Methods for Using Selection on Observed Variables to Address Selection on Unobserved Variables [1]
(Very Rough and not for Circulation)

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

February 5, 2010

[1]We have received helpful comments from seminar participants at Northwestern, University of Chicago, University of Pennsylvania, University of Wisconsin at Madison and Yale, the National Institute of Child Health and Development grant R01 HD36480-03 (Altonji and Taber), and the Economic Growth Center, Yale University.

Abstract

We develop new estimation methods for the causal effect of a variable 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. Our approach involves the use of factor model as a way to infer properties of unobserved covariates from the observed covariates. We propose a confidence interval estimator that covers the true value of the causal effect.

Notes: This is the same as the 100609v2 version, except the intro and conclusion have been edited, and the description of the monte carlo design that had been omitted by mistake has been added in. That description and the results need to be re-written to reflect new designs.

1 Introduction

Distinguishing between correlation and causality is the most difficult challenge faced by empirical researchers in the social sciences. Social scientists rarely are in a position to run a well controlled experiment. Consequently, they rely on a priori restrictions on the patterns of interaction among 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 a priori 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 estimation of bounds under assumptions that are weaker than the conventional ones.

In this paper we develop estimation strategies that may be helpful when strong prior information is unavailable regarding the exogeneity of either the variable of interest or instruments for that variable. This is the situation in many applications in economics and the other social sciences, including studies that range from the effectiveness of private schools, the effect of education on crime, the effect of crime on labor market outcomes, and the effects of obesity or excercise on health outcomes.

Our approach uses the degree of selection on observables 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

[1] See for example, Currie and Duncan (1995), Engen et al (1996), Poterba et al (1994), Angrist and Evans (1988), 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).

theoretical analysis confirming the intuition that such evidence can be informative in some situations. We provide ways to assess the degree of selection bias or omitted variables bias and in some situations provide ways to estimate bounds.

To fix ideas, let the outcome Y be a function of the latent variable Y^* which is determined as

$$\begin{aligned} Y^* &= \alpha T + W^{c\prime}\Gamma^c \\ &= \alpha T + W'\Gamma + \varepsilon, \end{aligned} \tag{1.1}$$

where T is an indicator an endogenous choice.[2] For example, in Altonji, Elder and Taber (2005, 2006, hereafter AET), Y is whether a student graduates from high school and T is whether the student attends a Catholic high school. The parameter α is the causal effect of T on Y^*, W^c is the vector of characteristics (observed and unobserved) that determine Y, and Γ^c is the causal effect of W on Y^*. In the second part of the equation W is the vector of observed variables, Γ is the corresponding subvector of Γ^c, and the error component ε is an index of the unobserved variables. In some applications a candidate instrument Z may be available, but for expositional purposes, we start with the case in which Z is T. Consider the linear projection of T onto $W'\Gamma$ and ε :

$$\text{Proj}(T|W'\Gamma, \varepsilon) = \phi_0 + \phi W'\Gamma + \phi_\varepsilon \varepsilon. \tag{1.2}$$

We formalize the idea that "selection on the unobservables is the same as selection on the observables" as

Condition 1.

$$\phi_\varepsilon = \phi.$$

One may contrast Condition 1 with the OLS condition

Condition 2.

$$\phi_\varepsilon = 0.$$

[2] We will focus on two special cases in this paper. The first is a continuous dependent variable in which $Y = Y^*$. The second is a binary variable in which $Y = 1(Y^* > 0)$.

Roughly speaking, Condition 1 says that the part of Y that is related to the observables and the part related to the unobservables have the **same** relationship with T. Condition 2 says that the part of Y related to the unobservables has **no** relationship with T.

As we discuss below, Condition 1 requires two types of assumptions. The first is that the elements of X are chosen at random from W^c. The second is that the number of elements in W and W^c are large, so that none of the elements dominates the distribution of T or the latent variable Y^*. These two assumptions are enough to establish asymptotic equality of the coefficients of the projection of T onto $W'\Gamma$ and ε. We provide a precise set of assumptions that are sufficient for Condition 1 as well as a formal proof that it holds.[3]

While the assumptions that lead to Condition 1 are strong and unlikely to hold exactly, they are no less objectionable than the OLS assumptions leading to Condition 2: $Cov(T, \varepsilon) = 0$ and $Cov(W, \varepsilon) = 0$.[4] As we discuss in more detail in Section 3, because there are only a limited number of factors that we expect to matter for a particular outcome, know how to collect, and can afford to collect, many relevant variables are left out. Furthermore, in many applications, the endogenous variable is correlated with many of the elements of W. Given the constraints that shape the choice of W and the fact that many of the elements of W are systematically related to T, it is unlikely that the many unobserved variables that determine ε are unrelated to T. This is basically what $Cov(T, \varepsilon) = 0$ requires. Furthermore, given that the W variables are typically intercorrelated, the assumption that $Cov(W, \varepsilon) = 0$ is likely to be a poor approximation to reality even though it is made in virtually all empirical studies in the social sciences.

We argue that Conditions 1 and 2 represent extreme assumptions about the degree of selection on unobservables and the truth is probably somewhere in between, with

Condition 3.

$$0 \leq \phi_\varepsilon \leq \phi \ \textit{if} \ \ \phi \geq 0$$
$$0 \geq \phi_\varepsilon \geq \phi \ \textit{if} \ \phi < 0$$

Unfortunately, Condition 1 and Condition 3 are not very helpful without additional assumptions about the relationship between W and the unobservable elements that determine ε. We propose two alternative estimators that differ in how they address this problem. We

[3] We take asymptotic approximations as the number of elements in W grows large.

[4] Technically these two assumptions are sufficient for Condition 2 but not necessary. However, the cases in which Condition 2 holds without them is involve an inprobable cancellation of biases.

refer to the first estimator as ES, for equality of selection on observables and unobservables. Roughly speaking, ES amounts to estimating the system consisting of (1.1) and (1.2) with the restriction $\phi_\varepsilon = \phi$ imposed to provide a lower (upper) bound on α if ϕ is greater (less) than 0. It requires a high level assumption that implies, roughly speaking, that the regression of T on $Y - \alpha T$ is equal to the regression of the part of T that is orthogonal to W on the corresponding part of $Y - \alpha T$. The high level assumption is required because the estimator does not make direct use of how the observed explanatory variables are interrelated to assess the consequences of omitted variables that affect both the treatment and the outcome. Essentially, it treats W as exogenous, in common with the vast IV literature that focusses on endogeneity of T 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^c are observed influences the distribution of the estimator. This estimator has been applied in Altonji, Elder and Taber (2005, 2006, and 2007, hereafter AET to study the effective of Catholic schools and the effectiveness of a medical procedure as well a number of other studies. We complete the theoretical analysis of the estimator that is presented in preliminary form in AET's unpublished 2002 paper.

The second estimator takes a more difficult but we believe more satisfactory approach, which is to explictly model the relationship among the elements of W^c. The general idea is that the observed elements W^c provide potentially useful information about the joint distribution of the elements of W^c and their relationships with Y, Z, and T. In particular, we develop a method of moments procedure that uses the bounds on selection embodied in Condition 3 but in addition assumes that W^c has a factor structure. For this reason, we refer to the estimator as ES-Factor. The estimator identifies an admissible set for α based on a model consisting of (1), a linear equation relating Z (or T) to W^c, the factor model of $W^c$and Condition 3. ES-Factor is not easy to describe in words, and so we delay a detailed description until after we present the model. However, a rough outline is follows. First, we use the covariance structure of elements of W^c to identify the factor loadings and idiosyncratic variance of the W_{ij}. Second, we use a moment condition relating Γ to $Cov(W, Y - \alpha T)$ conditional on an assumed value of α. The factor structure and the assumption that W is randomly selected from a larger subset W^* of W^c that is potentially observable permits us to estimate the missing moments that are need to obtain a consistent estimator of Γ. We define P_S to be the fraction of the elements of W^* that are included in W. Conditional on α,

three moments determine P_S and the projection coefficient ϕ, and, implicitly, the projection component ϕ_ε defined above. The model implies $0 < P_S \leq 1$ We construct a confidence interval for α from the values for which we cannot reject $0 < \phi_\varepsilon < \phi$ and $0 < P_S \leq 1$. We show that our estimator consistently identifies a set that contains the α. We also provide a parametric bootstrap procedure that may be used to construct a confidence interval for the set.

