Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables[1] (Preliminary and Incomplete)

Joseph G. Altonji
Timothy Conley
Todd E. Elder
Christopher R. Taber

May 26, 2011

[1]We have received helpful comments from seminar participants at Northwestern, University of Chicago, University of Michigan, University of Pennsylvania, University of Wisconsin at Madison, U.C. Santa Barbara, Georgetown, and Yale. We also thank Don Andrews, Joel Horowitz, and Xiaoxia Shi for helpful comments. We are grateful for financial support from the National Institute of Child Health and Development grant R01 HD36480-03 (Altonji and Taber) and from the Economic Growth Center, Yale University (Altonji).

Abstract

We develop new estimation methods for estimating causal effects based on the idea that the amount of selection on the observed explanatory variables in a model provides a guide to the amount of selection on the unobservables. We discuss two approaches, one of which involves the use of a factor model as a way to infer properties of unobserved covariates from the observed covariates. We construct an interval estimator that asymptotically covers the true value of the causal effect, and we propose related confidence regions that cover the true value with fixed probability.

1 Introduction

Distinguishing between correlation and causality is the most difficult challenge faced by empirical researchers in the social sciences. Social scientists are rarely in a position to run a well controlled experiment. Consequently, they rely on a priori restrictions about the relationships between the variables that are observed or unobserved. These restrictions are typically in the form of exclusion restrictions or assumptions about the functional form of the model, the distribution of the unobserved variables, or dynamic interactions. Occasionally, the restrictions are derived from a widely accepted theory or are supported by other studies that had access to a richer set of data. However, in most cases, doubt remains about the validity of the identifying assumptions and the inferences that are based on them. This reality has lead a number of researchers to focus on the estimation of bounds under weaker assumptions than those that are conventionally imposed.

In this paper, we develop estimation strategies that are useful in cases in which doubt remains about the exogeneity of instrumental variables or the treatment itself. This is the situation in many applications in economics and the other social sciences, with examples including the effectiveness of private schools, the effects of education on crime, the effects of crime on labor market outcomes, or the effects of obesity on health outcomes. Our approach uses the degree of selection on observed variables as a guide to the degree of selection on the unobservables. Researchers often informally argue for the exogeneity of an explanatory variable or an instrumental variable by examining the relationship between the instrumental variable and a set of observed characteristics, or by assessing whether point estimates are sensitive to the inclusion of additional control variables.[1] We provide a formal theoretical analysis confirming the intuition and providing conditions under which such evidence can be informative. It is important that we view this methodology as not an identification strategy itself in the sense in which one needs an instrument or treatment that is approximately exogenous in order for the bounds to be tight. If there is a lot of "selection on the observables" then the bounds can be very wide, but in the ideal case in which there is very little selection on observables, the bounds will be tight.

[1] See for example, Currie and Duncan (1995), Engen et al (1996), Poterba et al (1994), Angrist and Evans (1998), Jacobsen et al. (1999), Bronars and Grogger (1994), Udry (1996),Cameron and Taber (2001), or Angrist and Krueger (1999). Wooldridge's (2000) undergraduate textbook contains a computer exercise (15.14) that instructs students to look for a relationship between an observable (IQ) and an instrumental variable (closeness to college).

To fix ideas, let the Y_i be a continuous outcome of interest determined by:

$$Y_i = \alpha T_i + X_i'\Gamma_X + W_i^{c\prime}\Gamma^c + \xi_i \tag{1.1}$$

where T_i is a treatment variable.[2] The parameter of interest is α, the causal effect of T_i on Y_i. X_i is a vector of observed variables with coefficient vector Γ_X. X_i contains variables that are always observed, and W_i^c is a vector of additional characteristics that are relevant for determining the outcome which may or may not be observable to the econometrician. The final term, ξ_i, represents idiosyncratic shocks that are unrelated to the other components in the model. We use the notation $W_i'\Gamma$ to refer to the observed components of $W_i^{c\prime}\Gamma^c$ and $W_i^{u\prime}\Gamma^u$ to refer to its unobserved components. We can rewrite the model as:

$$Y_i = \alpha T_i + X_i'\Gamma_X + W_i'\Gamma + (W_i^{u\prime}\Gamma^u + \xi_i) \tag{1.2}$$

with the term in parentheses capturing all the unobservable components of the outcome.

The key idea in this paper is to model the relationship between T_i (or an instrument Z_i) and W_i^u. Our operational definition of "selection on unobservables is like selection on observables" is that the partial correlations of $W_i'\Gamma$ and $W_i^{u\prime}\Gamma^u$ with the treatment T_i are the same. The motivation for this involves thinking about the breakdown of exactly which characteristics are in W_i versus W_i^u as being determined by random chance. In addition, we view both W_i and W_i^u as having a large number of elements, none of which dominates in determining Y_i.[3] Dominant characteristics, like gender or schooling in a wage regression, are assumed always measured and in X_i. Finally, although the principal source of endogeneity bias here is that T_i is correlated with W_i^u, an additional source of bias stems from the correlation between W_i and W_i^u. In the context of a model for the determination of W, the correlations between the elements of W_i are informative about the nature of the correlation between W_i and W_i^u.

To illustrate the nature of the restrictions we use, consider the linear projection of T_i onto X_i, $W_i'\Gamma$ and $W_i^{u\prime}\Gamma^u$:

$$\text{Proj}(T_i|X_i, W_i'\Gamma, W_i^{u\prime}\Gamma^u) = \phi_0 + X_i'\phi_X + \phi W_i'\Gamma + \phi_u W_i^{u\prime}\Gamma^u. \tag{1.3}$$

[2]We will also discuss a binary dependent variable model in which the outcome is $1(Y_i > 0)$.

[3]We will utilize approximations that take the number of regressors in W^c (and W) to be large.

Our formalization of the idea that, after controlling for X_i, "selection on the unobservables is the same as selection on the remaining observables" is that:

Condition 1.

$$\phi_u = \phi.$$

One may contrast Condition 1 with the implication of the usual OLS orthogonality conditions:

Condition 2.

$$\phi_u = 0.$$

Roughly speaking, Condition 1 says that conditional on X_i, the part of Y_i that is related to the observables and the part related to the unobservables have the **same** relationship with T_i. Condition 2 says that the part of Y_i related to the unobservables has **no** relationship with T_i.

A projection like that in equation (1.3) will only be directly useful when an approximation for $W_i^{u\prime}\Gamma_i^u$ is available. When ξ_i is nonzero, the composite term $(W_i^{u\prime}\Gamma^u + \xi_i)$ is all that can be approximated. The analog of equation (1.3) is

$$\text{Proj}(T_i|X_i, W_i'\Gamma, (W_i^{u\prime}\Gamma^u + \xi_i)) = \phi_0 + X_i'\phi_X + \phi W_i'\Gamma + \phi_u(W_i^{u\prime}\Gamma^u + \xi_i). \tag{1.4}$$

With some abuse of notation, we continue to use ϕ_u as the last coefficient. Equal partial correlations of T_i with $W_i'\Gamma$ and $W^{u\prime}\Gamma^u$ in this projection will imply an inequality of ϕ and ϕ_u due to attenuation bias in the latter coefficient. This results in an intermediate condition 3 between the extremes of Conditions 1 and 2, defined as:

Condition 3.

$$0 \leq \phi_u \leq \phi \ \textit{if} \ \ \phi \geq 0$$
$$0 \geq \phi_u \geq \phi \ \textit{if} \ \phi < 0.$$

We propose two alternative estimators that differ in how they model the relationship between W_i and W_i^u. We refer to the first estimator as OU, which refers to using properties of observed ("O") covariates to infer the properties of unobserved ("U") covariates. OU amounts to estimating equation (1.2) using moment conditions that X and W_i are orthogonal to W_i^u and the restriction $\phi_u = \phi$. This estimates a lower (upper) bound on α if ϕ is

greater (less) than 0. It requires a high level assumption that implies, roughly speaking, that conditional on X_i, the coefficient of the regression of T_i on $(Y_i - \alpha T_i)$ has the same sign and is at least as large in absolute value as the coefficient of the regression of the part of T_i that is orthogonal to W_i on the part of $Y_i - \alpha T_i$ that is orthogonal to W_i. The high level assumption is required because the estimator does not make direct use of how the observed and unobserved explanatory variables are interrelated to assess the consequences of omitted variables that affect both the treatment and the outcome. Essentially, it treats W_i as exogenous, in common with the vast IV literature that focusses on endogeneity of T_i but treats the "controls"as exogenous. Furthermore, it does not provide a way to account for the fact that randomness in which elements of W_i^c are observed influences the distribution of the estimator. This estimator has been applied in Altonji, Elder and Taber (2005a, 2005b; hereafter, AET) to study the effectiveness of Catholic schools, as well as in a large number of other studies.[4] We complete the theoretical analysis of the estimator that is presented in preliminary form in AET (2002).

We also propose a second estimator that we believe is a more satisfactory approach because it relaxes the assumption that W_i is exogenous. In this second approach, we develop a method of moments procedure that uses the bounds on selection embodied in Condition 3 and also uses a factor structure to model the covariance between the observable and unobservable covariates. This structure allows us to infer properties of unobserved covariates based on the observed correlation structure of the observed covariates W_i. We show that this estimator, which we name OU-Factor, consistently identifies a set that contains α. We also provide a general bootstrap procedure that may be used to construct confidence regions for the identified set, as well as a less computationally demanding bootstrap procedure that seems to works well in practice.

Our paper is related in spirit to the rapidly growing emphasis in econometrics on partial identification and bound estimation. Some of these papers implicitly address omitted variables and selection bias. Indeed, we use the methods of Chernozhukov, Hong, and Tamer (2007) in studying the distribution of our estimator. Rosenbaum and Rubin (1983) and Rosenbaum (1995) propose examining the sensitivity of α to varying ϕ_u. As we've already noted, our paper has antecedents in the very large number of papers that examine the link

[4]AET and a number of subsequent papers measure the amount of selection on the index of observables that determine the outcome and then calculate a ratio of how large selection on unobservables would need to be in order to attribute the entire OLS estimate of α to selection bias. The approach that is closely related to the *OU* estimator.

between T_i (or an instrumental variable or an regression discontinuity indicator in an IV or regression discontinuity context) to the other covariates that influence Y_i and use the pattern as qualitative evidence about whether T_i is likely to be correlated with the omitted variables that influence Y_i. Our contribution is the development of a formal model of how the observed variables relate to the unobserved variables and the translation of the informal intuition that the patterns in the observables are informative about the unobservables into bounds estimators.[5]

The paper continues in Section 2, where we provide a formal model of which covariates are observed and which are unobserved. We provide an explicit set of assumptions under which Condition 1, Condition 2, and Condition 3 hold, and we elaborate on why Condition 3 is the most plausible of the three. In Section 3 we present the OU estimator. We also show that in general, Condition 1 is not sufficient to provide point identification of α. As a practical matter, this is not critical, because we focus on the use of Condition 3 to identify a range of admissible values for α. We then turn to the OU-Factor estimator based on specifying a factor structure for W_i^c. In Section 4 we provide some Monte Carlo evidence on the performance of OU and OU-Factor. We offer brief conclusions in Section 5.

2 Selection Bias and the Link Between the Observed and Unobserved Determinants of the Instrument and Outcome

In this section, we begin with a formal discussion of how the observables W_i are chosen from the full set W_i^c. This is the first step in developing a theoretical foundation for using the relationship between a potentially endogenous variable (or an instrument for that variable) and the observables to make inferences about the relationship between such a variable and the unobservables. In doing so, we provide a foundation for quantitatively assessing the importance of the bias from the unobservables. We then provide a set of conditions under which Condition 3 holds, which is central to OU and OU-factor.

[5] A large literature on survey non-response and to item nonresponse that leads to missing data on dependent variables or covariates for some observations, of which Kline and Santos (2010) is a recent example. We ignore item non-response and focus on missingvariables.

2.1 How are Observables Chosen?

We do not know of a formal discussion of how variables are chosen for inclusion in data sets. Here we make a few general comments that apply to many social science data sets. First, most large scale data sets such as the National Longitudinal Survey of Youth 1979, the British Household Panel, the Panel Study of Income Dynamics, and the German Socioeconomic Panel are collected to address many questions. Data set content is a compromise among the interests of multiple research, policy making, and funding constituencies. Burden on the respondents, budget, and access to administrative data sources serve as constraints. Obviously, content is also shaped by what is known about the factors that really matter for particular outcomes and by variation in the feasibility of collecting useful information on particular topics. Major data sets with large samples and extensive questionnaires are designed to serve multiple purposes rather than to address one relatively specific question. As a result, explanatory variables that influence a large set of important outcomes (such as family income, race, education, gender, or geographical information) are more likely to be collected. Because of limits on the number of the factors that we know matter, that we know how to collect, and that we can afford to collect, many elements of W_i^c are left out. This is reflected in the relatively low explanatory power of most social science models of individual behavior. Furthermore, in many applications, the treatment variable T_i is correlated with many of the elements of W_i^c.

These considerations suggest that Condition 2, which underlies single equation methods in econometrics, will rarely hold in practice. The optimal survey design for estimation of α would be to assign the highest priority to variables that are important determinants of *both* T_i and Y_i (it would also be to useful to collect potential instrumental variables that determine T_i but not Y_i). Condition 2 is based on the extreme assumption that surveys are sufficiently well designed to ensure that $\phi_u = 0$.

We next consider an assumption which is, in a sense, the other extreme. The constraints on data collection are sufficiently severe that it may be better to think of the elements of W_i as an approximately random subset of the elements of W_i^c, rather than being systematically chosen to eliminate bias. Indeed, a natural way to formalize the idea that "selection on the observables is the same as selection on the unobservables" is to treat observables and unobservables symmetrically by assuming that the observables are a random subset of a large number of underlying variables. We let the indicator S_j denote whether covariate j is

observed in the data set. We assume a symmetric treatment of observables and unobservables in this framework, that S_j is an *iid* binary random variable which is equal to one with probability P_S for all covariates in W_i^c.

In many applications a small set of exogenous variables may play a critical role in determining Y_i and T_i and are likely to be available in data sets appropriate for the research topic in question. These variables are represented by X_i. In AET's study of Catholic schools, Catholic religion is such a variable. We will also want to allow for the use of instrumental variables Z_i.

