Treatment Effect Estimation with Noisy Conditioning Variables

Kenichi Nagasawa*

September 10, 2021

Abstract

When estimating causal effects, controlling for confounding factors is crucial, but these characteristics may not be observed. A widely adopted approach is to use proxy variables in place of the unobserved ideal controls. However, this approach generally suffers from measurement error bias. In this paper, I develop a new identification strategy that addresses this issue. I use proxy variables to construct a random variable conditional on which treatment variables become exogenous. The key idea is that, under appropriate conditions, there exists a one-to-one mapping between the distribution of unobserved confounding factors and the distribution of proxies. To satisfy overlap/support conditions, I use an additional variable, termed excluded variable, which satisfies certain exclusion restrictions and relevance conditions. I also establish asymptotic distributional results for flexible parametric and nonparametric estimators of the average structural function. I demonstrate empirical relevance of my results by estimating causal effects of Catholic schooling on college enrollment.

Keywords: average structural function, bounded completeness, control functions, non-classical measurement errors.

*The University of Warwick, Department of Economics. Email: kenichi.nagasawa@warwick.ac.uk Address: Department of Economics, University of Warwick, Coventry, CV4 7AL, United Kingdom.

1 Introduction

In observational studies, controlling for confounding factors is crucial to identify causal effects of interest. The main challenge is that measuring these confounding elements are often difficult, if not impossible. A widely used approach to address this issue is to use proxy variables in place of the unobserved characteristics. For instance, to account for differences in unobserved worker's ability, researchers routinely include test scores in their regression controls. However, these measurements are generally contaminated with noise. In simple linear regression settings, it is well-known that using proxies as regressors induces attenuation bias. In nonparametric settings, Battistin and Chesher (2014) showed that average treatment effects are not identified under covariate measurement errors and that the direction of the bias cannot be determined a priori. Therefore, simply controlling for proxy variables does not lead to reliable inference on causal effects. In this paper, I provide a novel identification strategy for causal effects when control variables are measured with errors. This new approach has a wide range of applications since coarse measurements are prevalent in practice.

In its simplest form, the identification problem of interest is captured in the model

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_2(X^* - X) + \epsilon, \qquad \mathbb{E}[\epsilon|D, X^*, X] = 0$$

where D is the treatment of interest and X is a proxy variable for the unobserved confounding factor X^*. If X^* were observed, one would estimate the equation $Y = \beta_0 + \beta_1 D + \beta_2 X^* + \epsilon$. In practice, however, one regresses Y on D and X, using the error-ridden variable in place of X^*. The regression estimate using X suffers from measurement error bias, and the treatment effect β_1 is not consistently estimated. A textbook solution to this problem is to use an additional proxy variable as an instrumental variable (IV) for X. Denoting the second measurement of X^* by Z, the two-stage least squares (2SLS) method is equivalent to estimating

$$Y = \beta_0 + \beta_1 D + \beta_2 \mathbb{E}[X|D, Z] + \varepsilon, \qquad \mathbb{E}[\varepsilon|D, Z] = 0 \tag{1}$$

where $\mathbb{E}[X|D, Z]$ is used as it is the fitted value in the first-stage equation.

To motivate my identification result, I make two observations on (1). The first observation is that 2SLS is equivalent to the regression using the conditional mean of the proxy variable X as

an additional control. In a simple linear model, controlling for the mean of X suffices to remove the measurement error bias, but for more realistic models, one may want to include additional controls such as conditional variance and higher moments of X. In fact, my main identification result shows that flexibly controlling for the conditional distribution of X enables identification of treatment effects under appropriate overlap/support conditions. The identification result holds in a broad class of nonparametric models. For estimation and inference, I build on recent results in the program evaluation and control function literature (e.g., Arkhangelsky and Imbens, 2019; Chernozhukov et al., 2020) and impose structures on the proxy distribution and the outcome equation. My proposed procedure is easy to implement, only involving regression estimations and averaging over sample observations.