The paper continues in Section 2, where we provide the details of model of Y, T, and Z and the model of what is observed and what is unobserved. We provide an explicit set of assumptions under which Condition 1 holds, and elaborate on why Condition 3 is more likely. In Section 3 we 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 rather than on point identification of α. We then turn to estimation. In Section 4 we present the ES estimator. In Section 5 consider the ES-Factor estimator. We specify a factor structure for W^c

In Section 6 we provide some monte carlo evidence on the performance of ES and ES-Factor. We offer conclusions and a research agenda in Section 7.

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

As mentioned above many papers use the relationship between an endogenous variable or an instrumental variable and the observables to make inferences about the relationship between these variables and the unobservables. In this section we develop a theoretical foundation for this practice and provide a way to quantitatively assess the importance of the bias from the unobservables. In particular we show that modeling how the set of observed variables is determined can yield conditions that are useful for identification or the construction of bounds of treatment effects.

2.1 The Model

Our outcome variable is the variable Y. Our results in this section will generalize to more general latent variable models in which Y is a latent variable and the outcome is some function of the latent variable. However, to simplify the exposition we focus on the linear

outcome example here and in our estimation section below. As noted above, we define W^c to be the full set of variables that determine Y according to

$$Y = \alpha T + W^{c\prime}\Gamma^c, \tag{2.1}$$

where Γ^c is a conformable coefficient vector. We assume that Γ is random, but is drawn once and is identical for everyone in the population. However, W^c and T are random variables that vary across members of the population, so that each individual obtains an independent draw of W^c and T but common values of Γ and α.

Assume that some of the elements of W^c are observable to the econometrician and others are not (or that the econometrician does not know that some of the observed variables belong in the model for Y). Following the notation above, denote the observable portion of W^c as W and the corresponding elements of Γ^c as Γ so that

$$Y = \alpha T + W'\Gamma + \varepsilon, \tag{2.2}$$

where ε is unobserved. That is, for each potential covariate, W_j, let S_j be a dummy variable indicating whether W_j is observable. Then

$$W'\Gamma = \sum_{j=1}^{K^c} S_j W_j \Gamma_j, \quad \varepsilon = \sum_{j=1}^{K^c} (1 - S_j) W_j \Gamma_j. \tag{2.3}$$

Like Γ_j, S_j does not vary across the population.

As in the introduction we define the projection of some variable Z onto $W'\Gamma$ and ε to be

$$\text{Proj}(Z|W'\Gamma, \varepsilon) = \phi_0 + \phi W'\Gamma + \phi_\varepsilon \varepsilon. \tag{2.4}$$

Building on AET, we explore the use of the inequality restriction

$$\begin{aligned} 0 &\leq \phi_\varepsilon \leq \phi \text{ if } \phi > 0 \\ 0 &\geq \phi_\varepsilon \geq \phi \text{ if } \phi < 0 \end{aligned}$$

in the estimation of α. Unless stated otherwise, we will assume $\phi > 0$.

At this point the variable Z could be anything. In the estimation section below we will assume that Z is a potential instrumental variable for T, with the special case $Z = T$. Another possibility is that T could represent some nonlinear function of Z. For example if T is binary, an interesting case occurs when Z is latent and $T = 1(Z > 0)$.

2.2 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 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. 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. 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 of the limits on the number of the factors that we know matter and that we know how to collect and can afford to collect, many elements of W^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 endogenous variable is correlated with many of the elements of W.

These considerations suggest that the 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 and Y when choosing S. (It would also be to useful to collect potential instrumental variables that determine T but not Y.) However, many factors that influence Y and are correlated with T and/or W are left out.

The other extreme is that the constraints on data collection are sufficiently severe that it is better to think of the elements of W as a more or less random subset of the elements of W^c rather than a set that has been 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. In our notation this amounts to assuming that S_j is an *iid* binary random variable which is equal

to one with probability P_S. The outcome of S_j determines whether covariate W_j is observed. Of course, there are other ways to capture the idea of equality of selection on observables and unobservables. For example, P_S may vary across types of variables but have no systematic relationship with the values of Γ_j relative to the influence of the variables on T.

To the extent that the data set was designed for the study of the effect of T on Y, one might expect $\phi > \phi_\varepsilon$ in equation (2.4). Furthermore, in many problems Y is a future outcome and will depend on unobserved factors that are determined after Z and T are determined Consider the case of the effect of Catholic schools on 12th grade test scores and high school graduation studied by AER. In that case, ε will reflect variability in test performance on a particular day, which presumably has nothing to do with the decision to start Catholic high school. Furthermore, high school outcomes will be influenced by shocks that occur after eighth grade, which are excluded from W. These will influence high school outcomes but not the probability of starting a Catholic high school.

With these considerations in mind, we partition W^c into two categories of variables. The first, W^*, consists of K^* variables that affect Y and potentially Z and T and have a probability of being observed and used by the econometrician. The subvector W is observed and the subvector W^u is not. The second category consists of the $K^c - K^*$ vector W^{**}. These variables have a 0 probability of being observed and used. Without lost of generality, we will index the variables so that $j = 1, ..., K^*$ corresponds to W^* and $j = K^* + 1, ..., K^c$ corresponds to W^{**}. In this case,

$$\varepsilon = \sum_{j=1}^{K^c} (1 - S_j) W_j \Gamma_j \; + \sum_{j=K^*+1}^{K^c} W_j \Gamma_j \; = W^{u\prime} \Gamma^u + \xi$$

where Γ^u is the subvector of Γ^c that corresponds to W^u and $\xi = W^{**\prime} \Gamma^{**}$ For this reason, we use the inequality

$$0 < \phi_\varepsilon < \phi \tag{2.5}$$

as the basis for the estimation strategy developed below, which focusses on estimation of an admissible set for α that contains the true value rather on point estimation. (have to say something about prior informatoin about sign of selection)

2.3 Implications of Random Selection of Observables

We are now ready to consider the implications of random selection from W^c. We begin with the general case. We first allow the number of covariates W to get large and derive the probability limit of ϕ_ε/ϕ. We then consider three special cases.

For individual i, we define Y_i and Z_i as outcomes for a sequence of models indexed by K^* where there are K^* covariates that determine Y_i.[5] 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^*.

Define $\mathcal{G}^{K^*}$ as the information set consisting of the realizations of the S_j, the Γ_j, and the joint distribution of W_{ij} for $j = 1, ..., K^*$.

Assumption 1.

$$Y_i = \alpha T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}\Gamma_j + \xi_i \tag{2.6}$$

where (W_{ij}, Γ_j) *is unnconditionally stationary (indexed by* j*).*

We use slightly 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 in the case of some social science outcomes, a large number of components determine outcomes, with none dominating. 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 will differ across j. .

Let $\sigma_{j,\ell}^{K^*} = E\left(W_{ij}W_{i\ell} \mid \mathcal{G}^{K^*}\right)$. To guarantee that the $var(Y_i)$ does not blow up as K^* gets large, we assume that

Assumption 2.

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

and

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

[5] The "local to unity" literature in time series econometrics" (discussed in Stock, 1994) and the "weak instruments" literatures (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.

The next two assumptions guarantee that $cov(Z_i, Y_i)$ is well behaved as K^* grows.

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

$$E\left(Z_i W_{ij} | \mathcal{G}^{K^*}\right) = \frac{\mu_j^{K^*}}{\sqrt{K^*}}$$

then

$$E(\mu_j^{K^*} \Gamma_j) < \infty.$$

and

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

Assumption 4. *Mean zero W*

$$E\left(W_{ij} \mid \mathcal{G}^{K^*}\right) = 0.$$

Assumption 4 is inoccuous provided that the models for Y, T, and Z contain intercepts, which we surpress for notational convenience.

Finally we need some assumptions about the process under which observables are chosen. We have discussed the case above in which variables are chosen at random.

Assumption 5. *For $j = 1, ..., K^*$, S_j is independent and identicially distributed with* $0 < \Pr(S_j = 1) = P_s \leq 1$.

Assumption 6. *ξ is independent of Z and W^*.*

First we consider the general case and then derive some 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 W_{ij} is different from the random variables Γ_j^K 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 draws his own W_j.