There are many reasons to include indiosycratic shocks ξ_i in the framework. In many problems outcomes are determined considerably after the treatment T_i , characteristics X_i, or instruments Z_i are determined. Consider the case of the effect of deciding to attend Catholic high schools on 12th grade test scores studied by AET. All of the regressors used in AET are measured in eighth grade. High school outcomes will be influenced by shocks that occur during the four years of high school, many of which are unanticipated at the time of decision regarding whether to attend a Catholic school. Given this sequencing, these shocks influence high school outcomes but cannot affect the probability of starting a Catholic high school. In addition, ξ_i will be needed to reflect random variability in a student's performance which has nothing to do with the decision to attend Catholic high school. Similarly, in health applications, ξ_i may reflect health shocks (such as an accident or exposure to a virus) that occur after the treatment choice T_i has been made.

2.2 Implications of Random Selection of Observables

We are now ready to consider the implications of random covariate selection from W_i^c. We begin with the general case. We first derive the probability limit of ϕ_u/ϕ as the number of covariates in W_i^c becomes large. We then consider several special cases.

We define outcomes as being determined by a sequence of models indexed by K^*, where K^* is the number of elements of W_i^c.[6] A natural part of the thought experiment in which K^* varies across models is the idea that the importance of each individual factor declines with K^*. We take the dimensions of X_i and Z_i as fixed.

[6]The "local to unity" literature in time series econometrics" (e.g., Stock, 1994) and the "weak instruments" literatures (e.g., Staiger and Stock, 1997) are other examples in econometrics in which the asymptotic approximation is taken over a sequence of models, which in the case of those literatures, depend on sample size. However, in these cases the purpose of the model sequence is provide a better guide to the asymptotic distribution of estimator, which is quite different from the present case.

Define $\mathcal{G}^{K^*}$ as the information set consisting of the realizations of the S_j, coefficients Γ_j, and the joint distribution of W_{ij} conditional on $j = 1, ..., K^*$. That is, $E(W_{ij} \mid \mathcal{G}^{K^*})$ is the mean for a given j, where the expectation is only over i, but $E(W_{ij})$ is an unconditional expectation over both i and j. It may be helpful to think of this data generation process as operating in two steps. First the "model" is drawn: for a given K^*, the joint distribution of W_{ij}, T_i, Z_i, ξ_i, and S_j are drawn. $\mathcal{G}^{K^*}$ represents this draw. In the second stage of the data generating process, individual data are constructed from these underlying distributions.

The two steps combine to generate Y_i as is represented in Assumption 1.

Assumption 1.

$$Y_i = \alpha T_i + X_i'\Gamma_X + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}\Gamma_j + \xi_i \tag{2.1}$$

where (W_{ij}, Γ_j) *is unconditionally stationary (indexed by* j*), and* X_i *includes an intercept.*

We use (and slightly abuse) non-standard notation in Assumption 1. Rather than explicitly indexing parameters by K^*, we suppress a K^* index on (W_{ij}, Γ_j) and bring a $\frac{1}{\sqrt{K^*}}$ out in front of the sum. This scaling guarantees that no particular covariate will be any more important *ex ante* than the others. It embodies the idea that a large number of components determine most outcomes in the social sciences. Any variables that play an outsized role in Y_i and Z_i are assumed to always be in the set of special regressors X_i. Note that Assumption 1 involves unconditional stationarity. Conditional on $\mathcal{G}^{K^*}$, the variance of the W_{ij} and the contribution of the W_{ij} to the variance of Y_i will differ across j.

Throughout we will project all variables on X_i and take residuals to remove X_i from the regression. We will use "tildes" to denote the residuals from these projections, so we define

$$\begin{aligned}
\widetilde{W}_{ij} &\equiv W_{ij} - Proj(W_{ij} \mid X_i; \mathcal{G}^{K^*}) \\
\widetilde{T}_i &\equiv T_i - Proj(T_i \mid X_i; \mathcal{G}^{K^*}) \\
\widetilde{Z}_i &\equiv Z_i - Proj(Z_i \mid X_i; \mathcal{G}^{K^*}) \\
\widetilde{Y}_i &\equiv Y_i - Proj(Y_i \mid X_i; \mathcal{G}^{K^*})
\end{aligned}$$

where $Proj$ denotes a linear projection.[7] Let $\sigma_{j,\ell}^{K^*} = E\left(\widetilde{W}_{ij}\widetilde{W}_{i\ell} \mid \mathcal{G}^{K^*}\right)$. To guarantee that $var(Y_i)$ is bounded as K^* becomes large, we assume that

[7]Formally, the linear projection projection of a generic Y_i on a generic X_i is defined by $X_i'\delta$ where δ satisfies $E[(Y_i - X_i'\delta)X_i \mid \mathcal{G}^{K^*}] = 0$. Hereafter, this projection is meant to be the population projection conditional on G^{K^*}, i.e., for a very large N, but with K^* draw of $G^{K^*}$and fixed.

Assumption 2.

$$0 < \lim_{K^* \to \infty} \frac{1}{K^*} \sum_{j=1}^{K^*} \sum_{\ell=1}^{K^*} E(\sigma_{j,\ell}^{K^*} \Gamma_j \Gamma_\ell) < \infty \ ; \ \lim_{K^* \to \infty} Var \left(\frac{1}{K^*} \sum_{j=1}^{K^*} \sum_{\ell=1}^{K^*} \sigma_{j,\ell}^{K^*} \Gamma_j \Gamma_\ell) \right) \to 0 \ .$$

The next two assumptions guarantee that $cov(\widetilde{Z}_i, \widetilde{Y}_i)$ is well behaved as K^* grows.

Assumption 3. *For any $j = 1, ..., K^*$,define $\mu_j^{K^*}$ so that*

$$E\left(\widetilde{Z}_i \widetilde{W}_{ij} | \mathcal{G}^{K^*}\right) = \frac{\mu_j^{K^*}}{\sqrt{K^*}} \ .$$

We assume

$$E(\mu_j^{K^*} \Gamma_j) < \infty \ ; \ \lim_{K^* \to \infty} Var \left(\frac{1}{K^*} \sum_{j=1}^{K^*} \mu_j^{K^*} \Gamma_j \right) \to 0 \ .$$

Below we prove that Assumptions 2 and 3 are satisfied by a factor model for $\widetilde{Z}_i$ and $\widetilde{W}_{ij}$, which is central to the OU-Factor estimator. In the appendix, we also illustrate the assumptions using a second example in which the $\widetilde{W}_{ij}$ are linked across j through an MA model. The MA example is the most straightforward when one examines Assumptions 1 and 2 given that the two assumptions refer to observables as though they have a sequential ordering.

Finally, we provide assumptions about the process under which observables are chosen. Consider the case discussed above in which variables are chosen at random:

Assumption 4. *For $j = 1, ..., K^*$, S_j is independent and identically distributed with $0 < \Pr(S_j = 1) \equiv P_s \leq 1$. S_j is also independent of all other random variables in the model. If $var(\xi) \equiv \sigma_\xi^2 = 0$, then $P_S < 1$.*

Assumption 5. *ξ_i is mean zero and uncorrelated with $\widetilde{Z}_i$ and $\widetilde{W}_{ij}$.*

First we consider the relationship between ϕ and ϕ_u in the general case with nonzero $var(\xi_i)$ and then derive three key special cases.

Note that our asymptotic analysis is nonstandard in two respects. First, we are allowing the number of underlying explanatory variables, K^*, to get large. Second, the random variable $\widetilde{W}_{ij}$ is different from the random variables Γ_j and S_j in the following way. For each j we draw one observation on Γ_j and S_j which is the same for every person in the population; however, each individual i draws her own $\widetilde{W}_{ij}$.

Theorem 1. *Define ϕ and ϕ_u such that*

$$Proj\left(\tilde{T}_i \mid \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\tilde{W}_{ij}\Gamma_j, \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\tilde{W}_{ij}\Gamma_j+\xi_i; \mathcal{G}^{K^*}\right)$$
$$=\phi\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\tilde{W}_{ij}\Gamma_j\right)+\phi_u\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\tilde{W}_{ij}\Gamma_j+\xi_i\right).$$

Then under assumptions 1-3 and 4-5, if the probability limit of ϕ is nonzero, then

$$\frac{\phi_u}{\phi} \xrightarrow[K^*\to\infty]{p} \frac{\left(1-P_s\right)A}{\left(1-P_s\right)A+\sigma_\xi^2}$$

where

$$A \equiv \lim_{K^*\to\infty} E\left(\frac{1}{K^*}\sum_{j=1}^{K^*}\sigma_{j,j}^{K^*}\left(\Gamma_j\right)^2\right).$$

If the probability limit of ϕ is zero, then the probability limit of ϕ_u is also zero.

(Proof in Appendix-available from the Authors)

Next we consider three separate cases which we present as corollaries. We omit the proofs of these as they follow immediately from the proof of Theorem 1.

Corollary 1. *When $\sigma_\xi^2 = 0$,*

$$plim(\phi-\phi_u)=0.$$

Corollary 1 states that the coefficients of the projection of $\tilde{T}_i$ onto $\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\tilde{W}_{ij}\Gamma_j$ and $\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\tilde{W}_{ij}\Gamma_j$ approach each other with probability one as K^* becomes large. The other extreme is the case in which all the important control variables that affect both $\tilde{Z}$ and $\tilde{Y}$ are included in the model, so the variation in the composite error term $\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\tilde{W}_{ij}\Gamma_j+\xi_i$ arises from ξ only:

Corollary 2. *When $P_s = 1$,*

$$plim(\phi_u)=0.$$

What about the case in which selection on observables is stronger than selection on unobservables but there is still some selection on unobservables? This corresponds to the case in which $var(\xi) > 0$ and $P_s < 1$. The next Corollary considers this case:

Corollary 3. *When* $0 < P_s < 1$ *and* $\sigma_\xi^2 > 0$,

either

$$0 < \ plim(\phi_u) < plim(\phi),$$

or

$$plim(\phi) < plim(\phi_u) < 0,$$

or

$$0 = plim(\phi_u) = plim(\phi).$$

This Corollary plays a key role in the estimator below.

3 Estimators of α

We now discuss ways to estimate α. In Section 4.1 We set the stage by reviewing the OU estimator introduced in AET (2002, 2005). Then we present OU-Factor, beginning with the factor model of W_i^c that it requires.

But before turning to the estimators, we provide an explicit model for $\widetilde{Z}$ which we use for both estimators.

Assumption 6.

$$\widetilde{Z}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} \widetilde{W}_{ij}\beta_j + \psi_i, \tag{3.1}$$

where (i) ψ_i *is independent of all of the elements of* $\widetilde{W}_i^c$*.(ii)* β_j *is a stationary process with finite second moments.* β_j *may be correlated with* Γ_j*.*

It is convenient to rewrite the model for $\widetilde{Z}_i$ as

$$\widetilde{Z}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K} \widetilde{W}_{ij}\beta_j + u_i \tag{3.2}$$

where $u_i = \frac{1}{\sqrt{K^*}} \sum_{j=K+1}^{K^*} \widetilde{W}_{ij}\beta_j + \psi_i$, and all variables are residuals from linear projections onto the space of X_i.

3.1 The OU Estimator

We repeat the outcome equation here for convenience

$$\begin{aligned}\widetilde{Y}_i &= \alpha\tilde{T}_i + \widetilde{W}_i'\Gamma + (\widetilde{W}_i^{w\prime}\Gamma^u + \xi_i) \\ &\equiv \alpha\widetilde{T}_i + \widetilde{W}_i'\Gamma + \varepsilon_i\end{aligned}$$

Instrumental variables estimation of α uses the standard moment conditions $E(\widetilde{W}_i\varepsilon_i) = 0$ and the IV moment equation $E(\widetilde{Z}_i\varepsilon_i) = 0$. The simplest form of the OU estimator replaces the moment equation $E(\widetilde{Z}_i\varepsilon_i) = 0$ with condition 3. In most applications of *OU* to date either $T_i = 1(Z_i > 0)$ or $T_i = Z_i$, so we focus on this case:

$$\widetilde{T}_i = \widetilde{Z}_i = \frac{1}{\sqrt{K^*}}\widetilde{W}_i'\beta + u_i$$

A problem, however, is that despite the fact that mean independence of ε_i and $\widetilde{W}$ is maintained in virtually all observational studies of selection problems, is not likely to hold. Without it, α is not identified even if one has a valid exclusion restriction.[8] Our discussion of how the observables are determined makes clear that this is hard to justify in most settings-including ours. If the observables are correlated with one another, as in most applications, then the observed and unobserved determinants of Y_i are also likely to be correlated. This will lead to an inconsistent estimator whether one uses $E(\widetilde{Z}_i\varepsilon_i)$ as a moment condition or Condition 3. Note that this is not a problem with Theorem 1. That theorem did not require $E(\widetilde{Z}_i\varepsilon_i) = 0$. The problem is that this theorem involves the true value of Γ, but we need an assumption analogous to $E(\widetilde{Z}_i\varepsilon_i) = 0$ in order to consistently estimate Γ.

AET essentially assume away this problem (which the OU-factor does address). They assume that $E(\varepsilon_i \mid \widetilde{W}_i)$ is linear, and define G and e to be the slope vector and error term of the "reduced forms":

$$E\left(\widetilde{Y}_i - \alpha\widetilde{T} \mid \widetilde{W}\right) \equiv \widetilde{W}'G \tag{3.3}$$

$$\widetilde{Y} - E\left(\widetilde{Y} - \alpha\widetilde{T} \mid \widetilde{W}\right) \equiv e. \tag{3.4}$$

Let $\phi_{W'G}$ and ϕ_e be the coefficients of the projection of T on $W'G$ and e in a regression model that includes X. Note that under the assumption that if $E(\widetilde{Z}_i\varepsilon_i) = 0$, then $G = \Gamma$ and under the assumptions of Theorem 1, $0 \leq \phi_e \leq \phi_{W'G}$ when $\phi_{W'G} > 0$. AET show that this is true under the following more general (though not necessarily easy to interpret) condition:

[8]The exception is when the instrument is uncorrelated with W_i (and X_i) as well as ξ_i, as when the instrument is randomly assigned in an experimental setting.