For the second observation, note that the role of Z in identification is to create sufficient variation in $\mathbb{E}[X|D,Z]$ so that the standard rank condition holds. From this perspective, Z can be independent of X^* (or X), and instead Z can be correlated with D.[1] This point indicates that Z does not need to be a proxy for X^*, and the set of valid Z is larger than the set of proxy variables. For example, an IV for the treatment variable can be used as Z. Other candidates for Z will be discussed below. To emphasize that Z is not limited to a proxy variable, I call Z "excluded variable". Excluded variables are known as negative control exposure in the biostatistics literature (see e.g., Miao et al., 2018). It is worth noting that the type of exclusion restriction Z needs to satisfy is weaker than the standard IV exclusion restriction as Z is allowed to be correlated with X^*. This feature is practically important as justification of IV exclusion restrictions often requires considerable efforts. Below I describe formal requirements on excluded variables.

The main contribution of this paper is the novel identification result for treatment effects. In particular, I develop a new construction of control functions based on proxy and excluded variables.[2] The majority of existing results use IV for the treatment to construct a control function (for reviews, see Blundell and Powell, 2003; Matzkin, 2007; Wooldridge, 2015). My identification strategy instead uses a proxy variable for unobserved confounders. As my approach is applicable in observational settings, this paper provides a useful alternative to identification strategies that

[1] Even if $Z \perp\!\!\!\perp X$, the first-stage equation $X = \gamma_1 + \gamma_2 D + \gamma_3 Z + \eta$ may have $\gamma_3 \neq 0$ if $\text{Cov}(Z,D) \neq 0$ since D is correlated with X^* (and X).

[2] A control function is an estimable function of observed variables, conditional on which treatment variables become exogenous (Matzkin, 2007, p.5356). This conditional independence property of a control function is also known as balancing property in the causal inference literature.

rely on natural/quasi-experimental variation in datasets.

To demonstrate empirical relevance of my results, I apply my method to estimating causal effects of Catholic high school attendance on four-year college enrollment, building on the analysis of Altonji et al. (2005). The previous studies found that identification strategies based on IV might not be credible. By avoiding the use of IV, my identification result provides a useful alternative. I used test scores as proxy measurements of unobserved academic ability to construct a control function. Compared with standard methods, my approach corrects for selection bias and measurement error bias.

Related literature This paper adds to the extensive literature on control function methods. As already mentioned, many existing results construct a control variable from IV, whereas I use a proxy and an excluded variable. Another important distinction is that my result handles discrete and continuous treatment variables in a unified framework, whereas control function methods using IV often focus on a specific type of treatment variable. In particular, I do not impose strict monotonicity in the first-stage equation (e.g., Imbens and Newey, 2009). The distinction is practically important as monotonicity may not hold in some applications e.g., unordered discrete choice as a treatment variable. This difference arises as I do not explicitly model the first-stage equation and instead focuses on the relationship between unobserved heterogeneity and its proxy variable.

As discussed below, what my identification strategy essentially does is to find values of treatment and excluded variables $(d_1, z_1), (d_2, z_2)$ for which the conditional proxy distribution is invariant i.e., $\Pr[X \in \mathcal{A}|D = d_1, Z = z_1] = \Pr[X \in \mathcal{A}|D = d_2, Z = z_2]$ for all measurable sets $\mathcal{A}$. This feature is closely related to the exchangeablility condition used by Altonji and Matzkin (2005). In settings with group structure, Arkhangelsky and Imbens (2019) provided a framework where the exchangeability condition follows from model primitives. In this paper, I also provide a framework where the exchangeability-type condition holds but my model differs in not having an explicit group structure.

I motivated my identification strategy as a generalization of 2SLS estimation using repeated measurements. While the existing literature on non-linear measurement error models is extensive (for a review, see Schennach, 2020, and references therein), my approach is unique in the use of an excluded variable. The set of excluded variables contains the set of proxy variables. Thus, when

repeated measurements are available, my approach is applicable, but not vice versa. In addition, my approach is distinct from deconvolution methods.