Theorem 1. *Define ϕ and ϕ_ε such that*

$$\begin{aligned} &Proj_S\left(Z_i \mid \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j W_{ij} \Gamma_j, \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) W_{ij} \Gamma_j + \xi; \mathcal{G}^K\right) \\ &= \phi\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} S_j W_{ij} \Gamma_j\right) + \phi_\varepsilon\left(\frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} (1 - S_j) W_{ij} \Gamma_j + \xi_i\right). \end{aligned}$$

Then under assumptions 1-6 as K^* gets large, when the probability limits of ϕ is nonzero:

$$\frac{\phi_\varepsilon}{\phi} \xrightarrow{p} \frac{(1-P_s)\,A}{(1-P_s)\,A+\sigma_\xi^2}$$

where

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

When the probability limit of ϕ is nonzero then the probability limit of ϕ_ε is also zero

(Proof in Appendix)

From this 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 $\sigma_\xi^2 = 0$,*

$$plim(\phi - \phi_\varepsilon) = 0.$$

The case in which $var(\xi) = 0$, the case in which $W^c = W^*$, meaning that W is a random subset of all of elements of W^c. Corollary 2 states that the coefficients of the projection of Z_i onto $\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} S_j W_{ij}\Gamma_j$ and $\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*}(1-S_j)\,W_{ij}\Gamma_j$ approach each other with probability one as K^* gets large. This projection is meant to be the population projection (i.e., for a very large number of persons) but with K^* fixed.

The other extreme one may consider is one in which all the important control variables are included in the model, so the variation in the error term arises from ξ only. In this case

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

$$plim(\phi_\varepsilon) = 0.$$

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

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

either

$$0 < \ plim(\phi_\varepsilon) < plim(\phi),$$

$$plim(\phi) < plim(\phi_\varepsilon) < 0,$$

or

$$0 = plim(\phi_\varepsilon) = plim(\phi).$$

This final condition will provide the basis for the estimator below.

2.4 A Model for Z

Although Theorems 1 and 2 and Lemma 1 do not require assumptions about Z beyond those given above, going forward we specify that Z takes a form similar to that for Y_i. The model is

$$Z_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}\beta_j + \psi_i. \tag{2.7}$$

We assume

Assumption 7. *(i) ψ_i is independent of all of the elements of W^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 Z as

$$Z_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K} W_{ij}\beta_j + u_i \tag{2.8}$$

where $u_i = \frac{1}{\sqrt{K^*}} \sum_{j=K+1}^{K^*} W_{ij}\beta_j + \psi_i$.

2.5 Examples of Models of that Satisfy the Assumptions of Theorem 1.

Since Assumptions 2 and 3 are quite abstract, we provide examples of models that satisfy them. The key example is the factor model of the W_{ij}, which is central to our estimation strategy. To avoid repetition, we defer presentation of the factor model until Section 5. In the other example the W_{ij} are linked across j through an MA model.[6] The MA example is the most straight forward when one examines Assumptions 1 and 2 as we refer to observables as though they have a sequential ordering.

[6] Generalizing to a more general ARMA structure is conceptually straight forward, although the algebra becomes substantially more complicated.

The MA process for W_{ij} can be defined in several ways. What is important is that W_{ij} is unconditionally stationary. Conditional on $\mathcal{G}^{K^*}$ the distribution is not restricted. However, to simplify the exposition in this section we will assume that W_{ij} is conditionally stationary as well. Thus W_{ij} will have the same marginal distribution for all j. This is not a realistic assumption for the types of data sets typically used by economists, so we wish to stress that although we use it in developing the example, it is not required for theorem. The upshot is that will assume that across individual i, W_{ij} is generated by independent and identically distributed stationary $MA(q_w)$ processes

$$W_{ij} = \zeta_{ij} + \sum_{\ell=1}^{q_w} \mu_\ell \zeta_{ij-1}$$

where ζ_{ij} is i.i.d. with finite variance σ_ζ^2. The W_{ij} processes are also independent and of the Γ_j process, and we assume further that Γ_j is generated from a stationary process with finite fourth moments. We think of j as being ordered so that variables that measure related factors appear close to each in the j sequence.[7] Given our assumptions about the W_{ij} processes and Γ_j, it is almost immediate that Assumption 1 is satisfied. In the appendix we show that the model satisfies assumption 2. In the Appendix, we also prove that Assumption 3 is satified if Z_i is determined by (2.7) and assumption (7) holds.

3 The Implications of Condition 1 for Identification of α

In this section we show that Condition 1 provides identifying information but typically is not sufficient for point identification of α. To demonstrate, we assume the Y is determined by (2.2) but consider the special case in which $cov(\varepsilon, W) = 0$. Note that this assumption is satisfied by the factor model presented below in the special case in which the variance of the factors θ_i is 0. T can be binary or continuous. It is potentially endogenous and thus correlated with ε. The instrument Z may be correlated with ε.

Define π and define $\widetilde{\beta}$ so that

[7] For example, consider a study of educational attainment in which measures of student behavior (eg., absenteeism, suspensions, getting into fights, acting out in class, unprepared for class) are viewed as potentially important control variables. If these variables appear in sequence, the above model captures the fact that they are dependence and will have Γ_j coefficients that are related. Only a subset of the behavioral variables might actually be observed.

$$\text{Proj}\left(Z \mid W\right) = W'\widetilde{\beta}, \tag{3.1}$$

$$\text{Proj}\left(T \mid W, Z\right) = W'\pi + \lambda Z, \tag{3.2}$$

and redefine v and define u to be the residual components of Z and T, so that

$$v \equiv Z - W'\widetilde{\beta} \tag{3.3}$$

$$u \equiv T - W'\pi - \lambda Z. \tag{3.4}$$

Theorem 2. *Suppose that ε is independent of W. Under condition 1, the true value of α is a root of the cubic*

$$\begin{aligned}
0 = (\alpha - \alpha^*)^3 & \left[\frac{cov\left(v,(u+\lambda v)\right)}{var(\varepsilon)}\frac{var(W'\pi+\lambda W'\beta)}{var(W'\Gamma)} - \frac{cov\left(W'\beta,(W'\pi+\lambda W'\beta)\right)}{var(W'\Gamma)}\frac{var(u+\lambda v)}{var(\varepsilon)}\right] \\
+ (\alpha - \alpha^*)^2 & \left[\frac{cov(W'\beta,W'\Gamma)}{var(W'\Gamma)}\frac{var(W'\pi+\lambda W'\beta)}{var(W'\Gamma)} + 2\frac{cov\left(v,(u+\lambda v)\right)}{var(\varepsilon)}\frac{cov(W'\Gamma,(W'\pi+\lambda W'\beta))}{var(W'\Gamma)}\right. \\
& \left. - \frac{cov(W'\beta,W'\Gamma)}{var(W'\Gamma)}\frac{var(u+\lambda v)}{var(\varepsilon)} - 2\frac{cov\left(W'\beta,(W'\pi+\lambda W'\beta)\right)}{var(W'\Gamma)}\frac{cov(\varepsilon,(u+\lambda v))}{var(\varepsilon)}\right] \\
+ (\alpha - \alpha^*) & \left[\frac{cov\left(v,(u+\lambda v)\right)}{var(\varepsilon)} + 2\frac{cov(W'\beta,W'\Gamma)}{var(W'\Gamma)}\frac{cov(W'\Gamma,(W'\pi+\lambda W'\beta))}{var(W'\Gamma)}\right. \\
& \left. - \frac{cov\left(W'\beta,(W'\pi+\lambda W'\beta)\right)}{var(W'\Gamma)} - 2\frac{cov(W'\beta,W'\Gamma)}{var(W'\Gamma)}\frac{cov(\varepsilon,(u+\lambda v))}{var(\varepsilon)}\right].
\end{aligned}$$

Thus the identified set contains one, two or three values.

The theorem says that even even if $Cov(\varepsilon, W'\Gamma) = 0$, there are typically either three solutions (i.e. three values of α^* that we can not distinguish between) or there is a unique solution that equals α.

Theorem 3. *If we impose the same model as above but use T as the instrument we get a quadratic with two roots*

$$\begin{aligned}
\alpha^* &= \alpha \\
\alpha^* &= \alpha + \frac{var(\varepsilon)}{cov(u,\varepsilon)}
\end{aligned}$$

If the research knows the sign of $cov(u, \varepsilon)$, then α is identified.

Although there are two roots, this result is useful. When an applied researcher is worried about the bias in an IV estimator, including the case when $Z = T$, he or she often has a strong

prior about the sign of the bias, which is the sign of $cov(u, \varepsilon)$. Imposing an assumption about the sign of $cov(u, \varepsilon)$ on the data delivers point identification; if one imposes that $cov(u, \varepsilon)$ is positive (negative), then the smaller (larger) of the two solutions is the true value. One should make too much of this result, because in most applications variables represented by W^{**} will be present, $var(\xi)$ will be positive, and equality of selection will not hold. Consequently, we focus on the construction on construction of bounds based on Theorem 4 below rather than on point estimation.