Assumption 7.

$$\lim_{K^*\to\infty} \frac{\sum_{\ell=1}^{K^*} E\left(\widetilde{W}_{ij}\widetilde{W}_{ij-\ell}\right) E\left(\beta_j\Gamma_{j-\ell}\right)}{\sum_{\ell=1}^{K^*} E\left(\widetilde{W}_{ij}\widetilde{W}_{ij-\ell}\right) E\left(\Gamma_j\Gamma_{j-\ell}\right)} = \lim_{K^*\to\infty} \frac{\sum_{\ell=1}^{K^*} E\left(\widetilde{\widetilde{W}}_{ij}\widetilde{\widetilde{W}}_{ij-\ell}\right) E\left(\beta_j\Gamma_{j-\ell}\right)}{\sum_{\ell=1}^{K^*} E\left(\widetilde{\widetilde{W}}_{ij}\widetilde{\widetilde{W}}_{ij-\ell}\right) E\left(\Gamma_j\Gamma_{j-\ell}\right)}, \tag{3.5}$$

for the set of variables W_{ij} *in* $j = 1, ..., K^*$,

where $\widetilde{\widetilde{W}}_{ij}$ is the component of $\widetilde{W}_{ij}$ that is orthogonal to the observed variables (X_i, W_i), for all elements of W_i^*. Roughly speaking (3.5) says that the regression of T_i on $\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \xi_i\right)$ is equal to the regression of the part of $\widetilde{T}_i$ that is orthogonal to $\widetilde{W}_i$ on the corresponding part of $\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \xi_i\right)$. This condition holds under the standard assumption $E(\varepsilon_i \mid \widetilde{W}_i; \mathcal{G}^{K^*}) = 0$, in which case G and e_i equal Γ and ε_i, respectively. However, $E(\varepsilon_i \mid \widetilde{W}_i; \mathcal{G}^{K^*}) = 0$ is not necessary for (3.5).[9]

Theorem 2. *Define* $\phi_{\widetilde{W}'G}$ *and* ϕ_e *such that*

$$\begin{aligned} &Proj\left(\widetilde{Z}_i \mid \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\widetilde{W}_{ij}G_j, \frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\widetilde{W}_{ij}\Gamma_j + \xi_i; \mathcal{G}^{K^*}\right) \\ &\quad = \phi_{\widetilde{W}'G}\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j\widetilde{W}_{ij}\Gamma_j\right) + \phi_e\left(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\left(1-S_j\right)\widetilde{W}_{ij}\Gamma_j + \xi_i\right). \end{aligned}$$

Then under assumptions 1-5 and 7, as K^* *gets large, then*

$$\frac{\phi_e}{\phi_{\widetilde{W}'G}} \xrightarrow{p} \frac{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W}_{ij}\widetilde{W}_{ij-\ell}\right) E\left(\Gamma_j\Gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W}_{ij}\widetilde{W}_{ij-\ell}\right) E\left(\Gamma_j\Gamma_{j-\ell}\right) + \sigma_\xi^2}$$

if the probability limit of ϕ *is nonzero. If the probability limit of* $\phi_{\widetilde{W}'G}$ *is zero then the probability limit of* ϕ_e *is also zero.*

(Proof in Appendix available from Authors)

The upshot is that one can work with the system

[9] For example, one can show that (3.5) will also hold if $E\left(\beta_j\Gamma_{j-\ell}\right)$ is proportional to $E\left(\Gamma_j\Gamma_{j-\ell}\right)$ regardless of the correlations among the W_j.

$$
\begin{aligned}
\widetilde{Y}_i &= \alpha \widetilde{T}_i + \frac{1}{\sqrt{K^*}} \widetilde{W}_i' G + e_i. \\
\widetilde{T}_i &= \frac{1}{\sqrt{K^*}} \widetilde{W}_i' \beta + u_i \\
0 &\leq \left| \frac{cov(u_i, e_i | \mathcal{G}^{K^*})}{var(e_i \mid \mathcal{G}^{K^*})} \right| \leq \left| \frac{Cov(\widetilde{W}_i' \beta, \widetilde{W}_i' G | \mathcal{G}^{K^*})}{Var(\widetilde{W}_i' G | \mathcal{G}^{K^*})} \right|
\end{aligned}
$$

and estimate the set of α values that satisfy the above inequality restrictions. In practice, AET find that the lower bound is obtained when the equality of selection condition $\frac{cov(u_i, e_i | \mathcal{G}^{K^*})}{var(e_i | \mathcal{G}^{K^*})} = \frac{Cov(\widetilde{W}_i' \beta, \widetilde{W}_i' G | \mathcal{G}^{K^*})}{Var(\widetilde{W}_i' G | \mathcal{G}^{K^*})}$ is imposed and the upper bound corresponds to the case in which $\widetilde{T}_i$ is treated as exogenous, with $\frac{cov(u_i, e_i | \mathcal{G}^{K^*})}{var(e_i | \mathcal{G}^{K^*})} = 0$.

One can perform statistical inference accounting for variation over i conditional on which W_i are observed in the usual way, and we omit the details. However, there is no obvious way to account for random variation due to the draws of S_j which is another reason one might prefer OU-factor

3.2 OU-Factor: A Bounds Estimator Based on a Factor Model of $\widetilde{W}_{ij}$

3.2.1 A Factor Model of $\widetilde{W}_{ij}$

The biggest issue with the OU estimator is that it required assumption (3.5) which in general is hard to justify in a model in which the W_{ij} are chosen randomly from the set of W_i^c. Relaxing this assumption requires building a model of the relationship between the W_{ij} that we observe and the W_{ij} that we don't observe. We do this by building factor model of $\widetilde{W}_{ij}$, which is central to the estimator proposed below. The factor model is a convenient way to model the relationship among the covariates. We assume that $\widetilde{W}_{ij}$ has a factor structure

$$
\widetilde{W}_{ij} = \frac{1}{\sqrt{K^*}} \widetilde{F}_i' \Lambda_j + v_{ij}, \ j = 1, ..., K^*, \tag{3.6}
$$

where $\widetilde{F}_i$ is an r dimensional mean zero vector of factors. We treat r as finite, so while the dimension of $\widetilde{W}_{ij}$ grows, the number of factors remains constant. Recall that $\widetilde{W}_{ij}$ is the residual from the projection of W_{ij} upon X_i.We normalize the variance/covariance matrix of $\widetilde{F}_i$ be to the identity matrix. Define $\sigma_j^2 \equiv E(v_{ij}^2 \mid \mathcal{G}^{K^*})$, $j = 1, ..., K^*$. It is important to contrast our work with other models using factor structures such as Cuhna, Heckman, and

Schennach (2010). In much of this other work what is driving outcomes is the factors themselves. Our model is quite different. We continue to assume that outcomes are determined by $\widetilde{W}_{ij}$ itself. We use the factor structure only as a model of the covariance structure of W_i^c.

We define a model for $\tilde{T}_i$:

$$\widetilde{T}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S\widetilde{W}_{jij}\delta_j + \left[\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j)\, \widetilde{W}_{ij}\delta_j + \omega_i \right], . \tag{3.7}$$

For convenience repeat the equation for $\widetilde{Z}_i$ and $\widetilde{T}_i$:

$$\widetilde{Z}_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j\widetilde{W}_{ij}\beta_j + \left[\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j)\, \widetilde{W}_{ij}\beta_j + \psi_i \right], \tag{3.8}$$

$$\widetilde{Y}_i = \alpha\widetilde{T}_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j\widetilde{W}_{ij}\Gamma_j + \left[\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j)\, \widetilde{W}_{ij}\Gamma_j + \xi_i \right]. \tag{3.9}$$

The ω_i, ξ_i and ψ_i are assumed independent of all of the $\widetilde{W}_{ij}$ the instrument error term ψ_i is assumed to be correlated with the treatment error ω_i but not the outcome error ξ_i. The brackets in each of these expressions collect unobservable terms. Note that if all the elements of W_i^c were observed ($S_j = 1$ for all j), our framework reduces to the standard instrumental variables setup.

The stochastic structure of the model is that Λ_j, Γ_j, β_j and σ_j^2 differ across j, but are identical for all individuals in the population. We redefine $\mathcal{G}^{K^*}$to refer to aspects of the model of $\widetilde{W}_{ij}$, $\widetilde{T}_i$, $\widetilde{Y}_i$, and $\widetilde{Z}_i$, that do not vary across individuals:

$$\mathcal{G}^{K^*} = \left\{ (\Gamma_j, \beta_j, \delta_j, \Lambda_j, \sigma_j^2, S_j) \text{ for } j = 1, ..., K^* \right\}.$$

For estimation, we make the following additional assumptions.

Assumption 8. *(i)* $\left(\Gamma_j, \beta_j, \delta_j, \Lambda_j, \sigma_j^2\right)$ *is i.i.d with fourth moments; (ii) the error terms* $(\omega_i, \psi_i, \xi_i)$ *are mean zero with finite second moments and are independent of* $\widetilde{W}_i^c$*;* ψ_i *and* ξ_i *are uncorrelated,* ω_i *and* ψ_i *are correlated.*

Assumption 8 (ii) allows for there to be a component of $\widetilde{T}_i$, that is correlated with the instrument $\tilde{Z}_i$ but uncorrelated with the observed and unobserved determinants of $\widetilde{Y}_i$ allowing identification of α. In the Appendix we verify that the factor model of $\widetilde{W}_i$ in conjunction with the model (2.1) for $\widetilde{Y}_i$ (3.1) for $\widetilde{Z}_i$, and (3.7) for $\widetilde{T}_i$, satisfies Assumption 1 of Theorem 1.

3.2.2 An Estimator of an Admissible Set for α

We use the factor model to directly address the problem posed by the fact that the elements of $\widetilde{W}_i$ (as well as $\widetilde{T}_i$, and $\widetilde{Z}_i$) are likely to be correlated with the composite error term for $\widetilde{Y}_i$. We proceed under the following assumptions. First, the econometrician observes K, the number of observed covariates in $\widetilde{W}_i$ but not K^*, the number of and unobserved covariates in W_i^c. Second, the econometrician observes the joint distribution of Y_i, Z_i, T_i, X_i and $\{W_{ij} : S_{ij} = 1\}$. Third, we assume that $K/K^* \to P_{s0}$. Fourth, we assume that N becomes large faster than K^*, with $\frac{K^*}{N} \to 0$, so that we can take sequential limits. This seems like a good approximation in problems where K and K^* are large, but not for problems in which the number of variables that determine Y_i is small.

In general the model is not point identified, so we provide an estimator of a set that contains the true values. The key subset of the parameter vector of our model is $\theta = \{\alpha, \phi, P_s, \sigma_\xi^2)$. The parameter α is the key parameter denoted the treatment effect. P_s is the probability that $S_j = 1$, σ_ξ^2 is the variance of ξ_i, and ϕ is the coefficient in front of the observable index when we project $Z_i$onto the observables and the unobservables as in Theorem 1. The true value of θ is $\theta_0 = \{\alpha_0, \phi_0, P_{s0}, \sigma_{\xi 0}^2)$ which lies in the compact set $\bar{\Theta}$. Our approach is to estimate a set $\widehat{\Theta}$ that asymptotically will contain the true value θ_0. The key restrictions on the parameter set are

$$0 < P_{s0} \leq 1, \text{ and} \tag{3.10}$$

$$\sigma_{\xi 0}^2 \geq 0. \tag{3.11}$$

The case in which $P_{s0} = 1$ corresponds to the standard IV case represented by Condition 2, while $\sigma_{\xi 0}^2 = 0$ corresponds to the "unobservables like observables" case represented by Condition 1. We construct an estimate of the set of values of α by estimating the set of θ that satisfy all of the conditions and then projecting onto the α dimension. We then discuss construction of confidence intervals. While the upper and lower bound of the estimated set does not have to correspond to the cases in which $P_{s0} = 1$ and $\sigma_{\xi 0}^2 = 0$, in practice we find that it does.

It will be useful to make use of matrix notation. We assume that the variables are ordered so that $j = 1, .., K$ corresponds to the K observed covariates in W^c. Unless indicated otherwise,

- For a generic variable $B_i, i = 1, ..., N$, B will represent the $N \times 1$ vector.

- For a generic variable $B_j, j = 1, ..., K^*$, B will represent the $K \times 1$ vector of observable characteristics and B^* will represent the full $K^* \times 1$ vector.

- For a generic variable $B_{ij}, i = 1, .., N, j = 1, ..., K^*$, B will represent the $N \times K$ matrix of observable characteristics, $B^*$the full $N \times K^*$matrix of covariates, and B_i represents the $K \times 1$ vector of B_{ij} for a given i.

- We also employ the convention of using capital letters for matrices so, for example, the matrix version of v_{ij} will be written as V.

Given the large amount of notation we concentrate on the 1 factor case ($r = 1$), so $\widetilde{F}_i$ and Λ_j are scalars. We fully expect that the results generalize to the multiple factor case. We now present the estimator, which has two stages.

Stage 1

In the first stage we estimate the Λ_j and σ_j^2 for all observed variables. The moment conditions are the K equations

$$E\left(\widetilde{W}_{ij}^2|\mathcal{G}^{K^*}\right) = \frac{1}{K^*}\Lambda_j^2 + \sigma_j^2;\ j : S_j = 1\ , \tag{3.12}$$

and the $K \cdot (K-1)/2$ equations

$$E\left(\widetilde{W}_{ij_1}\widetilde{W}_{ij_2}|\mathcal{G}^{K^*}\right) = \frac{1}{K^*}\Lambda_{j_1}\Lambda_{j_2}\ ;\ j_1,\ j_2 : S_{j_1} = S_{j_2} = 1,\ j_1 \neq j_2\ . \tag{3.13}$$

This is a standard GMM problem. As N grows we will obtain $\sqrt{N}$ consistent estimates of $\frac{1}{\sqrt{K^*}}\Lambda_j$ for each j and for $\widehat{\sigma_j^2}$ by using the sample analogues to (3.12) and (3.13). Note that K^* is not known since it depends on the number of unobserved variables. However, the econometrician knows K. To simplify the exposition we define $\widehat{\lambda_j}$ to be the GMM estimate of the parameter $\sqrt{K} \times \frac{1}{\sqrt{K^*}}\Lambda_j \approx \sqrt{P_{S0}}\Lambda_j$ and λ to be the corresponding vector. In practice we just replace the left side of the equations by $\frac{1}{N}\sum_{i=1}^N \left(\widetilde{W}_{ij_1}\widetilde{W}_{ij_2}\right)$ and choose $\widehat{\lambda_j}$ and $\widehat{\sigma_j^2}$ as the values that minimize the unweighted squared difference between the values of $\frac{1}{N}\sum_{i=1}^N \left(\widetilde{W}_{ij_1}\widetilde{W}_{ij_2}\right)$ and the predictions summarized in the moment conditions above.