I use a completeness condition to formalize the relationship between proxies and unobserved heterogeneity, which was inspired by Hu and Schennach (2008). Their operator diagonalization technique has been successfully applied beyond measurement error contexts (e.g., Arellano et al., 2017; Bonhomme et al., 2019; Sasaki, 2015). A main distinction of my approach is that I require a rank condition on only one proxy variable, as opposed to two in the operator diagonalization technique. A cost of this weaker requirement is that I do not recover the distribution of unobserved confounding elements while the other approach does. In other words, I regard the distribution of unobserved heterogeneity as a nuisance parameter and focus on treatment effects. Also, I allow for measurement errors only in covariates whereas the operator diagonalization method handles measurement errors in treatment variables as well. Since there is a trade-off between the strengths of assumptions and results, I view the two approaches as complementary.

My results are also related to recent studies using proxy controls (e.g., Deaner, 2021; Miao et al., 2018). The conditional independence assumptions they impose are similar to those I use in this paper. The main distinction is that their approaches hinge on identifying reduced form parameters using the outcome equation by solving integral equations. In contrast, my approach addresses the endogeneity problem without using the outcome variable, and thus, it follows the spirit of design-based approach in the causal inference literature.

Roadmap In the next section, I describe the econometric model and discuss the identification results. Nonparametric and flexible parametric estimation methods are developed in Section 3, and Section 4 applies the results of this paper to estimating causal effects of Catholic schooling on college attendance using the National Longitudinal Study 1972 dataset. Section 5 concludes.

2 Econometric model and identification results

I employ the potential outcome notation. $\{Y(d) : d \in \mathcal{D}\}$ denotes the set of potential outcomes, $\mathcal{D}$ is the set of possible treatment levels, D is the realized treatment level, and X^* represents unobserved confounding factors. The treatment variable may have a discrete, continuous, or mixed distribution.

As a concrete example, let the outcome of interest be wage, D be educational attainment, and X^* be worker's ability. Interest lies in causal effects of educational attainment on wage level, and if we were to observe worker's ability X^*, causal effects may be identified by controlling for X^*. However, we do not observe worker's ability, creating the identification problem. In this context, a widely adopted approach is to control for variables X that proxies X^*. For instance, test scores are frequently used as a proxy for ability. Although controlling for proxy variables is a common empirical practice, it is theoretically unsatisfactory as measurement errors induce bias. Specifically, Battistin and Chesher (2014) showed that controlling for X instead of X^* does not identify the average treatment effects and that the sign of the bias cannot be determined a priori. They provided a method for sensitivity analysis over different magnitudes of the measurement error variance. In this paper, I develop a control function approach using a proxy variable X and what I call excluded variables, denoted by Z. As already mentioned, additional proxies can be used as excluded variables. In the wage example, researchers often observe multiple test scores (across time periods, or of different subjects) and one set of test scores can be used as X and others as Z. I discuss other candidates for Z after presenting identification results.

One parameter of interest is

$$\beta(d) = \mathbb{E}[\mathscr{T}(Y(d))], \qquad d \in \mathcal{D}$$

where $\mathscr{T}$ denotes a known transformation and the expectation is taken with respect to the marginal distribution of the potential outcome. Since the distribution of unobserved heterogeneity is held constant as d varies, the change in $\beta(d)$ represents the ceteris-paribus effect of the treatment on the mean outcome. As an example, take $\mathscr{T}(y) = y$ and $\mathcal{D} = \{0, 1\}$, and then, $\beta(1) - \beta(0)$ is the average treatment effect (ATE). When the treatment is continuous, $\beta(d)$ represents the dose-response function. Also, by taking $\mathscr{T}(y) = \mathbb{1}\{y \leq c\}$ for some $c \in \mathbb{R}$, $\beta(d)$ becomes the distribution structural function, from which the quantile structural function can be constructed. For concreteness, I focus on the identity map $\mathscr{T}(y) = y$ and in this case, $\beta(d)$ is referred to as the average structural function (ASF).

