Yes	No	200	Yes	1	0.942	0.887	0.978	0.943	0.356	0.301
XXXII	Yes	No	200	Yes	5	0.953	0.710	1.000	0.971	0.064	0.035

with which the same confidence interval covers θ_{cond}, that is, the frequency with which $(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))/\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$ is less than 1.96 in absolute value. This should in large samples be at least 0.95, and more than 0.95 in misspecified models according to our formal results. We also 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 $\theta_{\text{cond}}(\mathbf{X})$ should be 0.95.

In the first design (Design I) with a single covariate, 50 observations, a linear conditional expectation and a normal regressor and homoscedasticity, both variance estimators lead to coverage rates around 92%–93%, with the EHW variance doing slightly better. With five covariates (Design II), the difference between the two variance estimators (in favor of the EHW variance estimator) becomes more pronounced. Having a skewed distribution for the covariate with some high leverage values does not change the coverage rates very much in Design III. With 200 observations (Design V), the coverage rates become closer to the nominal coverage rates for both variance estimators. Given heteroscedasticity (Design IX), the EHW variance estimator does substantially better with a coverage rate of 91%, whereas the conditional variance estimator leads to confidence intervals with a coverage rate of 88% Allowing for misspecification of the regression function (Design XVII) changes the coverage rates substantially. The coverage rate, based on the EHW estimator, for θ_{pop}, is 90%. The coverage rate based on the conditional variance estimator, for θ_{cond}, is much closer to the nominal level, at 0.94.

Over the 32 designs, the worst performance of the EHW variance estimator is in Design XX, with misspecification and high leverage covariates, 50 observations, and 5 covariates, where the coverage rate is 79% instead of 95%. The worst performance of the conditional variance estimator is in Design XII, with a linear model, heteroscedasticity, five covariates, with high leverage, and 50 observations, with an actual coverage rate of 88%. It appears that the conditional variance estimator is more sensitive to heteroscedasticity, but less sensitive to the distribution of the covariates. Overall the worst case for the conditional variance estimator is substantially better than for the EHW variance estimator.

6.2 A Simulation Study of a Logistic Regression Model

Next, we do a similar simulation study in a nonlinear setting. Here, the outcome is a binary indicator. We estimate a logistic regression model specified as

$$\text{pr}(Y_i = 1 | X_{1i}, \ldots, X_{Ki}) = \frac{1}{1 + \exp(\theta_0 + \sum_{k=1}^{K} \theta_k \cdot X_{ki})}.$$

The data are generated through a model where a latent index Y_i^* satisfies

$$Y_i^* = \theta_0 + \sum_{k=1}^{K} \theta_k \cdot X_{ki} + \varepsilon_i,$$

and the observed outcome is the indicator that Y_i^* is nonnegative:

$$Y_i = I_{[Y_i^* \geq 0]}.$$

In the base case, there are 50 observations, and ε_i has a logistic distribution so that the logistic regression model is correctly specified. In this case there is a single covariate ($K = 1$), $\theta_1 = 1$, $\theta_0 = 0$, and the covariate has a standard Normal distribution with unit variance.

We can consider combinations of five modifications, similar to those in the linear model. First, we allow for the presence of four additional covariates ($K = 5$), with the additional covariates all having independent normal distributions with zero coefficients. Second, we change the distribution of the first covariate to include high leverage points by making it a mixture of a standard Normal distribution and a log normal distribution with parameters 0 and 0.5, and the probability of the log normal component equal to 0.1. Third, we change the sample size to 200. Fourth, we multiply the ε_i for all units by $\exp(1 - 0.5 \cdot X_{1i})$. In the linear case, this corresponds to introducing heteroscedasticity, but here this also implies misspecification of the logistic regression model. Finally, we directly misspecify the regression function by changing the specification of Y_i^* to