Stage 2

We estimate the rest of the parameters in a second stage. If we knew α_0 we could estimate Γ conditional on α_0 by taking advantage of the K moment conditions corresponding to the j for which $S_j = 1$,

$$\begin{aligned}\sqrt{K^*}E\left[\widetilde{W}_{ij}\left(\widetilde{Y}_i - \alpha_0\widetilde{T}_i)\right)|\mathcal{G}^{K^*}\right] &= \sqrt{K^*}E\left[\begin{array}{c}\left(\frac{1}{\sqrt{K^*}}\widetilde{F}_i\Lambda_j + v_{ij}\right)\cdot \\ \left(\frac{1}{\sqrt{K^*}}\sum_{\ell=1}^{K^*}\frac{1}{\sqrt{K^*}}\widetilde{F}_i\Lambda_\ell\Gamma_\ell + \frac{1}{\sqrt{K^*}}\sum_{\ell=1}^{K^*}v_{ij}\Gamma_\ell\right) + \xi_i|\mathcal{G}^{K^*}\end{array}\right] \\ &= \Lambda_j\left(\frac{1}{K^*}\sum_{\ell=1}^{K^*}\Lambda_\ell\Gamma_\ell\right) + \sigma_{vj}^2\Gamma_j \\ &\xrightarrow{p} \Lambda_j E(\Lambda_\ell\Gamma_\ell) + \sigma_{vj}^2\Gamma_j.\end{aligned}$$

We work with the sample analog of the above expression,

$$\left[\begin{array}{c}\sqrt{K^*}\frac{1}{N}\widetilde{W}'\left(\widetilde{Y} - \alpha_0\widetilde{T}\right)\end{array}\right] = \left[\begin{array}{c}\frac{1}{K}\frac{1}{P_{s0}}\widetilde{\lambda}\widetilde{\lambda}'\Gamma + \Sigma\Gamma\end{array}\right],$$

where Σ is the diagonal matrix composed of the σ_j^2 terms. Thus, for the parameter θ we can construct the estimator

$$\widehat{\Gamma}(\theta) \approx \left[\frac{1}{P_sK}\widetilde{\lambda}\widetilde{\lambda}' + \widehat{\Sigma}\right]^{-1}\frac{1}{N}\widetilde{W}'\left(\widetilde{Y} - \alpha\widetilde{T}\right), \tag{3.14}$$

where we define $\widehat{\Sigma}$ to be the diagonal matrix composed of the $\widehat{\sigma_j^2}$, which is estimated in the first stage.

Theorem 1 presents the main idea behind our approach which is a projection of Z_i onto the observable index and the unobservable index. One may show that the coefficient in from of the observables.

$$\phi_0 = \frac{\left[E(\Gamma_j\Lambda_j)E(\beta_j\Lambda_j) + E(\Gamma_j\beta_j\sigma_j^2)\right]\left[P_{s0}\left(1 - P_{s0}\right)E(\Gamma_j^2\sigma_j^2) + P_{s0}\sigma_{\xi0}^2\right]}{\sigma_{\xi0}^2\left[P_{s0}^2E(\Gamma_j\Lambda_j)^2 + P_{s0}E(\Gamma_j^2\sigma_j^2)\right] + \left[E(\Gamma_j\Lambda_j)^2 + E(\Gamma_j^2\sigma_j^2)\right]\left(1 - P_{s0}\right)P_{s0}E(\Gamma_j^2\sigma_j^2)}.$$

Our first two equations represent the moment conditions associated with a projection of Z_i upon the observable index $\widetilde{W}_i'\widehat{\Gamma}(\theta)$ and the unobservables which we can write as $\widetilde{Y}_i -$

$\alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)$. We define these as

$$q^1_{N,K^*}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \widetilde{W}_i'\widehat{\Gamma}(\theta) \times \left[\widetilde{Z}_i - \phi\widetilde{W}_i'\widehat{\Gamma}(\theta) - \phi\frac{(1-P_s)\widehat{\Gamma}(\theta)'\widehat{\Sigma}\widehat{\Gamma}(\theta)}{(1-P_s)\widehat{\Gamma}(\theta)\widehat{\Sigma}\widehat{\Gamma}(\theta) + P_s\sigma_\xi^2}\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\right] \tag{3.15}$$

$$q^2_{N,K^*}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left(\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\right) \times \left[\widetilde{Z}_i - \phi\widetilde{W}_i'\widehat{\Gamma}(\theta) - \phi\frac{(1-P_s)\widehat{\Gamma}(\theta)'\widehat{\Sigma}\widehat{\Gamma}(\theta)}{(1-P_s)\widehat{\Gamma}(\theta)'\widehat{\Sigma}\widehat{\Gamma}(\theta) + P_s\sigma_\xi^2}\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\right] \tag{3.16}$$

To understand the first two equations, note that when $\sigma_\xi^2 = 0$ they reduce to

$$\begin{aligned} q^1_{N,K^*}(\theta) &= \frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{W}_i'\widehat{\Gamma}(\theta)\left[\widetilde{Z}_i - \phi\widetilde{W}_i'\widehat{\Gamma}(\theta) - \phi\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\right]\right) \\ q^2_{N,K^*}(\theta) &= \frac{1}{N}\sum_{i=1}^{N}\left(\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\left[\widetilde{Z}_i - \phi\widetilde{W}_i'\widehat{\Gamma}(\theta) - \phi\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\right]\right). \end{aligned}$$

These are the orthogonality conditions of a linear regression of $\widetilde{Z}_i$ on $(\widetilde{W}_i'\widehat{\Gamma}(\theta))$ and $(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta))$ when the two regression coefficients are restricted to be the same. They are the empirical analogue of Corollary 1 of Theorem 1. Equations (3.15) and (3.16) are more complicated because the presence of ξ_i leads to attenuation bias on the regression coefficient on $\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)$.

When $P_S = 1$, the second equation reduces to

$$q^2_{N,K^*}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right)\left[\widetilde{Z}_i - \phi\widetilde{W}_i'\widehat{\Gamma}(\theta)\right]\right).$$

In this case $\widehat{\Gamma}(\theta)$ could be estimated as the coefficient vector from a linear regression of $\left(\widetilde{Y}_i - \alpha\widetilde{T}_i\right)$ on $\widetilde{W}_i$. (Our estimator is asymptotically equivalent to this with K^* fixed and N getting large.) In that case, $\widetilde{W}_i'\widehat{\Gamma}(\theta)$ would have to be orthogonal to the error term, so this equation would reduce further to

$$q^2_N(\alpha,\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{Y}_i - \alpha\widetilde{T}_i - \widetilde{W}_i'\widehat{\Gamma}(\theta)\right) \times \widetilde{Z}_i,$$

which is the standard IV moment equation.

The third equation essentially represents the fact that the total sum of squares must be the sum of its components:

$$q^3_{N,K^*}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{Y}_i - \alpha\widetilde{T}_i\right)^2 - \left[\left(\frac{\widehat{\Gamma}(\theta)'\widehat{\lambda}}{P_s}\right)^2 + \frac{\widehat{\Gamma}(\theta)'\widehat{\Sigma}\widehat{\Gamma}(\theta)}{P_s} + \sigma^2_\xi\right] \tag{3.17}$$

Turning to (3.17), $q^3_{N,K^*}(\theta)$ is the difference between the total sum of squares of $\left(\widetilde{Y}_i - \alpha\widetilde{T}_i\right)$ in the data for the hypothesized value of α and the sum of squares implied by the model estimate.

We will show that when evaluated at θ_0 these equations $q^1_{N,K^*}(\theta)$, $q^2_{N,K^*}(\theta)$, and $q^3_{N,K^*}(\theta)$ converge to zero as N and K^* grow.

We define the estimator $\widehat{\Theta}$ as the set of values of θ that minimize the criterion function

$$Q_{N,K^*}(\theta) = Kq_{N,K^*}(\theta)'\Omega q_{N,K^*}(\theta),$$

where

$$q_{N,K^*}(\theta) = \left[\begin{array}{ccc} q^1_{N,K^*}(\theta) & q^2_{N,K^*}(\theta) & q^3_{N,K^*}(\theta) \end{array}\right]'$$

and Ω is some predetermined symmetric positive definite weighting matrix and we can write the Cholesky decomposition as. $\Omega = LL'$.

3.3 Consistency of the Estimator

In this section we prove consistency using the standard methods from Chernozhukov, Hong, and Tamer (2007). Define $Q_0(\theta)$ as the probability limit of $Q_{N,K^*}(\theta)$ as N and K^* get large. Specifically we use sequential limits assuming that N grows faster than K^*. The identified set, Θ_I, is defined as the set of values that minimize $Q_0(\theta)$. We verify the conditions in Chernozhukov, Hong, and Tamer (2007) to show that the Hausdorff distance between $\widehat{\Theta}$ and Θ_I converges in probability to zero and that $\theta_0 \in \Theta_i$. Thus as the sample gets large our estimate of $\widehat{\Theta}$ will contain the true value with probability approaching 1.

We maintain the assumptions of the factor model W and Assumption (8). In addition we add Assumptions 9 and 10 below.

Assumption 9. *$\bar{\Theta}$ is compact with the support of P_s bounded below by $p^\ell_s > 0$.*

Assumption 10. *The dimension of $\tilde{F}_i$ is 1*

Define $d_h(\cdot,\cdot)$to be Hausdorff distance as defined in Chernozhukov, Hong, and Tamer (2007).

Theorem 3. *Assuming our factor model for W, and Assumptions 8-10,* $d_h(\widehat{\Theta}, \Theta_I)$ *converges in probability to zero and* $\theta_0 \in \Theta_I$.

(Proof in Appendix)

One can form a set estimator for α_0 just by taking the projection of $\widehat{\Theta}$ onto α. That is, we can define this set as

$$\widehat{A} \equiv \left\{\alpha : \text{ there exists some value of } (\phi, P_s, \sigma_\xi^2) \text{ such that } \{\alpha, \phi, P_s, \sigma_\xi^2\} \in \widehat{\Theta}\right\}$$

3.4 Constructing Confidence Intervals

In this section we discuss confidence interval construction. We start with the ideal procedure one would use given unlimited computing resources. We then discuss a more practical approach, which is the parametric bootstrap we use in the Monte Carlos below.

3.4.1 A General Procedure

Before discussing inference it is useful to step back and consider our basic approach. In terms of identification we have four parameters $(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ but only 3 equations: the population and limit of the sequence of models for (q_N^1, q_N^2, q_N^3).[10] However, we also have limits on the parameter space. In particular $0 < P_S \leq 1$ and $\sigma_\xi^0 \geq 0$. While we cannot get a point estimator for $(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$, we construct the set estimator $\widehat{\Theta}$ for this four dimensional parameter. Our set estimate for α_0 is just the set of α that lie within this identified set.

We can construct a confidence region in the analogous manner. That is, we could first construct a confidence set for $(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ and then let our confidence set for α be the values of α that lie within this set. The most natural way to construct the larger confidence set would be to "invert a test statistic." That is, we would first construct a test statistic $T(\theta)$ which has a known distribution under the null hypothesis: $\theta = \theta_0$. For each potential θ, we would construct an acceptance region of the test. When $T(\theta)$ lies within this acceptance region, θ would belong to this confidence set, otherwise it would not. Given the confidence

[10] In the definition of the estimator, we have not explicitly defined Λ,Γ,β, or Σ as parameters but express the estimates of these objects as functions of the data and θ. Because the dimension of these objects grows with K^*, it is easier to focus on the elements of θ when considering consistency and inference

set for the full parameter space, we take the confidence set to be the set of α that lie within this set. More formally let $T_{N,K^*}(\theta)$ be the estimated value of the test statistic and let $T^c(\theta)$ the critical value. Assuming we reject when the test statistic is larger than the critical value, the confidence set is defined as

$$\widehat{C}_{N,K^*} = \left\{\theta \in \Theta \mid \widehat{T}(\theta) \leq T^c(\theta)\right\},$$

and our estimated confidence region for α can be written as

$$\widehat{C}_{\alpha} = \left\{\alpha \in \mathbb{R} \mid \exists\, (\phi, P_S, \sigma_\xi) : (\alpha, \phi, P_S, \sigma_\xi) \in \widehat{C}_{N,K^*}\right\}.$$

There are many test statistics one could use and many ways to calculate the critical value. We consider the following algorithm based on the bootstrap. Consider testing the null hypothesis $\theta = \theta_0$.The most natural test statistic is the normalized criteria function, so that

$$T_{N,K^*}(\theta_0) = K\left(q^*_{N,K^*}(\theta_0) - q_{N,K^*}(\theta_0)\right)' \Omega \left(q^*_{N,K^*}(\theta_0) - q_{N,K^*}(\theta_0)\right)$$

where $q^*_{N,K^*}(\theta_0)$ represents the bootstrap distribution of $q_{N,K^*}(\theta_0)$. Such a test statistic would be computed as follows:

1. Estimate parameters to be used in generating data for the bootstrap. This involves using the data generation process for X_i as well. Specifically, from the empirical distribution of (X_i, W_i),

   (a) Estimate (Λ, Λ_X), Σ, and the data generating processes for F_i and v_{ij}.

   (b) Estimate

   $$\frac{\widehat{\Gamma}(\theta)}{\sqrt{K^*}} \equiv \left[\frac{1}{P_s K}\widetilde{\lambda}'\widehat{\lambda} + \widehat{\Sigma}\right]^{-1} \frac{1}{N}\widetilde{W}'\left(\widetilde{Y} - \alpha_0\widetilde{T}\right)$$
   $$\frac{\widehat{\beta}(\theta)}{\sqrt{K^*}} \equiv \left[\frac{1}{P_s K}\widetilde{\lambda}'\widehat{\lambda} + \widehat{\Sigma}\right]^{-1} \frac{1}{N}\widetilde{W}'\widetilde{Z}$$

   (c) For the hypothesized value of P_S, estimate the distribution of $(\xi_i, \psi_i, \omega_i)$.

2. Generate N_B bootstrap samples as follows for each sample.

   (a) Draw K observable covariates from the actual set of covariates (with replacement) with appropriate $\left(\widehat{\Gamma}_j, \widehat{\beta}_j, \widehat{\lambda}_j, \widehat{\Sigma}_{jj}\right)$.

(b) Draw $(K^* - K)$ unobservable covariates from the actual set of covariates (with replacement) with appropriate $\left(\widehat{\Gamma}_j, \widehat{\beta}_j, \widehat{\lambda}_j, \widehat{\Sigma}_{jj}\right)$.

(c) Now for $i = 1, N$ generate all of the (X_i, W_i^*) using the DGP for F_i, v_{ij} and v_{xi}.