2.1 Conditional independence result

I impose the following model restrictions.

Assumption 1. $X \perp\!\!\!\perp (D, Z)|X^*$.

Assumption 2. $Y(d) \perp\!\!\!\perp D|X^*$.

Assumption 3. $Y(d) \perp\!\!\!\perp Z|D, X^*$.

Assumption 4. *For each $d \in \mathcal{D}$, $\mathbb{E}[|Y(d)|] < \infty$. For some σ-finite measures $\lambda, \lambda_x, \lambda_d, \lambda_z$, the distribution of $\{X^*, X, D, Z\}$ is absolutely continuous with respect to the product measure $\lambda \times \lambda_x \times \lambda_d \times \lambda_z$. The conditional densities $f_{X|X^*}$, $f_{X^*|DZ}$ are uniformly bounded. If the support of X is uncountable, the conditional density $f_{X|DZ}$ is continuous in the X argument except at a finite number of points with probability one.*

Assumption 5. *For a real-valued, bounded, and λ-integrable function g, define the operator*

$$\Pi(g)(\cdot) = \int g(x^*) f_{X|X^*}(\cdot|x^*) d\lambda(x^*).$$

On the set of bounded and λ-integrable functions, Π is injective.

Assumption 1 states that given the "correctly measured" variable, its noisy measurement is independent of other variables. This type of restriction is common in the measurement error literature (see e.g., Assumption 2 in Hu and Schennach, 2008).

If we were to observe X^*, Assumption 2 would be the selection-on-observables assumption. This restriction would be intuitive if a researcher is willing to specify what X^* consists of (e.g, ability, motivation). Otherwise, X^* would denote some "index" of various unobserved characteristics. For instance, in the binary treatment case, a widely used framework is the threshold crossing model $D = \mathbb{1}\{\varphi(Z) \geq X^*\}$, where φ is some non-stochastic function and X^* is confined to be a scalar. In this model, a researcher can be agnostic about the identity of the underlying unobserved heterogeneity, but they impose the scalar index restriction. Being explicit about what X^* represents is crucial in assessing the plausibility of the identifying assumptions in specific contexts. Thus, a researcher should either specify elements of X^* or consider some explicit econometric model.[3]

[3] An example for continuous treatment variables is a random coefficient model: $D = Z'X^*$ where the dimensionality of X^* is restricted by the dimension of Z.

Assumption 3 is a version of exclusion restrictions for Z. It imposes that after conditioning on the treatment and "ideal" controls, the excluded variable Z has no impact on the outcome of interest. This restriction is weaker than the IV exclusion restriction because it allows for dependence between Z and X^*. Assumptions 2 and 3 together imply $Y(d) \perp\!\!\!\perp (Z, D)|X^*$, which is what I use in the end, but I state the two assumptions separately to clarify requirements on excluded variables Z.

Figure 1 collects the conditional independence restrictions using a directed acyclical graph (DAG). Deaner (2021) and Miao et al. (2018) have similar DAG representations in their identification arguments. As the DAG shows, the excluded variable Z may be correlated with the outcome only through the treatment and the unobserved confounding factor. The proxy variable X may be correlated with the outcome, but it has to be independent of Z and D conditional on X^*.

Assumption 4 is a set of mild regularity conditions. It accommodates various distributions of D, and the identification argument goes through without modifications for discrete, continuous, and mixed treatment variables. Also, it handles discrete and continuous X in a unified way.