$$Y_i^* = X_{1i} + \left(X_{1i}^2 - 1\right) + \varepsilon_i.$$

Table 3 presents the results for the 32 designs generated as combinations of these changes to the base design, based on 50,000 replications. There are some qualitative differences with the simulations for the linear case. There are generally bigger differences between the two variance estimators, $\hat{\mathbb{V}}_{\text{cond}}$ and $\hat{\mathbb{V}}_{\text{pop}}$. The coverage rates for confidence intervals, for θ_{pop} based on $\hat{\mathbb{V}}_{\text{pop}}$, and for $\theta_{\text{cond}}(\mathbf{X})$ based on $\hat{\mathbb{V}}_{\text{cond}}$, are closer to nominal levels. In contrast, inference for $\theta_{\text{cond}}(\mathbf{X})$ based on $\hat{\mathbb{V}}_{\text{pop}}$ leads to confidence intervals with substantially higher coverage, and inference for θ_{cond} based on $\hat{\mathbb{V}}_{\text{cond}}(\mathbf{X})$ leads to substantial undercoverage.

In general, inference for $\theta_{\text{cond}}(\mathbf{X})$ is less affected by the changes in the design than inference for θ_{pop}. For example, the worst design for θ_{pop} is still Design XX, with both misspecification and high leverage covariates, where the coverage rate is 0.930. For the conditional estimand, the worst designs are those with misspecification, with coverage rates around 0.924, still close to the nominal 0.95 level.

7. CONCLUSION

In this article, we discuss inference for conditional estimands in misspecified models. Following the work by Eicker (1967), Huber (1967), and White (1980a, 1980b, 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 A: PROOFS OF THEOREMS

A.1 Proof of Theorem 1

Given Assumptions 1 and 2, Theorem 2.6 in Newey and McFadden (1994) implies the first result. To prove the second result, let $\rho(x, \theta) = \mathbb{E}[\psi(Y_i, X_i, \theta) | X_i = x]$. Notice that $\mathbb{E}[\rho(X_i, \theta_{\text{pop}})] = 0$. Therefore, $\theta_{\text{cond}}(\mathbf{X})$ can be thought of as an extremum estimator that minimizes

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

We will prove $\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}} \xrightarrow{p} 0$ by showing that Assumption 2 also holds if we replace $\psi(Y_i, X_i, \theta)$ with $\rho(X_i, \theta)$. Because $\mathbb{E}[\rho(X_i, \theta)] = \mathbb{E}[\psi(Y_i, X_i, \theta)]$ it follows that part (i) in Assumption 2 holds also with $\rho(X_i, \theta)$ replacing $\psi(Y_i, X_i, \theta)$. Part (ii) of Assumption 2 follows from dominated convergence because, by Assumption 2(iii), $\mathbb{E}[\sup_{\theta \in \Theta} \|\psi(Y_i, X_i, \theta)\| | X_i] < \infty$ with probability one. To prove that part (iii) holds also after replacing $\psi(Y_i, X_i, \theta)$ with $\rho(X_i, \theta)$, notice that,

$$\|\rho(X_i, \theta)\| = \|\mathbb{E}[\psi(Y_i, X_i, \theta) | X_i]\| \leq \mathbb{E}\Big[\|\psi(Y_i, X_i, \theta)\| \,\Big|\, X_i\Big],$$

because the norm is a convex function by the triangle inequality. Therefore,

$$\begin{aligned}\sup_{\theta \in \Theta} \|\rho(X_i, \theta)\| &\leq \sup_{\theta \in \Theta} \mathbb{E}\Big[\|\psi(Y_i, X_i, \theta)\| \,\Big|\, X_i\Big] \\ &\leq \mathbb{E}\Big[\sup_{\theta \in \Theta} \|\psi(Y_i, X_i, \theta)\| \,\Big|\, X_i\Big].\end{aligned}$$

Taking expectations on both sides of the previous equation and using Assumption 2(iii), we obtain $\mathbb{E}[\sup_{\theta \in \Theta} \|\rho(X_i, \theta)\|] < \infty$. Now, Theorem 2.6 in Newey and McFadden (1994) implies $\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}} \xrightarrow{p} 0$ and, therefore, the second result of the theorem. □

A.2 Proof of Theorem 2

The first result follows from Theorem 3.4 in Newey and McFadden (1994).