4 Estimators of α

We now discuss ways to estimate α. We set the stage by reviewing the approach proposed in AET (2002, 2005). Then we present our approach. In section 4.2 we present the factor model of W_{ij} that underlies our approach. In Section ___ we present our estimator.

4.1 AET's Estimator

AET's estimation strategy is to simple: use (2.5) as an additional restriction on the system consisting of (**??**) and the equations (3.1) and (3.2) for T and Z. In their application to Catholic Schools, $T = 1(Z > 0)$ and so we focus on that case

$$T = 1(W'\beta + u)$$

They consider both high school graduation and test scores. In the case of graduation,

$$Y = 1(\alpha T + W'\Gamma + \varepsilon).$$

The problem, however, is that (2.5)) is not operational unless $E(\varepsilon|W) = 0$ because Γ is not identified. Mean independence of ε and W is maintained in virtually all studies of selection problems, because without it, α is not identified even if one has a valid exclusion restriction.[8] Our discussion of how the observables are arrived at makes clear that this is hard to justify in most settings. If the observables are correlated with one another, as in most applications, then the observed and unobserved determinants of Y are also likely to be correlated.

[8]The exception is when the instrument is uncorrelated with X as well as ξ, as when the instrument is randomly assigned in an experimental setting.

AET address the problem as follows. Assume that $E(\varepsilon|W)$ is linear. Define G and e to be the slope vector and error term of the "reduced form"

$$E\left(Y-\alpha T \mid W\right) \equiv W'G \tag{4.1}$$

$$Y - E\left(Y-\alpha T \mid W\right) \equiv e. \tag{4.2}$$

They provide sufficient conditions for the coefficients of the projection of T on $W'G$ and e to be equal and thus satisfy Condition 1. The sufficient conditions are the conditions of Theorem 1 and

$$\frac{\sum_{\ell=-\infty}^{\infty} E\left(W_j W_{j-\ell}\right) E\left(\beta_j \Gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(W_j W_{j-\ell}\right) E\left(\Gamma_j \Gamma_{j-\ell}\right)} = \frac{\sum_{\ell=-\infty}^{\infty} E(\tilde{W}_j \tilde{W}_{j-\ell}) E\left(\beta_j \Gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E(\tilde{W}_j \tilde{W}_{j-\ell}) E\left(\Gamma_j \Gamma_{j-\ell}\right)}, \tag{4.3}$$

where $\tilde{W}_j$ is the component of W_j that is orthogonal to the observed variables W, for all elements of W^c. [still screwed up. need to deal with ξ) Roughly speaking (4.3) says that the regression of T on $Y - \alpha T$ is equal to the regression of the part of T that is orthogonal to W on the corresponding part of $Y - \alpha T$. One can show that this condition holds under the standard assumption $E(\varepsilon \mid W) = 0$, in which case G and e equal Γ and ε, respectively. However, $E(\varepsilon \mid W) = 0$ is not necessary for (4.3).[9]

Based on the argument that selection on unobservables is likely to be weaker than selection on observables, they impose (2.5) rather than Condition 1. The upshot is that they work with the system

$$\begin{aligned} Y &= 1(\alpha T + W'G + e). \\ T &= 1(W'\beta + u) \\ 0 &\leq \frac{cov(u,e)}{var(e)} \leq \frac{Cov(W'\beta, W'G)}{Var(W'G)}. \end{aligned}$$

They estimate the set of α values that satisfy the above inequality restrictions. In practice, the lower bound is obtained when equality of selection condition $\frac{cov(u,e)}{var(e)} = \frac{Cov(W'\beta,W'G)}{Var(W'G)}$ is imposed and the upper bound is case in which T is treated as exogenous, with $\frac{cov(u,e)}{var(e)} = 0$.

[9] For example, one can show that (4.3) 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.

4.2 A Bounds Estimator Based on a Factor Model of W_{ij}

4.2.1 A Factor Model of $\mathbf{W}_{ij}$

We now present a factor model of W_{ij}, which is central to the estimator proposed below. The factor model is a convenient way to model the relationship among the covaritates. We assume that W_{ij} has a factor structure

$$W_{ij} = \frac{1}{\sqrt{K^*}} F_i' \Lambda_j + v_{ij}, \ j = 1, ..., K^* \tag{4.4}$$

where F_i is an r dimensional vector. We treat r as finite so while W_{ij} grows, the number of factors remains constant.

We normalize the variance/covariance matrix of F_i to the identity matrix. Define $\sigma_j^2 = E(v_{ij}^2 \mid j)$. When we refer to the "factor model", we will often mean the factor model of W_j, the model (1.1) for Y, and 2.7 which is the model for Z. We continue to assume that ξ_i and ψ_i are independent of all of the W_{ij} and of each other. They may also have factor structures, but the factors are uncorrelated with F_i. The stochastic structure of the model is that Λ_j, Γ_j, β_j and the varaince of v_{ij} differ across j, but are identical for all individuals in the population, $i = 1, ..., N$.

In this model we redefine $\mathcal{G}^{K^*}$ to refer to aspects of the model of W, Y, and Z, that do not vary acros individuals:

$$\mathcal{G}^{K^*} = \left\{ \Gamma_1, .., \Gamma_{K^*}, \beta_1, ..., \beta_{K^*}, \Lambda_1, ..., \Lambda_{K^*}, \sigma_1^2, ..., \sigma_{K^*}^2, S_1, ..., S_{K^*} \right\}.$$

In the appendix we show that this model satisfies assumptions 2 and 3. For estimation, we make the following additional assumptions.

Assumption 8. *(i)* $\left(\Gamma_j, \beta_j, \Lambda_j, \sigma_j^2\right)$ *is i.i.d with fourth moments; (ii) The components* ξ_i *and* ψ_i *of* Y *and* Z *respectively are independent of* W_i^c *and of each other.*

Assumption 8 (ii) implies that there is component of Z_i that is independent of the observed and unobserved determinents of Y. Without this there is no hope of identifying α using Z or a component of Z as a source of exogenous variation in T,because there is no exogenous variation. In the Appendix we verify that the factor model of W in conjunction with the model (2.6) for Y and (2.7) for Z satisfies Assumptions 1, 2, and 3 of Theorem 1.

4.2.2 An Estimator of an Admissible Set for α

In contrast to AET's approach, we use the factor model to directly address the problem posed by the fact that basic introspection about the variables available to the econometrician suggests that W as well as T and Z is correlated with the error term. We study identification under the following assumptions. First, we assume that the econometrician can observe the sequence of models indexed by $K^* = 1, ..., \infty$ and that for each model she observes K (but not K^*) as well as the joint distribution of Y_{iK}, Z_i, T_i and $\{W_{ij} : S_{ij} = 1\}$. The second is that asymptotically $K/K^* \to P_{s0}$. The third is that N gets 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 variables that determine Y is small.

In general the model is not point identified, so we provide an estimator of a set that contains α_0. Our approach is to set α to a hypothesized value, estimate that other parameters of the model and then test whether the data are consistent with the following restrictions implied by the model. The restrictions are

$$0 < P_{s0} \leq 1 \tag{4.5}$$

$$\sigma^2_{\xi 0} \geq 0 \tag{4.6}$$

where P_s and σ^2_ξ are parameters with the true values P_{s0} and $\sigma^2_{\xi 0}$. We construct an estimate of the set of values of α by examining, for each value of α whether the point estimates P_s and σ^2_ξ satisfy these conditions. We then go on to discuss construction of confidence intervals.