Assumption 5 formalizes the idea that the noisy measurement X has a strong relationship with X^*. This formulation follows Assumption 3 of Hu and Schennach (2008), and there have been a growing number of subsequent studies that use the same injectivity condition for identification. Hu and Schennach pointed out this assumption is analogous to the bounded completeness of the conditional distributions of X^* given X. Completeness conditions can be thought of as a generalization of the IV rank condition in linear models to nonparametric settings (Newey and Powell, 2003). Since this is a rank condition, the dimension of X should be at least as large as that of X^*. In the literature, there are several known sufficient conditions for completeness (e.g., Andrews, 2017; D'Haultfoeuille, 2011; Hu et al., 2017). For instance, if researchers are willing to impose the measurement error structure such as $X = \phi(X^* + \eta)$ where ϕ is invertible and $X^* \perp\!\!\!\perp \eta$, then primitive sufficient conditions for the bounded completeness exist. In a panel data setting, Wilhelm (2015) discussed justifications for completeness assumptions using past observations as proxies.

Completeness assumptions often lead to ill-posed inverse estimation problems because most existing studies use such conditions to compute the left-inverse of integral operators. My identification argument uses Assumption 5 to guarantee a unique solution to an integral equation but I do not need to estimate the inverse of the integral operator when implementing my procedure. Therefore,

I circumvent ill-posed inverse estimation problems associated with completeness conditions. This feature is appealing as ill-posed inverse problems may lead to poor finite-sample performance of estimation procedures.

With the above assumptions, I now state the first main result. Define

$$V = \{f_{X|DZ}(x|D, Z) : x \in \mathcal{X}\}$$

where the conditional density is with respect to the measure in Assumption 4 and $\mathcal{X}$ is the support of X. Since (D, Z) is a random vector, V is a stochastic process indexed by the support of X. The following lemma states that this stochastic process is a valid control function. A sketch of proof is presented in Section 2.5, and a formal proof is found in the supplemental appendix.

Theorem 1. *If Assumptions 1-5 hold, then*

$$Y(d) \perp\!\!\!\perp D|V \qquad \forall d \in \mathcal{D}.$$

Given this conditional independence result, I can identify objects such as the average and quantile structural functions provided that a common support condition holds (see e.g., Blundell and Powell, 2003; Imbens and Newey, 2009).

If the proxy X has a finite support with L points, Theorem 1 states that the random vector $(\Pr[X = x_\ell|D, Z])_{\ell=1}^{L}$ is a valid control function. A proxy variable with a finite support can satisfy the identifying assumptions when the unobserved confounding factor X^* also has a finite support. In labor economics and industrial organization, unobserved heterogeneity as finite discrete types is often used as a tractable modelling device (e.g., Keane et al., 2011). If researchers believe that discrete confounding factor is a good approximation to the underlying data generating process, then they can discretize proxy variables (if not already discrete) to construct a control function. This approach has an advantage that given the estimated control function $\widehat{V}$, one can directly apply standard techniques for estimation and inference of treatment effects.

When the support of X is large, Theorem 1 may not be directly applicable as controlling for the stochastic process may be infeasible. To overcome this issue, I follow the approach of Arkhangelsky and Imbens (2019) who modelled the conditional distribution of covariates for dimension reduction.

In particular, I impose that the conditional proxy distribution admits a finite-dimensional sufficient statistic.

Assumption 6. *There exists some fixed function θ such that*

$$f_{X|DZ}(x|d_1,z_1) = f_{X|DZ}(x|d_2,z_2) \quad \forall x \in \mathcal{X} \qquad \iff \qquad \theta(d_1,z_1) = \theta(d_2,z_2)$$

and the function θ is identifiable from the joint distribution of (X,D,Z).

With the additional structure of Assumption 6, the random vector $\theta(D,Z)$ possesses the balancing property and we set $V = \theta(D,Z)$ to identify treatment effects. The following lemma states this result.