(d) Using the DGP for ψ_i and ξ_i generate Z_i and $(Y_i - \alpha_0 T_i)$ *(*Note that we do not need to generate data on Y_{ii} and T_i themselves because only $\left(\widetilde{Y}_i - \alpha_0 \widetilde{T}_i\right)$enters the moment conditions that define the test statistic.*)*

(e) Given generated bootstrap data construct $q^*_{N,K^*}(\theta_0)$ and then the test statistic $Q_{N,K^*}(\theta)$. (This involves the intermediate steps of estimating Σ, λ and Γas well.)

3. From the bootstrap sample we can estimate the distribution of the test statistic and calculate the critical value given the size of the test.

For this critical value to be correct, we need that the bootstrap distribution of $T_{N,K^*}(\theta_0)$ provides a consistent estimate of the actual distribution of $T_{N,K^*}(\theta_0)$.

It will prove useful to define

$$\chi_j = \left[\begin{array}{cccccccccc} \Lambda_j\Gamma_j & \Lambda_j\beta_j & \Gamma_j\sigma_j^2\Gamma_j & \Gamma_j\sigma_j^2\beta_j & S_j\frac{\Lambda_j^2}{\sigma_j^2} & S_j\Gamma_j\Lambda_j & S_j\Gamma_j\Lambda_j\sigma_j^2 & S_j\beta_j\Lambda_j & S_j\beta_j\Lambda_j\sigma_j^2 & S_j\Gamma_j^2\sigma_j^2 \; S_j \end{array} \right]'$$

and

$$\chi_0 = E\left(\chi_j\right).$$

Our next goal to show that the limit of $q_{N,K^*}(\theta_0)$ as N gets large is a known function of only θ and the mean of χ_j.

Theorem 4. *Under Assumptions 8-10*

$$q_{n,K^*}\left(\theta\right) \underset{N\to\infty}{\overset{p}{\to}} f\left(\theta, \frac{1}{K^*}\sum_{j=1}^{K^*}\chi_j\right)$$

where f is a known function. As long as at our true parameter θ_0,

$$\frac{\partial L' f(\theta_0, E(\chi_j))}{\partial E(\chi_j)} \neq 0,$$

the bootstrap distribution of the test statistic is consistent.

(Proof in Progress)

The proof of this theorem is not quite done-we still need to verify that we have all the necessary regularity conditions.

The computational burden of computing $T^c(\theta)$ for the desired confidence level is likely to be very large. However, the moments that determine the criterion function of the model are continuous functions of θ. Consequently, $T^c(\theta)$ should be a smooth function of θ. We propose computing a modest number of draws of $Q_{N,K^*}(\theta)$ for each of the grid points of θ chosen and then approximating $T^c(\theta)$ by fitting a quantile regression model to the draws for the various values of θ. One can increase the number of grid points, number of draws, and the flexibility of the quantile regression model as needed to ensure that the approximation is accurate for the confidence level chosen. The restrictions $0 < P_S \leq 1$ and $0 \leq \sigma_\xi^2 < var(\tilde{Y}_i)$ as well as the fact that the sign of ϕ_u is known in some applications reduces the number of points that must be entertained.

3.4.2 A Simplified Bootstrap Procedure

Given the computational complexity of the above procedure, we also propose a less demanding alternative. An additional motivation for the alternative procedure stems from the fact that one often has a strong prior about the sign of the selection bias. We can obtain tighter bounds by imposing this prior (formally defined as "monotone selection" in Manski and Pepper, 2000). While our estimation interval can potentially be much more complicated, in simulations we consistently find a compact region with one end of the region occurring at the instrumental variable estimate ($P_S = 1$) and the other occurring at the "observables like unobservables" assumption ($\sigma_\xi = 0$). Without loss of generality we will assume positive selection bias so that the upper bound occurs under the constraint $P_S = 1$. We will also assume that the minimum value occurs at σ_ξ. We propose a parametric bootstrap procedure to construct one-sided confidence interval estimators for the lower and upper bounds of this set, denoted $\alpha_{\min}$ and $\alpha_{\max}$, respectively. For concreteness, suppose one choose a confidence level of $(1-\varphi)$. We construct these intervals such that the estimator $\hat{\alpha}_{\varphi,\min}$ has the nominal probability φ of being below $\alpha_{\min}$. The estimator $\hat{\alpha}_{\varphi,\max}$ has the nominal probability φ of exceeding $\alpha_{\max}$.

3.4.3 Construction of $\hat{\alpha}_{\varphi,\min}$

The procedure for estimating $\hat{\alpha}_{\varphi,\min}$ involves the following steps.

1. Estimate the model parameters under the assumption that $\sigma_\xi = 0$ by solving the system of equations

$$0 = q_N^1(\widehat{\alpha}_{min}, \widehat{\phi}, \widehat{P}_S, 0) = q_N^2(\widehat{\alpha}_{min}, \widehat{\phi}, \widehat{P}_S, 0) = q_N^3(\widehat{\alpha}_{min}, \widehat{\phi}, \widehat{P}_S, 0)$$

for $\widehat{\alpha}, \widehat{\phi}$, and $\widehat{P}_S$. In doing this we also obtain estimates of $\Lambda, \Lambda_X, \Sigma$, $\Sigma_X$and γ for X and the observable W_j.

2. Next estimate some additional parameters that will be used for generating the bootstrap sample.

   (a) Obtain estimates of the distributions for F_i, v_{ij}, and v_{xi} given the estimates of $[\hat{\Sigma}, \hat{\Lambda}_j]$. This can be done in a number of different ways. One could specify a parametric distribution and estimate the distribution parameters. Alternatively, one could do this completely nonparametrically. A third possibility is to take advantage of the fact that our estimator involves up to second moments of the variables, so only up to 4th moments of the distributions of these variables matter for the sampling distribution of $\hat{\alpha}_{\min}$. Instead of specifying parametric distributions, one could use a method of moments procedure to estimate up to the fourth moments from sample estimates of $E(\widetilde{W}_{ij}^r \widetilde{W}_{ij'}^s)$ and $\hat{\sigma}_v, \hat{\Lambda}_j$, $j = 1..., K$ for various values of r and s. One could then pick convenient parametric distributions for F_i and v_{ij}, $j = 1, ..., K$ and choose parameters of the distributions to match the relevant moments.[11] Call the estimates of the additional parameters of the F_i distribution $\hat{B}_F$ and the additional parameters of the v_{ij} distribution $\hat{B}_{v_j}$.[12] A similar procedure can be used to estimate additional parameters $\hat{B}_{vx}$ of the distribution of the vector v_{xi}

   (b) Next we need to estimate the distribution of $(\xi_i, \psi_i, \omega_i)$. We can use the same three approaches as in the previous case. To use the third we need estimates of

[11] Sticking with the one factor case and taking W_{ij} to be mean zero, using independence of θ_i and the v_{ij}, and using the fact that $var(\theta_i) = 1$, the moments are $E(W_{ij}^4) = \Lambda_j^4 E(\theta_i^4) + E(v_{ij}^4) + 4\Lambda^2 1 \sigma_{vij}^2$ and $E(W^2 ij W_{ij'}^2) = \Lambda_j^2 \Lambda_{j'}^2 E(\theta_i^6) + \sigma_{vj}^2 \sigma_{vj'}^2$ for all $j, j' \neq j$ pairs. The idea generalizes to the multiple factor case.

[12] An alternative is to use the K observed W_j, impose the estimates $\hat{\Lambda}_j$ and the estimates of $\hat{\sigma}_{vj}$, choose parametric distributions for θ_i $v_{1i}, ..., v_{Ki}$, and fit the parameters of those distributions. The chosen distributions should not impose constraints on the second and fourth moments. In principle, one could work with nonparametric distributions with the variance constrained to match the σ_{vj}^2. A nonparametric approach is unattractive from a computational point of view, and given that our estimators only involve first and second moments , it does offer any clear advantages.

fourth moments. To obtain them, one can use the fourth moments of $\widetilde{Y}_i - \hat{\alpha}\widetilde{T}_i$, $\widetilde{Z}_i$ and $\widetilde{T}_i$. Consider

$$E(\xi_i^4) = E(\widetilde{Y}_i - \alpha\widetilde{T}_i)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\widetilde{W}_{ij}\Gamma_j)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\widetilde{W}_{ij}\Gamma_j)^2\sigma_\xi^2.$$

We have the estimate of $\hat{\alpha}_{\min}$, so $E(\widetilde{Y}_i - \alpha\widetilde{T}_i)^4$ can be replaced with the corresponding sample moment. We also have estimates of $E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\widetilde{W}_{ij}\Gamma_j)^2$ and σ_ξ^2. One can use a similar procedure to estimate $E(\psi_i^4)$. The relevant moment condition is

$$E(\psi_i^4) = E(\widetilde{Z}_i)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\widetilde{W}_{ij}\beta_j)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}\widetilde{W}_{ij}\beta_j)^2\sigma_\psi^2 \ .$$

Note that this requires an estimate of $\widehat{\beta}$ and σ_ψ^2, but estimating these is analogous to estimating $\widehat{\gamma}$ and σ_ξ^2 where the dependent variable is now $\widetilde{Z}_i$ rather than $\widetilde{Y}_i - \alpha\widetilde{T}_i$. Estimation of δ, σ_ω^2 and $E(\omega_i^4)$ is analogous. We would then pick convenient parametric distributions for this joint distribution, and estimate parameters $B_{\xi,\psi,\omega}$. The joint distribution should not constrain the second and fourth moments unless one wishes to impose additional a priori information (such as normality) on it. We leave implicit the fact that $\hat{B}_{\xi,\psi,\omega}$ depends on $\hat{\alpha}_{\min}$.

3. Construct the Bootstrap sample. This involves a few different steps.

   (a) Using the estimates $[\hat{\beta}_j, \hat{\Gamma}_j, \hat{\sigma}_v, \hat{\Lambda}_j, \hat{B}_j]$, $j = 1, ..., K$, and the estimates $\hat{P}_S$, draw $\hat{K}^*$ values of $[\hat{\beta}_j, \hat{\Gamma}_j, \hat{\sigma}_{vj}, \hat{\Lambda}_j, \hat{B}_j]$ by sampling with replacement from the K estimated values. Let the first K correspond to the "observed" $W's$ for purposes of the bootstrap replication.

   (b) Using $(\hat{\sigma}_{vj}, \hat{\Lambda}_j, \hat{B}_j)$ and $\hat{B}_F$, generate $(F_i)^{(b)}$, $(v_{ij})^{(b)}$ and then $W_{ij}^{(b)}$, $i = 1...N.$, $j = 1, ..., \hat{K}^*$ where (b) denotes the bth bootstrap replication, $(b) = 1, ..., N_{boot}$.

   (c) Using the $\hat{K}^*$ values of $\hat{\beta}_j$, the associated K^* vectors $W_{ij}^{(b)}$, $\hat{\alpha}_{\min}$, and the draws of $\psi_i^{(b)}$, use $\widehat{B}_{\xi,\psi,\omega}$ to generate N values of $(Z_i^{(b)}, T_i^{(b)}, Y_i^{(b)})$.

4. For each bootstrap sample compute $\hat{\alpha}_{\min}^{(b)}$ by solving

$$0 = q_{N^{(b)}}^1(\widehat{\alpha}_{min}^{(b)}, \widehat{\phi}, \widehat{P}_S, 0) = q_{N^{(b)}}^2(\widehat{\alpha}_{min}^{(b)}, \widehat{\phi}, \widehat{P}_S, 0) = q_{N^{(b)}}^3(\widehat{\alpha}_{min}^{(b)}, \widehat{\phi}, \widehat{P}_S, 0)$$

on the bootstrap samples.

5. Calculate the φ^{th} quantile of the bootstrap sample of $\widehat{\alpha}_{min}$ and subtract the different between that and our point estimate from our point estimate of $\widehat{\alpha}_{min}$ to obtain the lower bound of our confidence set.

3.4.4 Construction of $\hat{\alpha}_{\varphi,\max}$

To obtain $\hat{\alpha}_{\varphi,\max}$, we assume that the largest value of $\hat{\alpha}$ that satisfies the restrictions of the model is obtained when one imposes the assumption that $\hat{P}_S = 1$ and ignores the possibility that unobserved $\widetilde{W}_{ij}$ that induce positive correlation between $\widetilde{T}_i$ and $\widetilde{Y}_i$. If one sets $\hat{P}_S$ to 1 in the matrix $\left[\frac{1}{\hat{P}_S \cdot K}\widehat{\lambda}'\widehat{\lambda} + \widehat{\Sigma}\right]$ and replaces the matrix with $\widetilde{W}'\widetilde{W}$ in equation 3.14) for $\Gamma(\hat{\theta})$, then the solution for $\hat{\alpha}$ is IV. Under the null, all of the W_j are observed. Thus we do not need to impose a model of how the W_j are related to each other to account for the effects of missing W_j. One can construct the one sided confidence interval estimate using the appropriate robust standard error estimator given assumptions about serial correlation and heteroskedasticity in ξ_i. Alternatively, one can use a conventional bootstrap procedure.

While the simplicity of the above approach is attractive, it has an important shortcoming. We have not been able to prove that OLS is the upper bound when P_S is less than 1 $Cov(W, \varepsilon) \neq 0$. This is because bias in $\hat{\Gamma}$ may lead to a partially offsetting bias in $\hat{\alpha}$.

4 Monte Carlo Evidence

In this section we present Monte Carlo evidence on the performance of the lower bound estimator $\hat{\alpha}_{\min}$ for the $OU - Factor$ and $\hat{\alpha}_{\max}$, which we estimate based on $\hat{\alpha}_{OLS}$ because in our context $\hat{\alpha}_{\max}$ turns out to be essentially the same as the OLS estimator.[13] We also present evidence on the performance of the lower bound estimator for OU, which we refer to $\hat{\alpha}_{OU}$ in the tables.

We assume that there are not X variables in the model ($\Gamma_X = 0$) so the equations of the model of Y_i, T_i, and W_{ij}:

[13]The OLS estimator is essentially the same as the estimate of α based on our moment equations with P_S set to 1. The two differ because we use the moments implied by the estimated factor structure rather than the actual variance covariance matrix of W in the moment condition for $\hat{\Gamma}$. In the designs we consider we found that the maximum value of $\hat{\alpha}$ consistent with $\sigma^2_\xi > 0$ occurred at $P_S = 1$, although we have not proved that this has to be the case for any model with a factor structure.

$$
\begin{aligned}
Y_i &= \alpha_0 T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}\Gamma_j + \xi_i \\
&= \alpha_0 T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j W_{ij}\Gamma_j + \frac{1}{\sqrt{K}} \sum_{j=1}^{K^*} (1 - S_j)\, W_{ij}\Gamma_j + \xi_i \\
W_{ij} &= \frac{1}{\sqrt{K^*}} F_i \Lambda_j + v_{ij} \\
T_i &= Z_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}^K \beta_j + \psi_i
\end{aligned}
$$

We focus on the case in which F_i is a scalar $(r = 1)$. We vary assumptions about P_S, the fraction of the W_{ij} variables that are included in the model.