With this procedure, we only need to verify the conditions of the model under the true value of α_0.In what follows we will simplify the notation by letting

$$\tilde{Y}_i \equiv Y_i - \alpha_0 T_i.$$

It will be useful to make use of matrix notation. We assume without loss of generality that the variables are ordered so that $j = 1, .., K$ corresponds to the K observed covariates in W^c. Unless indicated otherwise, (Chris, for consistency with the notation we introduce in the factor section, I think we need to write the model as $W\Gamma$ rather than $W'\Gamma$ Also, $W\beta$ instead of $W'\beta$ and $W\pi$ instead of $W'\pi$ I haven't changed this yet)

- 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 $K_o \times 1$ vector of B_{ij} fixing 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 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 $\Lambda_1, .., \Lambda_K$ and $\sigma^2_{v1}, ..., \sigma^2_{vK}$. The moment conditions are the K equations

$$E\left(W_{ij_1}W_{ij_2}\right) = \frac{1}{K^*}\Lambda^2_{j_1} + \sigma^2_{j_1};\ j_1 = 1, ..., K,\ j_1 = j_2 \tag{4.7}$$

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

$$E\left(W_{ij_1}W_{ij_2}\right) = \frac{1}{K^*}\Lambda^2_{j_1}\ ;\ j_1,\ j_2 = 1, ..., K,\ j_1 \neq j_2 \tag{4.8}$$

This is a standard GMM problem. As N grows we will obtain $\sqrt{N}$ consistent estimates of $\sqrt{P_{s0}}\Lambda_j$ for each j and for $\widehat{\sigma^2_j}$ by using the sample analogues to the (4.8 and 4.7) with the left hand side of the equations replaced by $\frac{1}{N}\sum_{i=1}^{N}\left(W_{i_1}W_{ij_2}\right)^2$ and choosing $\widehat{\sqrt{P_{s0}}\Lambda_j}$and $\widehat{\sigma^2_j}$ as the values that minimize the appropriately weighted difference the values of $\frac{1}{N}\sum_{i=1}^{N}\left(W_{i_1}W_{ij_2}\right)^2$ and the predictions summarized in the moment conditions. To simplify the exposition we define $\widehat{\lambda_j}$ to be the GMM estimate of the parameter $\sqrt{P_{s0}}\Lambda_j$ and $\widehat{\lambda}$ the corresponding vector.

Stage 2

We estimate the rest of the parameters in a second stage. To estimate Γ conditional on a hypothesized value for α_0, we take advantage of the moment condition

$$\begin{aligned}\sqrt{K^*}E\left[W_{ij}\tilde{Y}_i\right] &= \sqrt{K^*}E\left[\left(\frac{1}{\sqrt{K^*}}F_i\Lambda_j + v_{ij}\right)\left(\frac{1}{\sqrt{K^*}}\sum_{\ell=1}^{K^*}\frac{1}{\sqrt{K^*}}F_i\Lambda_\ell\Gamma_\ell + \frac{1}{\sqrt{K^*}}\sum_{\ell=1}^{K^*}v_{ij}\Gamma_\ell\right)\right] \\ &= \Lambda_j\left(\frac{1}{K^*}\sum_{\ell=1}^{K^*}\Lambda_\ell\Gamma_\ell\right) + \sigma_{vj}^2\Gamma_j \\ &\xrightarrow{p} \Lambda_j' E(\Lambda_\ell\Gamma_\ell) + \sigma_{vj}^2\Gamma_j\end{aligned}$$

We would like to use the sample analog of this:

$$\sqrt{K^*}\frac{1}{N}W'\widetilde{Y} \approx \frac{1}{K}\frac{1}{P_{s0}}\widehat{\lambda}\widehat{\lambda}'\Gamma + \Sigma\Gamma$$

which leads to

$$\Gamma^{K^*} \equiv \frac{\Gamma}{\sqrt{K^*}} \approx \left[\frac{1}{P_{s0}K}\widehat{\lambda}'\widehat{\lambda} + \widehat{\Sigma}\right]^{-1}\frac{1}{N}W'\tilde{Y}$$

where we introduce the notation Γ^{K^*} for$\frac{\Gamma}{\sqrt{K^*}}$ and define $\widehat{\Sigma}$ is the diagonal matrix composed of $\widehat{\sigma_j^2}$ which is estimated in the first stage. Since P_{s0} is not known, we express Γ^{K^*} as a function of P_S:

$$\hat{\Gamma}^{K^*}(P_S) \equiv \left[\frac{1}{P_sK}\widehat{\lambda}'\widehat{\lambda} + \widehat{\Sigma}\right]^{-1}\frac{1}{N}W'\tilde{Y}. \tag{4.9}$$

We leave implicit the fact that $\hat{\Gamma}^{K^*}$ also depends on the hypothesized value of α used to construct $\tilde{Y}$. Since$\hat{\Gamma}^{K^*}$ is a known function of P_S,we have three remaining parameters to estimate: $\theta = \left(P_s, \phi, \sigma_\xi^2\right) \in \Theta$ with true values $\theta_0 = \left(P_{s0}, \phi_0, \sigma_{\xi 0}^2\right)$, where $\sigma_{\xi 0}^2$ is the true value of $Var(\xi_i)$ and σ_ξ^2 is a hypothesized value. The definition of Θ imposes that $0 \leq P_s \leq 1$ and $\sigma_\xi \geq 0$. One may show that

$$\phi_0 = \frac{\left[E(\Gamma_j\Lambda_j)E(\beta_j\Lambda_j) + E(\Gamma_j\beta_j\sigma_j^2)\right]\left[P_{s0}\left(1-P_{s0}\right)E(\Gamma_j^2\sigma_j^2) + P_{s0}\sigma_{\xi 0}^2\right]}{\sigma_{\xi 0}^2\left[P_{s0}^2E(\Gamma_j\Lambda_j)^2 + P_{s0}E(\Gamma_j^2\sigma_j^2)\right] + \left[E(\Gamma_j\Lambda_j)^2 + E(\Gamma_j^2\sigma_j^2)\right]\left(1-P_{s0}\right)P_{s0}E(\Gamma_j^2\sigma_j^2)}.$$

Using this fact, we arrive at the following estimator for we define our estimator of based

on the following system of equations.

(4.10)

$$q_N^1(\alpha,\theta) = \frac{1}{N}\sum_{i=1}^{N} W_i'\widehat{\Gamma}^{K^*}(P_s) \times \left[Z_i - \phi W_i'\hat{\Gamma}^{K^*}(P_S) - \phi\frac{(1-P_s)\hat{\Gamma}^{K^*}(P_S)\hat{\Sigma}\hat{\Gamma}^{K^*}(P_S)}{(1-P_s)\hat{\Gamma}^{K^*}(P_S)\widehat{\Sigma}\hat{\Gamma}^{K^*}(P_S) + P_s\sigma_\xi^2}\left(\tilde{Y}_i - W_i'\hat{\Gamma}^{K^*}(P_S)\right)\right]$$

(4.11)

$$q_N^2(\alpha,\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - W_i'\hat{\Gamma}^{K^*}(P_S)\right) \times \left[Z_i - \phi W_i'\hat{\Gamma}^{K^*}(P_S) - \phi\frac{(1-P_s)\hat{\Gamma}^{K^*}(P_S)\,(P_s)'\,\widehat{\Sigma}\hat{\Gamma}^{K^*}(P_S)}{(1-P_s)\widehat{\Gamma}_{K^*}(P_s)'\,\widehat{\Sigma}\widehat{\Gamma}_{K^*}(P_s) + P_s\sigma_\xi^2}\left(\tilde{Y}_i - W_i'\hat{\Gamma}^{K^*}(P_S)\right)\right]$$

(4.12)

$$q_N^3(\alpha,\theta) = \frac{1}{N}\sum_{i=1}^{N}\tilde{Y}_i^2 - \left(\frac{\hat{\Gamma}^{K^*}(P_S)\widehat{\lambda}}{P_s}\right)^2 - \frac{\hat{\Gamma}^{K^*}(P_S)\widehat{\Sigma}\hat{\Gamma}^{K^*}(P_S)}{P_s} - \sigma_\xi^2$$

subject to $\theta \in \Theta$.

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

$$\begin{aligned} q_N^1(\alpha,\theta) &= \frac{1}{N}\sum_{i=1}^{N}\left(W_i'\hat{\Gamma}^{K^*}(P_S)\left[Z_i - \phi W_i'\hat{\Gamma}^{K^*}(P_S) - \phi\left(\tilde{Y}_i - W_i'\hat{\Gamma}^{K^*}(P_S)\right)\right]\right) \\ q_N^2(\alpha,\theta) &= \frac{1}{N}\sum_{i=1}^{N}\left(\left(\tilde{Y}_i - W_i'\hat{\Gamma}^{K^*}(P_S)\right)\left[Z_i - \phi W_i'\hat{\Gamma}^{K^*}(P_S) - \phi\left(\tilde{Y}_i - W_i'\hat{\Gamma}^{K^*}(P_S)\right)\right]\right) \end{aligned}$$

These are the classic moment conditions of a regression of Z_i on $(W_i'\hat{\Gamma}^{K^*}(P_S))$ and $(Y_i - \alpha T_i - W_i'\hat{\Gamma}^{K^*}(P_S))$ restricting the regression coefficients be the same. They are the empirical analogue of Corollary 1 of Theorem 1. In the general case the equations are more complicated because the error term ξ leads to attenuation bias.