To prove the second result, we will first establish the joint asymptotic distribution of $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}})$ and $\sqrt{N}(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}})$, and then we use this result to derive the asymptotic distribution of

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})) = \sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \\ - \sqrt{N}(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}}).\end{aligned} \quad \text{(A.1)}$$

By Assumptions 3(ii) and (iv) and Lemma 3.6 in Newey and McFadden (1994) we obtain that, for x in a set of probability one, $\rho(x, \theta)$ is continuously differentiable with respect to θ in an open neighborhood $\mathcal{N}$ of θ_{pop}, with

$$\frac{\partial\rho(x,\theta)}{\partial\theta'} = \mathbb{E}\left[\frac{\partial\psi(Y_i,X_i,\theta)}{\partial\theta'}\,\middle|\, X_i = x\right].$$

Notice that

$$\begin{aligned}&\psi(Y_i,X_i,\theta_{\text{pop}})'\psi(Y_i,X_i,\theta_{\text{pop}}) = \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'\\ &\quad\times|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i] + (\psi(Y_i,X_i,\theta_{\text{pop}}) - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})\\ &\quad\times|X_i])'(\psi(Y_i,X_i,\theta_{\text{pop}}) - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i])\\ &\quad+2E[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i](\psi(Y_i,X_i,\theta_{\text{pop}}) - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]).\end{aligned}$$

Taking expectation eliminates the cross-product term, which implies

$$\mathbb{E}\left[\|\rho(X_i,\theta_{\text{pop}})\|^2\right] \le \mathbb{E}\left[\|\psi(Y_i,X_i,\theta_{\text{pop}})\|^2\right] < \infty.$$

Using convexity of the norm, we obtain

$$\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial\rho(x,\theta)}{\partial\theta'}\right\| \le \mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial\psi(Y_i,X_i,\theta)}{\partial\theta'}\right\|\,\middle|\, X_i = x\right].$$

Taking averages on both sides of the last equation and using Assumption 3(iv) we obtain:

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial\rho(x,\theta)}{\partial\theta'}\right\|\right] < \infty.$$

Notice also that

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

which is nonsingular by Assumption 3(v).

As a result, Theorem 3.4 in Newey and McFadden (1994) holds for the estimator that minimizes

$$\left(\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\psi(Y_i,X_i,\theta_1)\\ \rho(X_i,\theta_2)\end{pmatrix}\right)'\left(\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\psi(Y_i,X_i,\theta_1)\\ \rho(X_i,\theta_2)\end{pmatrix}\right)$$

with respect to θ_1 and θ_2. Applying Theorem 3.4 of Newey and McFadden (1994), we obtain

$$\sqrt{N}\begin{pmatrix}\hat\theta - \theta_{\text{pop}}\\ \theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}}\end{pmatrix} \xrightarrow{d} N\left(0,\,\Gamma_{\text{joint}}^{-1}\mathbb{V}_{\text{joint}}(\Gamma_{\text{joint}}^{-1})'\right),$$

where $\mathbb{V}_{\text{joint}}$ is equal to

$$\begin{pmatrix}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})\psi(Y_i,X_i,\theta_{\text{pop}})']} & \mathbb{E}_{[}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]]}\\ \mathbb{E}_{[}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]]} & \mathbb{E}_{[}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]}\mathbb{E}_{[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]]}\end{pmatrix},$$

and

$$\Gamma_{\text{joint}} = \begin{pmatrix}\Gamma & 0\\ 0 & \Gamma\end{pmatrix}.$$

Now, because Equation (A.1), we obtain,

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

where $\mathbb{V}_{\text{gmm,cond}} = \Gamma^{-1}\Delta_{\text{cond}}(\Gamma^{-1})'$, and

$$\begin{aligned}\Delta_{\text{cond}} &= \mathbb{E}\Big[\psi(Y_i,X_i,\theta_{\text{pop}})\psi(Y_i,X_i,\theta_{\text{pop}})'\Big]\\ &\quad-\mathbb{E}\Big[\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]\Big]\\ &= \mathbb{E}\left[\mathbb{V}(\psi(Y_i,X_i,\theta_{\text{pop}})|X_i)\right].\end{aligned}$$

□

A.3 Proof of Corollary 1

The result follows directly from $\mathbb{V}(\psi(Y_i,X_i,\theta_{\text{pop}})|X_i) = \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i] - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]$. □

We next state a lemma from Abadie and Imbens (2010) that will be useful in what follows.