4.1 W parameters

The distributions of the variables that determine W_{ij} are

$$
\begin{aligned}
&F_i \sim N(0,1) \\
&v_{ij} \sim N(0, \sigma_{vj}^2);\ \sigma_{vj} \sim U(1.0, 2.0) \\
&\Lambda_j = \bar{\Lambda} + \tilde{\Lambda}_j \\
&\tilde{\Lambda}_j \sim U(-\tilde{\Lambda}_{\max}, \tilde{\Lambda}_{\max})
\end{aligned}
$$

For this specification,

$$
\begin{aligned}
E[Cov(W_j, W_{j'})|j \neq j'] &= \frac{1}{K^*} E(\Lambda_j \Lambda_{j'}) = \frac{1}{K^*} \bar{\Lambda}^2 \text{ and} \\
E[Var(W_j)] &= \frac{1}{K^*} \bar{\Lambda}^2 + \frac{1}{3K^*} [\tilde{\Lambda}_{\max}]^2 + E(\sigma_{vj}^2),
\end{aligned}
$$

where the expectations are defined over j and j'. We report $\frac{E[Cov(W_j, W_{j'})]}{E[Var(W_j)]}$ in the tables below.

4.2 Parameters of the Y_j and T_j Equations

Γ_j and β_j have expected values μ_Γ and μ_β, respectively, and depend on a common component ε_j and the components $\varepsilon_{\Gamma j}$ and $\varepsilon_{\beta j}$ that are specific to Γ_j and β_j. They are determined by

$$\begin{aligned}\Gamma_j &= \mu_\Gamma + \frac{g_\varepsilon}{[g_\varepsilon^2 + (1-g_\varepsilon)^2]^{.5}}\varepsilon_j + \frac{(1-g_\varepsilon)}{[g_\varepsilon^2 + (1-g_\varepsilon)^2]^{.5}}\varepsilon_{\Gamma j} \\ \beta_j &= \mu_\beta + \frac{b_\varepsilon}{[b_\varepsilon^2 + (1-b_\varepsilon)^2]^{.5}}\varepsilon_j + \frac{(1-b_\varepsilon)}{[b_\varepsilon^2 + (1-b_\varepsilon)^2]^{.5}}\varepsilon_{\beta j},\end{aligned}$$

where ε_j, ε_{Γ_j}, and ε_{β_j} are uniform random variables with mean 0 and variance 1. They are mutually independent and independent across j.

The parameters g_ε and b_ε determine relative weights on ε_j and the idiosyncratic terms $\varepsilon_{\Gamma_j}, \varepsilon_{\beta_j}$, thereby determining the covariance between Γ_j and β_j. The weights are normalized so that $var(\Gamma_j) = var(\beta_j) = 1$ regardless of the choice of g_ε and b_ε. g_ε^2 and b_ε^2 are the shares of the variances accounted for by the common component ε_j, respectively. For the above design,

$$\begin{aligned}E(\Gamma_j \cdot \beta_{j'}) &= \mu_\Gamma\mu_\beta + \frac{g_\varepsilon \cdot b_\varepsilon}{[g_\varepsilon^2 + (1-g_\varepsilon)^2]^{.5} \cdot [b_\varepsilon^2 + (1-b_\varepsilon)^2]^{.5}}\ , j = j' \\ &= \mu_\Gamma\mu_\beta\ , j \neq j'\end{aligned}$$

$$\begin{aligned}cov(\Gamma_j, \beta_{j'}) &= corr(\Gamma_j, \beta_{j'}) = \frac{g_\varepsilon \cdot b_\varepsilon}{[[g_\varepsilon^2 + (1-g_\varepsilon)^2]^{.5} \cdot [b_\varepsilon^2 + (1-b_\varepsilon)^2]^{.5}].}, j = j' \\ &= 0, j \neq j'.\end{aligned}$$

$$\begin{aligned}E(\Gamma_j \cdot \Gamma_{j'}) &= \mu_\Gamma\mu_\Gamma + 1, j = j' \\ &= \mu_\Gamma\mu_\Gamma, j \neq j' \\ E(\beta_j \cdot \beta_{j'}) &= \mu_\beta\mu_\beta + 1, j = j' \\ &= \mu_\beta\mu_\beta, j \neq j'\end{aligned}$$

Below we consider the effects of varying g_ε and b_ε, and we also consider a case in which $\beta_j = 0$ for all j.

4.3 Additional Parameter Values

We also examine the sensitivity of the estimates to the importance of ψ and ξ, the idiosyncractic components of T and Y, respectively. To do this, we vary σ_ξ^2 so as to vary the expected fraction of the variance of the unobservable component of Y that is due to ξ. That

is, we choose σ_ξ^2 to manipulate

$$R_\xi^2 \equiv E\left[\sigma_\xi^2/(\frac{1}{K^*}Var(\sum_{j=K_0+1}^{K^*} W_j\Gamma_j|\Gamma) + \theta\sigma_\xi^2)\right],$$

where the expectation is defined over the joint distribution of Γ, β, and W. Similarly, we set σ_ξ^2 to control

$$R_\psi^2 \equiv E\left[\sigma_\psi^2/(\frac{1}{K^*}Var(\sum_{j=1}^{K^*} W_j\beta_j|\beta) + \sigma_\psi^2)\right].$$

We report R_ψ^2 and R_ξ^2 in the tables below. Note that for a given value of R_ξ^2, the value of σ_ξ^2 will depend on the choice of P_S, but ϕ and ϕ_u will not. We view this as an attractive parameterization because we are primarily concerned with ensuring that ϕ and ϕ_u do not depend on P_S.[14] The expected values of ϕ and ϕ_u at the true α are complicated functions of the parameters of the data generation process, so we simply compute the average values in each design as well as the average estimate of $\hat{\phi}$ at $\hat{\alpha}_{\min}$. Note that the bias in OLS declines with P_S because ψ assumes an increasing important role as the source of variance in T_i that is orthogonal to the observed W_j. However, the variance of ψ also rises when the covariance among the W_j is increased and when we change μ_β.

For all experiments, we set $N = 2000$ and report results based on 1000 Monte Carlo replications. The bootstrap estimates of the .10 one-sided confidence interval estimate is based on 1000 bootstrap replications for each Monte Carlo replication. We set K^* to 100 and α_0 to 1.0 in all the experiments reported, and we set R_ψ^2 to 0.5 in all experiments except Table 1, where it is set to 1. We vary P_S, R_ξ^2, $\bar{\Lambda}$, $\tilde{\Lambda}_{\max}$, μ_B, μ_Γ, g_ε, and b_ε across experiments. Specifically, we set P_S of 0.2, 0.4, and 0.8 and we set R_ξ^2 to 0, 0.2, and 0.4. We vary μ_B, μ_Γ, g_ε, and b_ε such that $E(\beta_j\Gamma_j)$ takes on several different values. Finally, we vary $\bar{\Lambda}$ and $\tilde{\Lambda}_{\max}$. In one set of case, we set $\bar{\Lambda} = 0$, which means that $E[Corr(W_{ij}, W_{ij'})] = 0$ if $j \neq j'$. In the other set of cases, $E[Corr(W_{ij}, W_{ij'})] = 0.2$ if $j \neq j'$.

[14] If we fix $Var(\xi_i)$ at a nonzero value, the ratio ϕ_ε/ϕ approaches 0 (the case in which OLS is unbiased) as P_S approaches 1. In assessing how variation in P_S matters, we wish to hold constant the degree to which selection on observables is similar to selection on unobservables. For each Monte Carlo experiment we set σ_ψ^2 and σ_ξ^2 to the fixed values

$$\begin{aligned}
\sigma_\xi^2 &= E\left[\frac{R_\xi^2}{1-R_\xi^2}\frac{1}{K^*}Var(\sum_{j=K_0+1}^{K^*} W_j\Gamma_j|\Gamma)\right] \\
\sigma_\psi^2 &= E\left[\frac{R_\psi^2}{1-R_\psi^2}\frac{1}{K^*}Var(\sum_{j=1}^{K^*} W_j\beta_j|\beta)\right]
\end{aligned}$$

given the values of the other parameters of the experiment.

4.4 Monte Carlo Results

We first consider a baseline case in which T_i is randomly assigned. Table 1 reports results for a design in which $\beta_j = 0$ for all j ($\mu_\beta = 0$, $var(\varepsilon_{\beta j}) = 0$, and $b_\varepsilon = 0$), which means that T does not depend on the W_j. For these designs, $\hat{\alpha}_{OLS}$ is unbiased because $E(\phi) = E(\phi_u) = 0$. We report the median as our measure of central tendency, and we also report the 10th and 90th percentile values in order to show a measure of dispersion. The median values of ϕ, ϕ_u, and $\widehat{\phi}$ across replications are shown in the top three rows of the table.

The estimates of $\hat{\alpha}_{OLS}$ are tightly distributed around 1.0 in all three cases. The dispersion declines with P_S, reflecting a smaller variance of the unobserved components of Y as P_S increases. The values of $\hat{\alpha}_{OU}$ and of $\hat{\alpha}_{\min}$ are also tightly distributed around 1.0, although they are estimated less precisely than the OLS coefficients. When $P_S = 0.2$, the 90th-10th differential of $\hat{\alpha}_{\min}$ is roughly double that of the 90th-10th differential for $\hat{\alpha}_{OLS}$, but when $P_S = 0.8$, the three estimators have similar dispersion.

We turn next to designs in which OLS estimates of α_0 are biased. In Table 2a, we set $\mu_\beta = \mu_\Gamma = 0.3$, which leads to bias in $\hat{\alpha}_{OLS}$ in the specifications we consider. In the first three columns we chose b_ε and g_ε so that $E(\Gamma_j \beta_j) = 0.3$. The median of $\widehat{a}_{OLS}$ is 1.256 when $P_S = 0.2$ and 1.101 when $P_S = 0.8$. The decline in bias as P_S increases reflects the fact that the fraction of the variance in T_i that is uncorrelated with the excluded W_j rises with P_S. $\hat{\alpha}_{\min}$ is essentially unbiased in all three cases, with the dispersion declining with P_S. In the last three columns we increase b_ε and g_ε so that $E(\Gamma_j \beta_j) = 0.6$ (i.e., $Corr(\Gamma_j, \beta_j) = .51$). For each value of P_S, the bias in OLS increases relative to the cases in which $E(\Gamma_j \beta_j) = 0.3$. Interestingly, the $\widehat{\alpha}_{OU}$ and $\widehat{\alpha}_{\min}$ estimators are less noisy compared to the $E(\Gamma_j \beta_j) = 0.3$ case. When $E(\Gamma_j \beta_j) = 0.6$ and $P_S = 0.8$, as shown in column 6, $\widehat{\alpha}_{OU}$ and $\widehat{\alpha}_{\min}$ have no more sampling error than the OLS estimator.

Table 2b repeats the calculations found in Table 2a but introduces a factor structure such that $E[Corr(W_{ij}, W_{ij'})] = 0.2$ if $j \neq j'$. We impose this correlation by setting $\bar{\Lambda}$ to 3.4. In order to keep $E[Var(W_{ij})]$ constant relative to the $\bar{\Lambda} = 0$ case, we reduce $\widetilde{\Lambda}_{\max}$ from 6.2 to 2.0. The bias in OLS tends to be lower for this design, primarily because the regressors that are included do a better job of controlling for the omitted W_j when the correlation among the W_j is higher. Intuitively, as $E[Corr(W_{ij}, W_{ij'})] \to 1$, it does not matter which regressors are actually observed and which are not. The increase in the correlation across W_j is also associated with an improvement in the performance of $\hat{\alpha}_{\min}$ relative to $\hat{\alpha}_{OU}$. In particular,

$\hat{\alpha}_{OU}$ is downward biased in all of the designs apart from the one shown in the final column. This is likely due to the fact that the $\hat{\alpha}_{OU}$ estimator is based on the assumption that the restriction $\phi = \phi_u$ based on the true Γ_j carries over to the coefficient vector Γ^P of the projection of $Y_i - \alpha_i T$ on the observables W_i. However, the positive correlation between the observed and unobserved covariates results in positive omitted variables bias (on average) in the observed $\hat{\Gamma}_j$, because the unobserved covariates are positively correlated with Y. Since the observed covariates are also positively correlated with T in these designs, the positive bias on the estimates of Γ_j leads the projection of T on $W_i\Gamma^P$ to overstate the amount of selection bias, inducing a negative bias in the $\hat{\alpha}_{OU}$ estimates. This negative bias also affects the OLS estimator, partially counteracting the positive bias caused by the positive correlation of T with the unobserved elements of W. As a result, the positive bias in the OLS estimates is smaller in Table 2b than in Table 2a.

As is evident from the table, $\hat{\alpha}_{\min}$ performs very well in the presence of a factor structure. It has a median value very close to 1 and a sampling error that is similar to OLS. Presumably, the superior performance of $\hat{\alpha}_{\min}$ relative to $\hat{\alpha}_{OU}$ for the parameter values in Table 2b is due to the fact that explicitly accounting for the factor structure eliminates the positive bias on the estimates of Γ_j, which in turn eliminates the negative bias in the estimate of α_0. However, the difference in performance between $\hat{\alpha}_{\min}$ and $\hat{\alpha}_{OU}$ is only large in a few designs, such as that given by the first two columns in the table.