When $P_S = 1$ and recalling that $\tilde{Y}_i \equiv Y - \alpha T_i$, the second equation is

$$q_N^2(\alpha,\theta) = \frac{1}{N}\sum_{i=1}^{N}\left([Y_i - \alpha T_i] - W_i'\hat{\Gamma}^{K^*}(1)\right) \times \left[Z_i - \phi W_i'\hat{\Gamma}^{K^*}(1)\right]$$

In this case $\hat{\Gamma}$ could be estimated as the regression coefficient of a regression of $Y_i - \alpha T_i$ on W_i. (Our estimator is asymptotically equivalent to this with K fixed and N getting large.) In that case $W_i'\hat{\Gamma}$ would have to be orthogonal to the error term, so this equation would

become

$$q_N^2(\alpha,\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \alpha T_i - W_i'\hat{\Gamma}\right)\times Z_i$$

which is the standard IV moment equation.

Turning to (4.12), $q_N^3(\theta)$ is the difference between the sample value of $var(\tilde{Y}_i)$ for the hypothesized value of α and the variance implied by the model estimate.

We define the estimator $\hat{\theta}(\alpha^*)$ as the set of values of θ that minimize

$$\sum_{\ell=1}^{3}\left[q_\ell\left(\alpha^*, \widehat{\theta}\left(a^*\right)\right)\right]^2.$$

We have not been able to prove that the solution to the moment equations is unique for a given value of α, which is another reason why we allow the estimator of α to be a set. In practice, we have yet to find an example which is not point identified (both in finite samples and at the asymptotic limit). In Theorems 4 and 5 below, we present the asymptotic properties of $\hat{\theta}(\alpha^*)$, assuming that in a neighborhood of α_0, $\theta(\alpha^*)$ is point identified.

4.3 Asymptotic Properties of the Estimator

In this section we show that at the true value of α, α_0, the value of θ_0 is an element of $plim(\widehat{\theta}(\alpha_0))$. We then consider the asymptotic distribution of $\widehat{\theta}(\alpha_0)$.

Assumption 9. *F_i, ξ_i, and ψ_i are all mean 0 and i.i.d. across individuals and are independent of each other with finite second moments. v_{ij} is mean zero and i.i.d. across individuals and covariates with finite variance. The vector $(\Gamma_j, \Lambda_j, \beta_j)$ is i.i.d. across covariates with finite second moments.*

Assumption 10. *The support of the parameters is compact with the support of P_s bounded below by $P_s^\ell > 0$.*

Assumption 11. *The dimension of F_i is 1*

Theorem 4. *Under Assumptions 9-11, at the true value α_0, $\theta_0 \in plim(\widehat{\theta}(\alpha_0))$.*

(Proof in Appendix)

Since θ_0 satisfies the restrictions (4.5,4.6), the theorem implies directly that as K^* goes to infinity and K^*/N goes to 0, the probability that $\widehat{\theta}(\alpha^*)$ satisfies the restrictions is 1 if

$\alpha^* = \alpha_0$. Consequently, our estimator identifies an admissible set for α_0 consisting of the points α^* such that $plim\widehat{\theta}(\alpha^*)$ satisfies (4.5,4.6).

Next we consider asymptotic normality of the estimator.

It will prove useful to define

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

and

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

In the next theorem we show that the limit of $q_n(\alpha, \theta)$ as N gets large is a known function of only θ and the mean of χ_j.This property makes the asymptotic distribution straight forward to figure out.

Theorem 5. *Under Assumptions (4.5,4.6),*

$$q_n\left(\alpha_0, \theta_0\right) \underset{N\to\infty}{\overset{p}{\to}} f\left(\alpha, \theta, \frac{1}{K}\sum_{j=1}^{K}\chi_j\right)$$

where f is a known function. Assume in addition that in the neighborhood of α_0, $\theta(\alpha^*)$ is point identified. Then

$$\sqrt{K}\left(\widehat{\theta} - \theta_0\right) \sim N(0, H'G'VaP_s(\chi_j)GH)$$

where

$$G = \frac{\partial f\left(\alpha_0, \theta_0, \chi_0\right)}{\partial \chi}$$

and

$$H = \left[\frac{\partial f\left(\Theta_0, \chi_0\right)}{\partial \Theta}\right]^{-1}.$$

(Proof in Appendix)

4.4 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 more practical approach which is the parametric bootstrap we use in the Monte Carlo section below.

4.4.1 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). However, we also have limits on the parameter space. In particular $0 < P_S \leq 1$ and $\sigma_\xi^0 \geq 0$. In principle while we cannot get a point estimator for $(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$, we can construct a set estimator for this four dimensional parameter. Our set estimate for α_0 is just the set of α that lie within this identified set.

In practice our estimator is based on the sample analogue of this idea. That is, in principle would would solve for the full set of $(\alpha, \phi, P_S, \sigma_\xi)$ in its parameter set that satisfy $q_N^1 = q_N^2 = q_N^3 = 0$. We would then take our identified region of α as the values of α that lie in this set. That is the full set can be written as

$$\widehat{S}_N = \left\{(\alpha, \theta) \in \mathbb{R} \times \Theta \mid q_N^1(\alpha, \theta) = q_N^2(\alpha, \theta) = q_N^3(\alpha, \theta) = 0\right\},$$

and our estimated set for α can be written as

$$\widehat{\alpha} = \left\{\alpha \in \mathbb{R} \mid (\alpha, \Theta) \cap \widehat{S}_N \neq \emptyset\right\}.$$

Our estimator can be considered as one implemention of this idea. That is we solve for the set $\widehat{S}_N$ in practice by first setting α to some value α^*. We then search for values values of $\theta \in \Theta$ such that $q_N^1(\alpha^*, \theta) = q_N^2(\alpha^*, \theta) = q_N^3(\alpha^*, \theta) = 0$. If we can find such a θ, then α^* belongs in the identified set $\widehat{\alpha}$. By varying α^* across the support of α we can solve for $\widehat{\alpha}$.[10]

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 $T(\alpha, \phi, P_S, \sigma_\xi)$ for which we know its distribution under the null hypothesis: $(\alpha, \phi, P_S, \sigma_\xi) = (\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$.[11] For each $(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ we would construct an acceptance region of the test. When $T(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ lies within this acceptance region, $T(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ would belong to this confidence set, otherwise it would not. Given the confidence set for the full parameter space, we take the confidence set of to be the set of α that lie within this set. More formally

[10] The only real short cut that we take is that there may be multiple values of θ that solve the equations. If we find any value, the set is not empty and so α lies within the estimated region and we do not search for more values. Clearly, modifying the procedure to do this would be straight forward.

[11] A natural choice for a test statistic would be the objective function $\sum_{\ell=1}^{3}\left[q_N^\ell(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)\right]^2$.

let $\widehat{T}(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ be the estimated value of the test statistic and let $T^c(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0)$ the critical value. Assuming we reject when the test statistic is larger than the critical value the confidence set is defined as

$$\widehat{C}_N = \left\{ (\alpha, \theta) \in \mathbb{R} \times \Theta \mid \widehat{T}(\alpha, \theta) \leq T^c(\alpha_0, \phi_0, P_S^0, \sigma_\xi^0) \right\},$$

and our estimated confidence region for α can be written as

$$\widehat{C}_\alpha = \left\{ \alpha \in \mathbb{R} \mid (\alpha, \Theta) \cap \widehat{C}_N \neq \emptyset \right\}.$$

In the appendix we present the algorithm one would use to do this in practice.

4.4.2 A Parametric Boot Strap Procedure

As a practical matter implementing the procedure above in impractical as testing the null over a four dimensional grid is computationally very difficult. A second issue is 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 monotone selection in Manski and Pepper, 2000). While our potential estimation interval can potentially be much more complicated, for the simulations we have run, we consistently find a compact region with one end of the region occuring at the instrumental variable estimate ($P_S = 1$) and the other occuring at the "observable like unobservable restriction" ($\sigma_\xi = 0$). 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 occur at σ_ξ. We propose a parametric bootstrap procedure to construct a one sided confidence interval estimators for α_{low} and $\alpha_{\max}$. The estimator $\hat{\alpha}_{.10low}$ has 10% probability of being below α_{low}. The estimator $\hat{\alpha}_{.10,\max}$ has a 10% nominal probability of exceeding $\alpha_{\max}$.

In doing this we need to add a bit more structure to the model as we are going to need to be able to generate data for T_i. This requires writing a data generating model for T_i and we use the analogue to that for Y_i and Z_i.

$$T_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}\delta_j + \omega_i.$$

Note that in the spirit of instrumental variables we allow ω_i to be correlated with both ξ_i and φ_i. However, we maintain the assumption that ξ_i and ψ_i are independent.