Lemma A.1. (Lemma 1, Abadie and Imbens 2010, p. 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 variables, with V_i scalar, and with compact support for W_i. For some positive integer n, and for $j = 1,2,\ldots,n$, let $\mu_p(w) = \mathbb{E}[V_i^p|W_i = w]$ be Lipschitz in w with constant C_p. Then for all nonnegative k, m such that $\max(k,m) \le n/2$,

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

A.4 Proof of Lemma A.2

First we show

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

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

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

Therefore,

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

$$= \left| \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} \mu_k(W_i)\left[\mu_m\left(W_{\ell_W(i)}\right) - \mu_m(W_i)\right]\right\}\right|$$

$$\le \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N} |\mu_k(W_i)|\cdot\left|\mu_m\left(W_{\ell_W(i)}\right)-\mu_m(W_i)\right|\right\}$$

$$\le \sup_w |\mu_k(w)|\cdot \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N} C_m \left\|W_i - W_{\ell_W(i)}\right\|\right\}$$

$$= o(1),$$

by Lemma A.1 and dominated convergence. This finishes the proof of (A.2).

Next, we will show that

$$\mathbb{E}\left\{\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k\cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\left(V_i^k\middle|\, W_i\right)\cdot\mathbb{E}\left(V_i^m\middle|\, W_i\right)\right]\right]^2\right\} = o(1), \quad \text{(A.3)}$$

which, together with (A.2), proves the claim in the Lemma. First, we expand the square:

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

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

By (A.2), this is equal to

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

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[V_i^{2k}\cdot V_{\ell_W(i)}^{2m}\right] + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\right] \quad \text{(A.4)}$$
$$-\left\{\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}^2 + o(1).$$

Consider the first term in (A.4). Using the independence of V_i and $V_{\ell_W(i)}$ conditional on $\mathbf{W}$ we have

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[V_i^{2k}\cdot V_{\ell_W(i)}^{2m}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left[V_i^{2k}|W_i\right]\cdot\mathbb{E}\left[V_{\ell_W(i)}^{2m}|W_{\ell_W(i)}\right]\right]$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\mu_{2k}(W_i)\cdot\mu_{2m}(W_{\ell_W(i)})\right] \le \frac{C}{N'}$$

because the terms are bounded by the Lipschitz condition on $\mu_p(x)$ for all p at least equal to $2k$ and $2m$. Therefore, the first term in (A.4) is $o(1)$, and the entire expression is

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\right] - \left\{\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}^2 + o(1). \quad \text{(A.5)}$$

We write the expectation of the first term conditional on $\mathbf{W}$ as

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\,\mathbf{W}\right]\right]$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\,\mathbf{W}\right]\right]$$
$$+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)=j \text{ or } \ell_W(j)=i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\,\mathbf{W}\right]\right].$$

The number of terms in the second sum is limited by the "kissing number" the number of units a given unit can be the closest match for (Miller et al.1997; see also Abadie and Imbens 2010), which depends on the dimension of W_i. Let the kissing number be denoted by $\overline{L}$. Then, for given i there is only one j such that $\ell_W(i) = j$, and at most $\overline{L}$ j such that $\ell_W(j) = i$. With each term in the second sum bounded by $\mathbb{E}[V_i^{m+k}|W_i]\cdot\mathbb{E}[V_i^{m+k}|W_i]$, which is bounded, the second sum is bounded by

$$\mathbb{E}\left[\frac{\overline{L}}{N}\cdot\mathbb{E}[V_i^{m+k}|W_i]^2\right] = o(1).[4pt]$$

Hence

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\,\mathbf{W}\right]\right]$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\,\mathbf{W}\right]\right] + o(1)$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_{\ell_W(i)}^m|W_{\ell_W(i)}\right)\right.$$
$$\left.\times\ \mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_{\ell_W(j)}^m|W_{\ell_W(j)}\right)\right\} + o(1). \quad \text{(A.6)}$$