In Table 3, we relax the assumption that the observables are a random set of all the unobservables by setting $R^2_\xi = 0.2$. In the left panel, $\bar{\Lambda} = 0$ and $\widetilde{\Lambda}_{\max} = 6.2$, as in Table 2a. Not surprisingly, allowing for a positive variance of ξ has no effect on the median of OLS. However, the lower bound estimators $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ are now both biased downward because the assumption that $\phi = \phi_u$ no longer holds. This is easy to see in the first column, in which the median of ϕ_u across replications is 0.353, roughly 80 percent of the median of ϕ (0.438). In other words, selection on unobservables is now only 80 percent as large as selection on observables. When $E(\Gamma_j\beta_j) = 0.3$ and the factor structure is such that $E[Corr(W_{ij}, W_{ij'})] = 0$, the medians of $\hat{\alpha}_{OU}$ vary from 0.784 to 0.975 depending on P_S, and the corresponding medians of $\hat{\alpha}_{\min}$ vary from 0.878 to 0.979. However, the sampling variance of the $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ estimators is fairly wide when P_S is small. When we increase b_ε and g_ε so that $E(\Gamma_j\beta_j) = 0.6$, the positive bias in OLS increases, as was the case in Table 2a, while there is no systematic change for the other estimators. The sampling variances of $\hat{\alpha}_{OU}$ and

$\hat{\alpha}_{\min}$ are wider in this case than in the analogous cases in Table 2a (in which the assumption $\phi = \phi_u$ holds). We do not fully understand this pattern, but in spite of it, the lower bound estimators usefully complement OLS.

The right panel of Table 3 sets $\bar{\Lambda}$ and $\widetilde{\Lambda}_{\max}$ so that $E[Corr(W_{ij}, W_{ij'})] = 0.2$. The median values of $\hat{\alpha}_{\min}$ do not change very much relative to the case of independent W_j, but the sampling distribution narrows substantially. This likely reflects the fact that when the W_j are correlated, it is easier to "fill in" for the effects of missing covariates using the OU-Factor moment conditions, so that it matters less which elements of W^* are actually observed.

Table 4 is analogous to Table 3, except now $R^2_\xi = 0.4$, thereby lowering ϕ_u relative to ϕ. The median of OLS is essentially unchanged relative to the cases in which R^2_ξ is 0 or 0.2, which is not surprising. As one would expect, the medians of $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ decline in all cases, with the largest declines occurring when $P_S = 0.2$. The medians of $\widehat{\alpha}_{\min}$ range between 0.288 to 0.890 when $E[Corr(W_{ij}, W_{ij'})] = 0$. The sampling variability of the $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ estimators also increases relative to Table 3. As expected, the sampling variance of $\widehat{\alpha}_{\min}$ modestly improves when $E[Corr(W_{ij}, W_{ij'})]$ increases from 0 to 0.2.

Table 5 summarizes an experiment in which $\mu_\beta = 1$, $\mu_\Gamma = 5$ and $g_\varepsilon = b_\varepsilon = 0$. For this specification $E(\Gamma_j \beta_j) = 5$, and Γ_j and β_j are uncorrelated. In the first three columns, $E[Corr(W_{ij}, W_{ij'})] = 0$ and $R^2_\xi = 0$. OLS is badly upward-biased in these designs, with the median of $\hat{\alpha}_{OLS}$ equaling 2.109 when $P_S = 0.2$, 1.929 when $P_S = 0.4$ and 1.419 when $P_S = 0.8$. The medians of $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ range between 0.889 and 1.065, although they have a substantial sampling variance. In the middle three columns, $E[Corr(W_{ij}, W_{ij'})] = 0.2$. The bias in OLS declines but is still substantial when $P_S = 0.2$. Both α_{OU} and $\alpha_{\min}$ perform well in these designs, as they are tightly distributed around the true value of α. In the last three columns we keep $E[Corr(W_{ij}, W_{ij'})] = 0.2$ and set $R^2_\xi = 0.2$. The medians of $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ are roughly 0.26 when $P_S = 0.2$ and roughly 0.95 when $P_S = .8$, and the estimators have a relatively tight distribution. Overall, the designs in Table 5 highlight the fact that $\hat{\alpha}_{OU}$ and $\hat{\alpha}_{\min}$ can perform very well in cases in which OLS is badly biased upward, particularly when $\phi = \phi_u$ holds. When $|\phi| > |\phi_u|$, so that selection on observables is stronger than selection on unobservables, the lower bound estimators yield values below the true α_0, as expected, but the resulting bounds are often useful.

Finally, in Table 6 we explore the performance of the simplified bootstrap procedure for

six designs described above. All results in the table are based on 1000 Monte Carlo replications, each of which includes 1000 bootstrap replications. The two columns in panel A correspond to columns 2 and 4 of Table 2a, in which $\bar{\Lambda} = 0$, so that $E[Corr(W_{ij}, W_{ij'})] = 0$, and $R^2_\xi = 0$. In the first column, in which $E(\Gamma_j \beta_j) = 0.3$, the empirical size, given by $\Pr(\hat{\alpha}_{0.10,\min} < \alpha)$, is 0.087, based on a nominal size of 0.10. When $E(\Gamma_j \beta_j) = 0.6$, the empirical size equals 0.090, so that in both cases the confidence region given by $(\hat{\alpha}_{0.10,\min}, \hat{\alpha}_{0.10,\max})$ excluded α_0 in slightly less than 10 percent of cases. The table also reports the median of the estimated standard error of $\hat{\alpha}_{\min}$ across Monte Carlo replications, where the standard error in each replication is calculated across all 1000 bootstrap replications. In both cases, this estimated median standard error is slightly smaller than the standard deviation (across Monte Carlo replications) of $\alpha_{\min}$. The fact that the bootstrapped distribution of $\hat{\alpha}_{\min}$ is slightly more disperse than the analogous distribution across Monte Carlo replications is likely the cause of the underrejection described above, i.e., that the empirical sizes of the tests are slightly smaller than the nominal size.

In panel B, $E[Corr(W_{ij}, W_{ij'})] = 0.2$, and designs in the two columns correspond to columns 2 and 4 of Table 2b. Again, coverage rates are close to the nominal size of 0.10, and median standard error estimate is in the ballpark of the standard deviation across replications of $\hat{\alpha}_{\min}$.

Finally, in panel C, $R^2_\xi = 0.2$. In these cases, the estimated sampling variances of $\hat{\alpha}_{\min}$ are slightly lower than the standard deviations across replications. While one might expect that this pattern would lead to over-rejection, i.e., empirical sizes greater than 0.10, the opposite case holds: in the first column, the empirical size is 0.038, and in the second it is only 0.001. This underrejection occurs because the $\phi = \phi_u$ condition does not hold, so that the $\hat{\alpha}_{\min}$ estimator is a conservative one – the lower bound given by $\hat{\alpha}_{\min}$ will systematically lie below α_0, which is a restatement of the fact that the estimates of $\hat{\alpha}_{\min}$ in Table 3 were biased downward. As a result, the confidence region given by $(\hat{\alpha}_{0.10,\min}, \hat{\alpha}_{0.10,\max})$ will include α_0 in more than 90 percent of cases.

On the whole, the Monte Carlo results may be summarized as follows. First, the medians of $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{OU}$ are close to 1 when the assumption of equality of selection on observed and unobserved variables is correct ($R^2_\xi = 0$). There are some differences in performance depending upon the specifics of the experiment, particularly the strength of the factor structure, but overall the two perform similarly. The sampling variances are narrower when the

stronger is the factor structure, i.e., when $E[Corr(W_{ij}, W_{ij'})] = 0.2$. Second, both $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{OU}$ typically lie below the value of α_0 when $\phi > \phi_u$. This is to be expected, because both estimators are based on the assumption that $\phi = \phi_u$ and are to be interpreted as lower bound estimators if $\phi > \phi_u > 0$ (in the case $\phi > 0$). Third, the gap between the lower bound estimators and α_0 declines with P_S, which is also to be expected. Fourth, the $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{OU}$ estimators are usually less precise than is α_{OLS}. The loss of precision depends on the design and is negligible in the case in which T is randomly assigned (as in Table 1). For some designs, such as some of the cases with a strong factor structure in Table 2b, the sampling variance of $\widehat{\alpha}_{\min}$ is actually smaller than that of $\hat{\alpha}_{OLS}$. Overall, the distribution of $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{OU}$ are sufficiently precise to provide useful information about α in all of the cases that we consider.

5 Conclusion

In many situations, exclusion restrictions, functional form restrictions, or parameter restrictions are not sufficiently well grounded in theory or sufficiently powerful to provide a reliable source of identification. What can one do?

As we noted in the introduction, it is standard procedure to look for patterns in the relationship between an explanatory variable or an instrumental variable and the observed variables in the model when considering exogeneity. We provide a theoretical foundation for thinking about the degree of selection on observed variables relative to unobserved variables, and we propose two estimators that make explicit use of the pattern of selection in the observables to bound the treatment effect. We contrast the standard IV or OLS assumption that the researcher has chosen the control variables so that the instrument (or the treatment itself) are not related to the unobservables with the assumption that the control variables are randomly chosen from the full set variables that influence the outcome, and argue that the truth is likely to lie somewhere in between.

Our estimators build on Theorem 1, which concerns the coefficients of the projection of an outcome on the regression indices of the observables and the unobservables. A number of assumptions are required, but roughly speaking, the theorem says that when the number of observed and unobserved variables that influence the outcome are large, the coefficient on the index of unobservables will lie between 0 and the coefficient on the index of observables. Both OU and the $OU - Factor$ estimators identify bounds by imposing the inequality restriction

on the econometric model for the outcome. However, in the likely case that the observed and unobserved variables are related, the coefficients on the control variables will to suffer from omitted variables bias, invalidating the restriction and the case for bounds. The OU estimator combines Theorem 1 with a high level assumption about the link among the observed and unobserved variables. The $OU - Factor$ estimator adds the assumption that the observed and unobserved explanatory variables have a factor structure, which provides additional moment restrictions that permit one to account for the effects of omitted variables. We show that the estimator identifies a set that asymptotially contains the true value of the treatment parameter. We derive the asymptotic distribution of the $OU - Factor$ estimator and present a parametric bootstrap approach to statistical inference. Our Monte Carlo simulations are generally encouraging, particularly for $OU - Factor$.

There is a very long research agenda. More Monte Carlo evidence is needed in the context of real world applications and data sets. Thus far we have not applied the $OU - Factor$ estimator, and we have not performed Monte Carlo studies for designs with multiple factors. The OU estimator has the advantage of simplicity and has already been used in a number of applications. However, a way to account for randomness in which explanatory variables are included in W when constructing confidence intervals is needed. Ultimately, we believe that incorporating a formal model of the relationships among the observed and unobserved variables in W^c is the more promising long-run research path. The linear factor model that we employ in developing the $OU - Factor$ estimator is a natural way to do this, but it is also restrictive. Other models of the joint distribution of the covariates should be explored. We only touch upon the case of heterogeneous treatment effects and so far we have only considered models in which the index that determines the outcome is an additively separable function.

More generally, we think of OU and $OU - Factor$ as a start for an investigation into a broader class of estimators based on the idea that if one has some prior information about how the observed variables were arrived at, then the joint distribution of the outcome, the treatment variable, the instrument, and the observed explanatory variables are informative about the distribution of the unobservables.

In closing, we caution against the potential for misuse of the idea of using observables to draw inferences about selection bias, whether through an informal comparison of means or through the estimators we propose. The conditions required for Theorem 1 imply that it

is dangerous to infer too much about selection on the unobservables from selection on the observables if the observables are small in number and explanatory power, or if they are unlikely to be representative of the full range of factors that determine an outcome.

References

Altonji,Joseph G., 1988."The Effects of Family Background and School Characteristics on Education and Labor Market Outcomes," unpublished manuscript, Northwestern University, 1988.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, 2002. "The Effectiveness of Catholic School," unpublished manuscript, Northwestern University.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, 2005(a)."Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy,* 13(1): 151-84.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, 2005(b)."An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling", *Journal of Human Resources*, 40(4): 791-821.

Angrist J., and W. Evans, 1998. "Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" *American Economic Review*, 88, 450-477.

Angrist, Joshua D., and Alan B. Krueger, 1999."Empirical Strategies in Labor Economics," *Handbook of Labor Economics Vol. 3A,* Ashenfelter and Card (eds.), North Holland.

Bronars, Stephen G. , and Jeff Grogger, "The Economic Consequences of Unwed Motherhood: Using Twins as a Natural Experiment," *American Economic Review* 80(1994), 1141-56.

Cameron, Stephen V. and Heckman, James J., 1998."Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy,* 106, 262-333.

Cameron, Stephen V. and Christopher R. Taber, 2004."Estimating Borrowing Constraints Using the Returns to Schooling," *Journal of Political Economy*, 112(1), 132-182.

Chernozhukov, V., H. Hong, and E. Tamer, 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243-1284.

Coleman, James S., Thomas Hoffer, and Sally Kilgore, 1982. *High School Achievement: Public, Catholic, and Private Schools Compared* (New York, NY: Basic Books, Inc.).

Coleman, James S., and Thomas Hoffer, 1987. *Public and Private Schools: The Impact of Communities* (New York, NY: Basic Books, Inc., 1987).

Cuhna, Flavio, James. Heckman, and Susznne Schennach, 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, vol. 78(3), 883-931.

Currie, Janet, and Thomas Duncan, 1990."Does Head Start Make a Difference?" *American Economic Review*, 85, 341-64.

Engen, Eric, William Gale, and John Karl Sholz, 1996."The Illusory Effects of Saving Incentives on Saving," *Journal of Economic Perspectives,* 10, 113-138.

Evans, William N., and Robert M. Schwab, 1995."Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics,* 110, 947-974.

Goldberger, Arthur S., and Glen C. Cain, 1982."The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer and Kilgore Report," *Sociology of Education*, LV, 103-122.

Grogger, Jeff, and Derek Neal, 2000."Further Evidence on the Benefits of Catholic Secondary Schooling," *Brookings-Wharton Papers on Urban Affairs*, 151-193.

Heckman, James .J., 1990. "Varieties of Selection Bias," *American Economic Review,* 80.

Heckman, J., and Robb, R., 1985. "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer eds.,*Longitudinal Analysis of Labor Market Data.* Cambridge, Cambridge University Press.

Imbens, G., and Angrist, J., 1994."Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-75.

Jacobsen, Joyce P., James W. Pearce III, and Joshua L. Rosenbloom, 1999."The Effect of Childbearing on Married Women's Labor Supply and Earnings," *Journal of Human Resources* 34(3), pp. 449-474.

Manski, C., 1989."Anatomy of the Selection Problem," *Journal of Human Resources,* 24, 343-360.

Manski, C., 1994."The Selection Problem," in C. Sims (ed) *Advances in Econometrics: Sixth World Congress* (Cambridge: Cambridge University Press).

Manski, C., and J. Pepper, 2000."Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica,* 68(4), 997-1010.

McLeish, D. L., 1975. "A Maximal Inequality and Dependent Strong Laws," *The Annals of Probability*, 3, 829-839.