4.4.3 Construction of $\hat{\alpha}_{.10low}$

The procedure for estimating $\hat{\alpha}_{.10low}$ involves the following steps.

1. Estimate the parameters under the model under the assumption that $\sigma_\xi = 0$. We do this 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 $\widehat{\alpha}, \widehat{\varphi}$, and $\widehat{P}_S$. In doing this we also obtain estimates of Λ, Σ, and γ for the observable covariates.

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

   (a) Obtain estimates of the distributions for F_i, v_{ij} 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 4rth 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 $E(W_{ij}^r W_{ij'}^s)$ 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 θ_i and v_{ij}, $j = 1, ..., K$ and choose parameters of the distributions to match the relevant moments.[12] Call the estimates of the additional parameters of the θ_i distribution $\hat{B}_\theta$ and the additional parameters of the v_{ij} distribution $\hat{B}_{v_j}$.[13]

[12] 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.

[13] An alternative is to using 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 is constrained to match the σ_{vj}^2. A nonparametric approach is unattractive from computational point of view. Given that our estimators only involves second moments it does offer any clear advantages.

(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 fourth moments. To obtain them, one can use the fourth moments of $Y_i - \hat{\alpha}T_i$, Z_i and T_i. Consider

$$E(\xi_i^4) = E(Y_i - \alpha T_i)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} W_{ij}\Gamma_j)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} W_{ij}\Gamma_j)^2\sigma_\xi^2.$$

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

$$E(\psi_i^4) = E(Z_i)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} W_{ij}\beta_j)^4 - E(\frac{1}{\sqrt{K^*}}\sum_{j=1}^{K^*} W_{ij}\beta_j)^2\sigma_\psi^2 \ .$$

Note that this requires an estimate of $\widehat{\beta}$ and σ_ψ^2 but estimating these is completely analogous to estimating $\widehat{\gamma}$ and σ_ξ^2 where the dependent variable is now Z_i rather than $Y_i - \alpha T_i$. 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 choose parameters $B_{\xi,\psi,\omega}$. The joint distribution should not contrain the second and fourth moments unless one wishes to impose additional appriori 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 $\hat{B}_\theta$, generate $(\theta_i)^{(b)}$, $(v_{ij})^{(b)}$ 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

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

   on the bootstrap samples.

5. Calculate the 90^{th} quantile of the bootstrap sample of $\widehat{\alpha}_{min}$ and subtract that from our point estimate of $\widehat{\alpha}_{min}$ to obtain the lower bound of our confidence set.

4.4.4 Construction of $\hat{\alpha}_{.90\,\max}$

To obtain $\hat{\alpha}_{.90\,\max}$, 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 W_j that induce positive correlation between T_i and 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 $W'W$ in equation 4.9) for $\Gamma(\hat{P}_S)$, 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 boostrap procedure.

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

5 Monte Carlo Evidence

In this section we present Monte Carlo evidence on the performance of $\hat{\alpha}_{\min}$, $\hat{\alpha}_{AET}$, and $\hat{\alpha}_{OLS}$. We present $\hat{\alpha}_{OLS}$ because in our context $\hat{\alpha}_{\max}$ turns out to be essentially the same as the OLS estimator.[14] We also report the standard deviations, standard error estimates, and the .10 lower bound to the one-sided confidence intervals for $\hat{\alpha}_{\min}$. We do not present such bounds for $\hat{\alpha}_{AET}$.

[14]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 $\sigma^2_\xi > 0$ occured at $P_S = 1$, although we have not proved that this has to be the case for any model with a factor structure. (Todd–we have to check this]

In discussing the design, we first restate the equations of the model of Y_i, T_i, and W_{ij}:

$$\begin{aligned} Y_i &= \alpha_0 T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}\Gamma_j + \xi_i \\ &= \alpha_0 T_i + \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K_o} W_{ij}\Gamma_j + \frac{1}{\sqrt{K}} \sum_{j=K_o+1}^{K^*} W_{ij}\Gamma_j + \xi_i \\ W_{ij} &= \frac{1}{\sqrt{K^*}} \theta_i' \Lambda_j + v_{ij} \\ T_i &= Z_i = \frac{1}{\sqrt{K^*}} \sum_{j=1}^{K^*} W_{ij}^K \beta_j + \psi_i \end{aligned}$$

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

5.1 W parameters

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

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

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.

5.2 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. We have normalized the weights 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.

5.3 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

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

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_ε will not. We view this as an attractive parameterization because we are primarily concerned with ensuring that ϕ and ϕ_ε do not depend on P_S.[15] The expected values of ϕ and ϕ_ε 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 $\hat{\alpha}_{\min}$.

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, R_ψ^2 to 0.5, and α_0 to 1.0 in all the experiments reported, and 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 $E(\beta_j\Gamma_j) = 0.09$, 0.3, and 0.6. Finally, we vary $\bar{\Lambda}$ and $\tilde{\Lambda}_{\max}$. In one case, we set $\bar{\Lambda} = 0$, which means that that $E[Corr(W_{ij}, W_{ij'})] = 0$ if $j \neq j'$. In the other case, $E[Corr(W_{ij}, W_{ij'})] = 0.2$ if $j \neq j'$.

5.4 Monte Carlo Results

We first consider a baseline case in which T_i is randomly assigned. Table MC1 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

[15] 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

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

given the values of the other parameters of the experiment.

does not depend on the W_j. For these designs, $\hat{\alpha}_{OLS}$ is unbiased because $E(\phi) = E(\phi_\varepsilon) = 0$. We use the median as our measure of central tendency but also report the 10th and 90th percentile values. We use the 90th-10th differential as a measure of dispersion. The median values of ϕ, ϕ_ε, and $\widehat{\phi}$ across replications are shown in the 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 $\hat{\alpha}_{AET}$ 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. The results are not very sensitive to the value of P_s. (to be checked.)[16]

We turn next to designs in which OLS estimates of α_0 are biased. In Table MC2a, we set $\mu_\beta = \mu_\Gamma = 0.3$, which leads to bias OLS estimates for the specification we consider. To see this, note that even if $b_\varepsilon = g_e = 0$, so that the elements of β_j and Γ_j are uncorrelated, OLS will be biased if $P_S < 1$ because $E(\beta_j \Gamma_j) = 0.09$. In the top panel of the table, $\bar{\Lambda} = 0$, so that $E[Corr(W_{ij}, W_{ij'})] = 0 \; \forall \; j \neq j'$. We consider the $b_\varepsilon = g_e = 0$ case in the first three columns of the table. In the first column, with $P_S = 0.2$, ϕ and ϕ_ε are small. (The median of $\phi = -.071$ and the median of $\phi_\varepsilon = .03$. For this design $\phi = \phi_\varepsilon$ and both are positive, so the difference reflects sampling error.) The bias in OLS in this case is small regardless of the value of P_S. The precision of the OLS estimator is also essentially invariant to the value of P_S.[17] In contrast, the performance of the $\hat{\alpha}_{AET}$ and $\hat{\alpha}_{\min}$ estimators improves as P_S increases. $\hat{\alpha}_{AET}$ exhibits some downward bias when $P_S = 0.2$, but $\hat{\alpha}_{\min}$ is approximately unbiased in all cases. $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{AET}$ are noisier than OLS but not dramatically so when $P_S = 0.8$.

In the next three columns of the table, we chose b_ε and g_ε so that $E(\Gamma_j \beta_j) = 0.3$. Not surprisingly, the upward bias in OLS is higher than the corresponding cases in the first three columns of the table, with the median of $\widehat{a}_{OLS}$ rising to 1.256 when $P_S = 0.2$ and 1.101 when $P_S = 0.8$. Again, $\hat{\alpha}_{\min}$ is essentially unbiased in all three cases, with the dispersion

[16] Note for the future. Our setup for the Monte Carlos does not really allow us to capture the intuition that a really well-informed researcher will put in the variables that matter for Y and T first. We vary How B and G are related for all Ws at the same time, and we vary the role of psi. We should revisit this, particularly if we can successfully modify the theory to allow the Prob(Sj) to vary with j, as Chris and I were discussing while at NBER

[17] It is surprising that the bias in OLS is not monotone decreasing in P_S. Sampling error may be the reason, but phenomen shows up in several places in the tables.

declining with P_S. The last three columns increase b_ε and g_ε so that $E(\Gamma_j\beta_j)$ to 0.6 and $Corr(\Gamma_j, \beta_j) = .51$. 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 $\widehat{\alpha}_{AET}$ and $\widehat{\alpha}_{\min}$ estimators are less noisy as $E(\Gamma_j\beta_j)$ increases. When $E(\Gamma_j\beta_j) = 0.6$ and $P_S = 0.8$ (column 6) shown in the last column, the $\widehat{\alpha}_{AET}$ and $\widehat{\alpha}_{\min}$ estimators have no more sampling error than the OLS estimator. The bootstrap confidence interval estimator is reasonably close to the 90% nominal value. (Check when we have results.)