Because of the Lipschitz condition on $\mu_p(w) = \mathbb{E}[V_i^p|W_i = w]$ it follows that

$$\left|\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_{\ell_W(i)}^m|W_{\ell_W(i)}\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_{\ell_W(j)}^m|W_{\ell_W(j)}\right)\right.$$
$$\left.-\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right|$$
$$\le C\cdot\max_i\|W_i - W_{\ell_W(i)}\|\cdot\max_j\|W_j - W_{\ell_W(j)}\|,$$

which goes to zero by Lemma A.1. Hence (A.6) is

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\right.$$
$$\left.\times\ \mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right\} + o(1). \quad \text{(A.7)}$$

Next we show that this is equal to

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\right.$$
$$\left.\times\ \mathbb{E}\left(V_j^m|W_j\right)\right\} + o(1). \quad \text{(A.8)}$$

The difference between (A.7) and (A.8) is

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j|j=i \text{ or } \ell_W(i)=j, \text{ or } \ell_W(j)=i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\right. \left.\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right\}. \quad \text{(A.9)}$$

All terms in this sum are bounded by the Lipschitz condition. By the bound on the kissing number and the boundedness of the expectations, it follows that (A.9) is $o(1)$. Next,

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

and thus

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

by (A.2). This finishes the proof of (A.3), 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 compact support for W_i. Suppose that $\mu_p(w) = \mathbb{E}[V_i^p|W_i = w]$ is Lipschitz in w with constant C_p for $p \leq 4$. Define

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

Then:

$$\hat{\mathbb{V}}_{\text{cond}} \xrightarrow{p} \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]. \quad \text{(A.10)}$$

A.5 Proof of Lemma A.3

To prove $\hat{\mathbb{V}}_{\text{cond}} \xrightarrow{p} \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]$, we show

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

Without loss of generality we focus on the case with V scalar:

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

and

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

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

$$\frac{1}{N}\sum_{i=1}^{N}V_{\ell_W(i)}^2 \xrightarrow{p} \mathbb{E}\left[V_i^2\right], \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N}V_i\cdot V_{\ell_W(i)} \xrightarrow{p} \mathbb{E}\left[\mathbb{E}\left(V_i|\,W_i\right)^2\right]. \quad \text{(A.11)}$$

The first part of (A.11) follows from applying Lemma A.2 with $k = 0$ and $m = 2$, and the second part follows from applying Lemma A.2 with $k = m = 1$. □

A.6 Proof of Theorem 3

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 $\hat{\theta} \xrightarrow{p} \theta_{\text{pop}}$ and $\psi(Y_i, X_i, \theta)$ is differentiable in θ, it follows that $\hat{\Delta}_{\text{cond}} - \tilde{\Delta}_{\text{cond}} \xrightarrow{p} 0$. Therefore, $\hat{\mathbb{V}}_{\text{gmm,cond}} = \hat{\Gamma}^{-1}\hat{\Delta}_{\text{cond}}(\hat{\Gamma}')^{-1} \xrightarrow{p} \Gamma^{-1}\Delta_{\text{cond}}(\Gamma')^{-1} = \mathbb{V}_{\text{gmm}}$, cond. □

SUPPLEMENTARY MATERIALS

The supplementary materials contain the proof of Theorem 2 and Corollary 1 under asymptotic equicontinuity condition and an application to quantile regression.

[Received October 2012. Revised May 2014.]

REFERENCES

Abadie, A., and Imbens, G. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267. [1601,1603]

——— (2008), "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76, 1537–1557. [1603]

——— (2010), "Estimation of the Conditional Variance in Paired Experiments," *Annales dEconomie et de Statistique*, 91, 175–187. [1603,1611,1612]

Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2014), *Finite Population Causal Standard Errors*, NBER Working Paper 20325, Cambridge, MA: National Bureau of Economic Research. [1601]

Andrews, D. W. K. (1994), "Empirical Process Methods in Econometrics," in *Handbook of Econometrics* (Vol. IV), eds. R. F. Engle, and D. L. McFadden, Amsterdam: Elsevier Science. [1606]

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006), "Quantile Regression Under Misspecification, With an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563. [1607]