Murnane, Richard J., 1984."A Review Essay—Comparisons of Public and Private Schools: Lessons from the Uproar," *Journal of Human Resources* 19, 263–77.

Murphy, Kevin M., and Robert H. Topel, 1990."Efficiency Wages Reconsidered: Theory and Evidence," in Y. Weiss and R. Topel eds., *Advances in the Theory and Measurement of Unemployment.* New York, St. Martin's Press, 204-40.

Neal, Derek, 1997."The Effects of Catholic Secondary Schooling on Educational Attainment," *Journal of Labor Economics* 15, 98-123.

Poterba, James, Steven Venti, and David Wise, 1994."Targeted Retirement Saving and the Net Worth of Elderly Americans," *American Economic Review,* 84, 180-185.

Rosenbaum, Paul R., 1995. *Observational Studies*, Springer-Verlag, New York, NY.

Rosenbaum, Paul R., and Donald Rubin. 1983. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study

with Binary Outcome." Journal of the Royal Statistical Society Series B: 45(2): 212–18

Staiger, Douglas, and James Stock, 1997."Instrumental Variables Regression with Weak Instruments," *Econometrica,* 65(3), 557-586.

Stock, James, 1994."Unit Roots, Structural Breaks and Trends," *Handbook of Econometrics, Volume 4,* Engle and Mcfaddend eds., Elsevier Science, 2740-2841.

Udry, Christopher, 1996."Gender, Agricultural Production, and the Theory of the Household", *Journal of Political Economy*, 104, 1010-1046.

White, Halbert, 1984. *Asymptotic Theory for Econometricians,* Academic Press, Inc.

Table 1: Monte Carlo Results Based on Designs in which Z is Randomly Assigned (all β terms=0)
Factor structure: E(Corr(W_j,$W_{j'}$))=0

	P_S=0.2	P_S=0.4	P_S=0.8
Median of φ	0.004	-0.003	-0.002
Median of φ_ε	0.001	-0.001	0.001
Median of estimated φ (at α_{min})	0.002	-0.002	-0.002
α_{OLS}			
10th percentile	0.943	0.986	0.988
Median	1.005	1.002	1.000
90th percentile	1.049	1.017	1.014
α_{OU}			
10th percentile	0.911	0.981	0.985
Median	1.000	1.006	1.000
90th percentile	1.092	1.021	1.013
α_{min}			
10th percentile	0.909	0.981	0.984
Median	1.001	0.994	1.000
90th percentile	1.117	1.010	1.014

Notes: In all specifications, α_0=1, E(Γ)=0, all β terms=0, N=2000, and K*=100.

Table 2a: Monte Carlo Results Based on Designs in which $R^2_\xi=0$
Factor structure: $E(Corr(W_j,W_{j'}))=0$

	$E(\beta*\Gamma)=0.3$			$E(\beta*\Gamma)=0.6$		
	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$
Median of φ	0.432	0.453	0.453	0.848	0.820	0.823
Median of φ_ε	0.456	0.446	0.452	0.815	0.831	0.817
Median of estimated φ at α_{min}	0.415	0.465	0.451	0.786	0.816	0.815
α_{OLS}						
10th percentile	1.167	1.102	1.038	1.376	1.264	1.119
Median	1.256	1.180	1.101	1.477	1.351	1.181
90th percentile	1.345	1.256	1.170	1.550	1.421	1.249
α_{OU}						
10th percentile	0.307	0.749	0.920	0.621	0.797	0.947
Median	0.940	1.007	1.004	0.987	0.997	1.004
90th percentile	1.422	1.269	1.086	1.382	1.156	1.057
α_{min}						
10th percentile	0.624	0.810	0.907	0.739	0.837	0.941
Median	0.993	0.998	1.006	1.001	1.002	1.004
90th percentile	1.371	1.202	1.096	1.195	1.107	1.061

Notes: In all specifications, $\alpha_0=1$, $E(\Gamma)=E(\beta)=0.3$, N=2000, K*=100, $R^2_\psi=0.5$, and $R^2_\xi=0$.

Table 2b: Monte Carlo Results Based on Designs in which $R^2_\xi=0$
Factor structure: $E(Corr(W_j,W_{j'}))=0.2$

	$E(\beta*\Gamma)=0.3$			$E(\beta*\Gamma)=0.6$		
	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$
Median of φ	0.735	0.772	0.782	0.924	0.929	0.928
Median of φ_ε	0.767	0.727	0.752	0.920	0.932	0.914
Median of estimated φ at α_{min}	0.750	0.785	0.803	0.923	0.943	0.941
α_{OLS}						
10th percentile	1.084	1.029	0.955	1.224	1.192	1.060
Median	1.137	1.116	1.042	1.294	1.293	1.137
90th percentile	1.228	1.202	1.158	1.448	1.425	1.262
α_{OU}						
10th percentile	0.639	0.616	0.777	0.864	0.872	0.880
Median	0.795	0.866	0.938	0.966	0.979	0.991
90th percentile	0.914	1.057	1.121	1.067	1.066	1.078
α_{min}						
10th percentile	0.864	0.889	0.933	0.911	0.923	0.959
Median	0.989	0.993	1.002	0.983	0.999	1.005
90th percentile	1.105	1.088	1.049	1.068	1.058	1.041

Notes: In all specifications, $\alpha_0=1$, $E(\Gamma)=E(\beta)=0.3$, N=2000, K*=100, $R^2_\psi=0.5$, and $R^2_\xi=0$.

Table 3: Monte Carlo Results Based on Designs in which R^2_ξ=0.2

	Factor structure: $E(Corr(W_j,W_{j'}))=0$						Factor structure: $E(Corr(W_j,W_{j'}))=0.2$					
	$E(\beta*\Gamma)=0.3$			$E(\beta*\Gamma)=0.6$			$E(\beta*\Gamma)=0.3$			$E(\beta*\Gamma)=0.6$		
	P_S=0.2	P_S=0.4	P_S=0.8	P_S=0.2	P_S=0.4	P_S=0.8	P_S=0.2	P_S=0.4	P_S=0.8	P_S=0.2	P_S=0.4	P_S=0.8
Median of φ	0.438	0.444	0.454	0.850	0.816	0.832	0.909	0.899	0.806	1.097	1.077	0.956
Median of φ_ε	0.353	0.347	0.337	0.632	0.640	0.631	0.564	0.533	0.580	0.689	0.670	0.697
Median of estimated φ at α_{min}	0.412	0.495	0.474	0.670	0.718	0.791	0.678	0.712	0.802	0.755	0.823	0.915
α_{OLS}												
10th percentile	1.165	1.138	1.039	1.378	1.334	1.120	1.112	1.091	1.026	1.208	1.202	1.087
Median	1.257	1.224	1.101	1.476	1.407	1.181	1.188	1.157	1.068	1.300	1.284	1.130
90th percentile	1.345	1.314	1.171	1.551	1.479	1.248	1.280	1.231	1.134	1.397	1.379	1.217
α_{OU}												
10th percentile	0.213	0.258	0.884	0.033	0.439	0.889	0.333	0.505	0.725	0.379	0.556	0.852
Median	0.868	0.784	0.975	0.783	0.806	0.953	0.649	0.752	0.922	0.713	0.812	0.948
90th percentile	1.388	1.276	1.067	1.346	1.121	1.013	0.834	0.877	0.970	0.863	0.901	0.982
α_{min}												
10th percentile	0.227	0.576	0.865	0.275	0.493	0.882	0.447	0.690	0.904	0.380	0.641	0.907
Median	0.879	0.878	0.979	0.738	0.812	0.952	0.788	0.871	0.969	0.732	0.840	0.963
90th percentile	1.516	1.215	1.077	1.063	0.950	1.017	1.026	1.202	1.014	0.903	0.922	0.994

Notes: In all specifications, α_0=1, $E(\Gamma)=E(\beta)$=0.3, N=2000, K*=100, R^2_ψ=0.5, and R^2_ξ=0.2.

Table 4: Monte Carlo Results Based on Designs in which R^2_ξ=0.4

	Factor structure: E(Corr(W_j,$W_{j'}$))=0						Factor structure: E(Corr(W_j,$W_{j'}$))=0.2					
	E(β*Γ)=0.3			E(β*Γ)=0.6			E(β*Γ)=0.3			E(β*Γ)=0.6		
	P_S=0.2	P_S=0.4	P_S=0.8	P_S=0.2	P_S=0.4	P_S=0.8	P_S=0.2	P_S=0.4	P_S=0.8	P_S=0.2	P_S=0.4	P_S=0.8
Median of φ	0.440	0.442	0.455	0.841	0.812	0.827	1.076	0.988	0.831	1.311	1.204	0.993
Median of φ_ε	0.259	0.252	0.249	0.461	0.465	0.467	0.385	0.395	0.412	0.489	0.466	0.493
Median of estimated φ at α_{min}	0.402	0.487	0.504	0.512	0.585	0.744	0.537	0.622	0.785	0.565	0.673	0.865
α_{OLS}												
10th percentile	1.188	1.149	1.028	1.364	1.313	1.125	1.130	1.117	1.023	1.219	1.204	1.084
Median	1.270	1.228	1.093	1.466	1.392	1.165	1.198	1.164	1.084	1.284	1.281	1.140
90th percentile	1.355	1.328	1.166	1.565	1.507	1.250	1.270	1.243	1.134	1.390	1.372	1.186
α_{OU}												
10th percentile	-0.449	-0.280	0.750	-0.656	-0.185	0.712	-0.317	0.016	0.789	-0.353	0.034	0.786
Median	0.705	0.551	0.887	0.446	0.437	0.825	0.294	0.513	0.876	0.267	0.537	0.891
90th percentile	1.418	1.232	1.028	1.486	0.977	0.911	0.595	0.720	0.927	0.610	0.736	0.943
α_{min}												
10th percentile	-0.505	-0.010	0.762	-0.628	-0.225	0.703	-0.299	0.175	0.817	-0.555	0.041	0.796
Median	0.657	0.655	0.890	0.288	0.418	0.832	0.377	0.590	0.914	0.297	0.553	0.901
90th percentile	1.702	1.582	1.021	1.080	0.699	0.920	0.932	0.789	0.966	0.662	0.739	0.954

Notes: In all specifications, α_0=1, E(Γ)=E(β)=0.3, N=2000, K*=100, R^2_ψ=0.5, and R^2_ξ=0.4.

Table 5: Monte Carlo Results Based on Designs in which E(β)=1 and E(Γ)=5

	Factor structure: $E(Corr(W_j,W_{j'}))=0$			Factor structure: $E(Corr(W_j,W_{j'}))=0.2$			Factor structure: $E(Corr(W_j,W_{j'}))=0.2$		
	$R^2_\xi=0.0$			$R^2_\xi=0.0$			$R^2_\xi=0.2$		
	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$	$P_S=0.2$	$P_S=0.4$	$P_S=0.8$
Median of φ	0.197	0.192	0.191	0.201	0.202	0.196	0.548	0.365	0.216
Median of φ_ε	0.190	0.192	0.202	0.200	0.191	0.199	0.097	0.079	0.105
Median of estimated φ at α_{min}	0.198	0.190	0.192	0.201	0.197	0.197	0.175	0.185	0.193
α_{OLS}									
10th percentile	2.009	1.823	1.266	1.402	1.164	1.027	1.350	1.151	1.024
Median	2.109	1.929	1.419	1.480	1.223	1.047	1.479	1.212	1.054
90th percentile	2.221	2.025	1.541	1.593	1.290	1.073	1.575	1.300	1.076
α_{OU}									
10th percentile	-0.027	0.502	0.826	0.915	0.944	0.980	0.060	0.469	0.923
Median	0.924	1.062	1.019	0.990	0.984	1.000	0.267	0.568	0.951
90th percentile	1.613	1.428	1.200	1.057	1.043	1.018	0.424	0.655	0.974
α_{min}									
10th percentile	0.014	0.476	0.764	0.854	0.919	0.945	0.007	0.437	0.885
Median	0.889	1.065	1.038	0.990	1.003	1.001	0.255	0.544	0.947
90th percentile	1.597	1.469	1.199	1.106	1.062	1.050	0.431	0.665	0.993

Notes: In all specifications, $\alpha_0=1$, $E(\Gamma)=1$, $E(\beta)=5$, $N=2000$, $K^*=100$, and $R^2_\psi=0.5$.

Table 6: Monte Carlo Evidence on the Performance of the Simplified Bootstrap Procedure

Panel A: $E(Corr(W_j, W_{j'}))=0.0$ and $R^2_{\xi}=0.0$

	$E(\beta^*\Gamma)=0.3$	$E(\beta^*\Gamma)=0.6$
$Pr(\alpha_{min,0.10}<\alpha)$ (Empirical Coverage Rate)	0.087	0.090
Median(Standard error(α_{min}))	0.378	0.169
Standard deviation(α_{min})	0.224	0.134
Median($\alpha_{min,0.10}$)	0.750	0.803
10th percentile(α_{min})	0.810	0.837

Panel B: $E(Corr(W_j, W_{j'}))=0.2$ and $R^2_{\xi}=0.0$

	$E(\beta^*\Gamma)=0.3$	$E(\beta^*\Gamma)=0.6$
$Pr(\alpha_{min,0.10}<\alpha)$ (Empirical Coverage Rate)	0.103	0.083
Median(Standard error(α_{min}))	0.178	0.068
Standard deviation(α_{min})	0.220	0.059
Median($\alpha_{min,0.10}$)	0.875	0.907
10th percentile(α_{min})	0.889	0.923

Panel C: $E(Corr(W_j, W_{j'}))=0.2$ and $R^2_{\xi}=0.2$

	$E(\beta^*\Gamma)=0.3$	$E(\beta^*\Gamma)=0.6$
$Pr(\alpha_{min,0.10}<\alpha)$ (Empirical Coverage Rate)	0.038	0.001
Median(Standard error(α_{min}))	0.267	0.077
Standard deviation(α_{min})	0.338	0.111
Median($\alpha_{min,0.10}$)	0.724	0.725
10th percentile(α_{min})	0.690	0.641

Notes:

1) In all designs, $\alpha_0=1$, $E(\Gamma)=E(\beta)=0.3$, $N=2000$, $K^*=100$, $R^2_{\psi}=0.5$, and $P_S=0.4$.

2) Estimates in the rows labeled $Pr(\alpha_{min,0.10}<\alpha)$ are the empirical sizes of the rejection regions described in the text, based on a nominal size of 0.10. These quantities (and all others in the table) are based on 1000 Monte Carlo replications, each of which includes 1000 bootstrap replications.