Table MC2b repeats the calculations found in Table MC2a 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, perhaps 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 that comes from θ is associated with an improvement in the performance of $\hat{\alpha}_{\min}$ relative to $\hat{\alpha}_{AET}$. In particular, $\hat{\alpha}_{AET}$ is substantially downward biased unless $E(\Gamma_j\beta_j) = 0.6$ or $P_S = 0.8$. This may be due to fact that the $\hat{\alpha}_{AET}$ estimator is based on the assumption that the restriction $\phi = \phi_\varepsilon$ 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. The positive correlation between the observed and unobserved covariates that is present in these designs results in positive omitted variables bias on the observed $\hat{\Gamma}_j$. The bias arises 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 may lead the projection of T on $W_i\Gamma^P$ to overstate the amount of selection bias, inducing a negative bias in the AET estimates of α_0. This negative bias also affects the OLS estimator, partially counteracting the positive bias caused by correlation of T with the unobserved elements of W. This is why the positive bias on the OLS estimates is smaller in Table MC2b than in Table MC2a.

Most importantly, $\hat{\alpha}_{\min}$ performs very well in the presence of a factor structure. It had a median close to 1 in all cases and a 90th-10th differential that is similar to OLS in the cases in which $E(\Gamma_j\beta_j) = 0.3$ or 0.6. This superior performance of $\hat{\alpha}_{\min}$ relative to $\hat{\alpha}_{AET}$ 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.

In Table MC3a, we relax the assumption that the observables are a random set of all the unobservables by setting $R^2_\xi = 0.2$. In the top panel, $\bar{\Lambda} = 0$ and $\widetilde{\Lambda}_{\max} = 6.2$, as in Table MC2a. Not surprisingly, allowing a positive variance for ξ has no effect on the median of OLS. However, the lower bound estimators $\hat{\alpha}_{AET}$ and $\hat{\alpha}_{\min}$ are now both downward biased for α because the assumption that $\phi = \phi_\varepsilon$ no longer holds. This is easiest to see in the three cases in which P_S equals 0.8; in all three cases ϕ_ε is approximately equal to 0.8ϕ; in other words, selection on unobservables is now only 80 percent as large as selection on observables. When $E(\Gamma_j\beta_j) = 0.3$, the median of $\hat{\alpha}_{AET}$ varies from 0.907 to 0.975 depending on P_S, and the corresponding median values of $\hat{\alpha}_{\min}$ are 0.976, 0.956, and 0.979. However, the sampling variance of the $\hat{\alpha}_{AET}$ and $\hat{\alpha}_{\min}$ estimators is quite 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 in table MC2a, while there is no systematic change for the other estimators. The sampling variances of $\hat{\alpha}_{AET}$ and $\hat{\alpha}_{\min}$ are wider in this case than in the analogous cases in Table MC2a (in which the assumption $\phi = \phi_\varepsilon$ holds.). We do not fully understand this pattern, but in spite of it, the lower bound estimators usefully complement OLS. (Add a sentence about the confidence interval estimates when we have them).

Table MC3b again allows for correlation among the elements of W_j by setting $\bar{\Lambda}$ and $\widetilde{\Lambda}_{\max}$ so that $E[Corr(W_{ij}, W_{ij'})] = 0.2$. Relative to the iid case, the performance of $\hat{\alpha}_{\min}$ improves substantially, with median values that are close to 1.0 for all cases. The sampling distribution narrows substantially, perhaps reflecting the fact that when the W_j are correlated, it is easier to "fill in" for the effects of missing covariates using our moment conditions, so that it matters less which elements of W^* are actually observed. Relative to the values in Table MC3a, the negative bias of the $\hat{\alpha}_{AET}$ estimator increases and the positive bias of the $\hat{\alpha}_{OLS}$ declines, again reflecting positive correlation between the observed and unobserved elements of W^*.

Finally, Tables MC4a and MC4b are analogous to tables MC3a and MC3b, except now we set $R^2_\xi = 0.4$, thereby lowering ϕ_ε 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.. The performance of $\hat{\alpha}_{AET}$ is poor in all three cases in which $P_S = 0.2$, with large sampling errors and negative bias. The medians of $\widehat{\alpha}_{\min}$ range between 0.786 and 0.982, but this estimator is noisy relative to OLS except when $P_S = 0.8$ and $E(\Gamma_j\beta_j) = 0.6$. As we saw earlier in a comparison of Tables MC3a and MC3b, the performance of $\widehat{\alpha}_{\min}$ improves substantially when $E[Corr(W_{ij}, W_{ij'})] = 0.2$. There appears to be negative bias in all cases, but this

bias is typically small relative to the positive bias in $\widehat{\alpha}_{OLS}$. The negative bias in $\widehat{\alpha}_{AET}$ is substantial in most cases, reflecting the fact that $\phi > \phi_{\varepsilon}$ as well as the positive correlation between the observed and unobserved elements of W.

In Table MC5 we explore the performance of the bootstrap procedure for a few of the designs.

The monte carlo results may be summarized as follows. First, the median of $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{AET}$ are similar when there is no factor structure, although $\widehat{\alpha}_{\min}$ is less dispersed, particularly when $P_S = .2$. $\widehat{\alpha}_{\min}$ performs much better than $\widehat{\alpha}_{AET}$ when there a factor structure. Second, both $\widehat{\alpha}_{\min}$ and $\widehat{\alpha}_{AET}$ are biased down when $\phi > \phi_{\varepsilon}$. This is to be expected, because both estimators are based on the assumption that $\phi = \phi_{\varepsilon}$ and are to be interpreted as lower bound estimators if $\phi > \phi_{\varepsilon} > 0$ (in the case $\phi > 0$). Third, the downward bias in $\widehat{\alpha}_{\min}$ when $\phi > \phi_{\varepsilon}$ is reduced considerably when there is a factor structure, at least in the cases we consider. Fourth, precision is worse than with OLS. The loss of precision depends on the design and is neglible in the case in which T is randomly assigned (Table MC1). However, $\widehat{\alpha}_{\min}$ is sufficiently precise to provide useful information about α in all of the cases that we consider.

6 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 observed variables.and 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 be in between. Our estimators build on Theorem 1, which concerns the coefficients of the projection of outcome on the regression indices of the observables and the

unobservables it. 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 ES and the $ES - Factor$ 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 ES estimator combines Theorem 1 with a high level assumption about the link among the observed and unobserved variables. The $ES - Factor$ estimator adds the assumption that the observed and unobserved explanatory variables have a factor structure. The factor structure provides additional moment restrictions that permit one to account for the effects of omitted variables. We show that the estimator identifies a set that contains the true value of the treatment parameter. We derive the asymptotic distribution of the $ES - Factor$ estimator and present a parametric bootstrap approach to statistical inference. Our monte carlo simulations are generally encouraging, particularly for $ES - Factor$.

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 $ES - Factor$ estimator, and we have not performed monte carlo studies for designs with multiple factors. The ES 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 W_j is the more promising long run research path. The linear factor model that we employ in developing The $ES - Factor$ estimator is a natural way to do this, but it 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 the index that determines the outcome is an additively separable function.

More generally, we think of ES and $ES - Factor$ as a start for 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., "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, "The Effectiveness of Catholic School," under revision, Northwestern University, 1999.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "IV Strategies for Evaluating the Catholic School Effect", (2001, in progress).

Angrist and Evans, "Children and their Parent's Labor Supply: Evidence from Exogenous Variation in Family Size" *American Economic Review*, 88 (1988), 450-477.

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

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., "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy,* 106(1998), 262-333.

Cameron, Stephen V. and Christopher R. Taber, "Estimating Borrowing Constraints Using the Returns to Schooling," unpublished manuscript, Nortwestern University, 2001.

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

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

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

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

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

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

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

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

Heckman, J., and Robb, R., "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, 1985

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

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

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

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

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

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

Murphy, Kevin M., and Robert H. Topel, "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, 1990, 204-40.

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

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

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

Staiger, Douglas, and James Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica,* (65) No. 3 (1997), 557-586.

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

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

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

Wooldridge, Jeffrey, *Introductory Econometrics*, South-Western College Publishing, 2000.