Angrist, J., and Pischke, S. (2009), *Mostly Harmless Econometrics*, Princeton, NJ: Princeton University Press. [1602]

Chow, G. (1984), "Maximum-Likelihood Estimation of Misspecified Models," *Economic Modelling*, 1, 134–138. [1606]

Efron, B. (1982), *The Jacknife, the Bootstrap and Other Resampling Plans*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [1602]

Efron, B., and Tibshirani, (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall. [1602]

Eicker, F. (1967), "Limit Theorems for Regression With Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), Berkeley, CA: University of California Press, pp. 59–82. [1601,1602,1610]

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press. [1604]

Goldberger, A. (1991), *A Course in Econometrics*, Cambridge, MA: Harvard University Press. [1604]

Huber, P. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), Berkeley, CA: University of California Press, pp. 221–233. [1601,1602,1606,1610]

Imbens, G. (1997), “One-Step Estimators for Over-Identified Generalized Method of Moments Models,” *Review of Economic Studies*, 61, 655–680. [1605]

Imbens, G., and Kolesár, M. (2012), *Robust Standard Errors in Small Samples: Some Practical Advice*, NBER Working Paper 18478, Cambridge, MA: National Bureau of Economic Research. [1602]

Imbens, G., Rubin, D., and Sacerdote, B. (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. [1604]

Imbens, G. W., Spady, R. H., and Johnson, P. (1998), “Information Theoretical Approaches to Inference in Moment Condition Models,” *Econometrica*, 66, 333–357. [1605]

Koenker, R., and Bassett, G. (1978), “Regression Quantiles,” *Econometrica*, 46, 33–50. [1606]

Miller, G. L., Teng, S., Thurston, W., and Vavasis, S. A. (1997), “Separators for Sphere-Packings and Nearest Neighbor Graphs,” *Journal of the ACM*, 44, 1–29. [1612]

Müller, U. (2013), “Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix,” *Econometrica*, 81, 1805–1849. [1601]

Newey, W., and McFadden, D. (1994), “Estimation in Large Samples,” in *The Handbook of Econometrics* (Vol. 4), eds. D. McFadden, and R. F. Engle, Amsterdam: Elsevier. [1605,1606,1610,1611]

Newey, W., and Smith, R. (2004), “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72, 219–255. [1605]

Pakes, A., and Pollard, D. (1989) “Simulation and the Asymptotics of Optimization Estimators,” *Econometrica*, 57, 1027–1057. [1606]

Pencavel, J. (1986), “Labor Supply of Men: A Survey,” in *Handbook of Labor Economics*, eds. O. Ashenfelter, and R. Layard, North Holland: Elsevier, pp. 3–102. [1604]

Powell, J. L. (1984) “Least Absolute Deviations Estimation for the Censored Regression Model,” *Journal of Econometrics*, 25, 303–325. [1606]

Qin, J., and Lawless, J. (1994), “Generalized Estimating Equations,” *The Annals of Statistics*, 20, 300–325. [1605]

Sachs, J., and Warner, A. (1997), “Fundamental Sources of Long-Run Growth,” *American Economic Review*, 87, 184–188. [1601,1607]

Tibshirani, R. (1986), Comment on “Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis” by J. Wu, *The Annals of Statistics*, 14, 1335–1339. [1603]

VanderVaart, A. (2000), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [1604]

White, H. (1980a), “Using Least Squares to Approximate Unknown Regression Functions,” *International Economic Review*, 21, 149–170. [1601,1602,1610]

——— (1980b), “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817–838. [1601,1602,1610]

——— (1982), “Maximum Likelihood Estimation of Misspecified Models,” *Econometrica*, 50, 1–25. [1601,1606,1610]

Wooldridge, J. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press. [1604,1605]

Wu, J. (1986), “Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis,” *The Annals of Statistics*, 14, 1261–1295. [1603]

Yatchew, A. (1997), “An Elementary Estimator of the Partial Linear Model,” *Economic Letters*, 57, 135–143. [1601,1603]

——— (1999), “An Elementary Nonparametric Differencing Test of Equality of Regression Functions,” *Economic Letters*, 62, 271–278. [1601,1603]