Lemma 1. *Under Assumptions 1-6,*

$$Y(d) \perp\!\!\!\perp D|\theta(D,Z).$$

To see what form θ may take, first consider some parametric family of distributions (e.g., normal, Poisson, etc.), represnted as $f(x,\vartheta)$ where ϑ denotes the parameter vector. Then, we can take $f_{X|DZ}(x|d,z) = f(x,\theta(d,z))$ for some function θ. Since the conditional proxy distribution is completely characterized by $\theta(d,z)$, the above restriction holds. Next, consider the case $X \in \mathbb{R}$ and

$$X = m(D,Z) + \sigma(D,Z)\eta, \qquad \eta \perp\!\!\!\perp (D,Z)$$

where η is some random variable whose distribution is left unspecified and $m(D,Z), \sigma^2(D,Z)$ are the conditional mean and variance of X. Equivalently, the conditional density of X given (D,Z) is $f(\frac{x-m(d,z)}{\sigma(d,z)})$ where f is the density of η, which is left unspecified. Thus, $\theta(D,Z) = (m(D,Z), \sigma^2(D,Z))$. This location-scale family modelling can be extended to multi-dimensional X in a straightforward way.

For another example, we may consider

$$f_{X|DZ}(x|D,Z) = \sum_{k=1}^{\infty} \zeta_k(D,Z) e_k(x) \tag{2}$$

where $\{\zeta_k, e_k\}_{k\geq 1}$ is a collection of non-stochastic functions. When the conditional density is smooth, we can use a set of approximating functions (e.g., polynomials) to express the density in an infinite sum. Alternatively, when $\mathcal{X}$ is a closed interval and $f_{X|DZ}(x|D,Z)$ is bounded and continuous in x, the Karhunen–Loève theorem implies the above expansion. If (2) holds, then $\theta(D,Z) = \{\zeta_k(D,Z)\}_{k\geq 1}$ satisfies Assumption 6, although this does not reduce the dimensionality of the problem. For dimension reduction, one may assume that there exists a finite number of $\zeta_k(D,Z)$ terms conditional on which Assumption 6 holds. This restriction is true in the cases of parametric and location-scale families. When it does not hold exactly, it may still be a good approximation of the underlying data generating process.

In the sequel, I focus on the location-scale family model of proxy variables for concreteness. With this modelling choice, my identification strategy first estimates

$$V = \{\mathbb{E}[X|D,Z], \mathbb{E}[X^2|D,Z]\}$$

and using this first-step estimate as an additional control, treatment effects can be estimated with standard approaches. If researchers find it necessary to include additional control variables, then they can choose a more flexible model of the proxy distribution, which may include other distributional features such as higher moments. Since the choice of V is essentially the problem of selecting appropriate controls in the main regression, one can leverage existing results on treatment effect estimation in high-dimensional settings (e.g., Belloni et al., 2017).

2.2 Identification of causal effects: overlap/support condition

The conditional independence result $Y(d) \perp\!\!\!\perp D|V$ is not sufficient for identification of causal effects such as the ASF. The remaining important condition is the overlap/support condition. Specifically, to nonparametrically identify treatment effects, the conditional support of V given $D = d$ needs to equal the marginal support of V for relevant treatment level d (see e.g., Assumption 2 of Imbens and Newey, 2009). As a necessary condition of this requirement, Z has to be correlated with either (i) X^* (X) or (ii) D: otherwise $\mathbb{E}[X|D,Z]$ would be independent of Z and the common support condition would fail. Yet, non-zero correlation between Z and (X^*, D) is not sufficient, and the support invariance property often requires a large support of Z. This large support condition can

be stringent in empirical applications, and there have been various proposals to address this issue (e.g., Florens et al., 2008; Newey and Stouli, 2019). In this paper, I build on these existing results to propose feasible estimation procedures. In particular, I follow the approach of Chernozhukov et al. (2020) (CFNSV henceforth) who considered flexible parametric estimation of various causal effects.

A version of CFNSV's approach postulates that

$$\mathbb{E}[Y|D,V] = \psi\big(p(D,V)'\gamma_0\big) \tag{3}$$

where ψ is a known, strictly monotonic link function, p is a vector of user-chosen transformations, γ_0 is the parameter to be estimated, and $'$ denotes the matrix transpose. For a continuously distributed outcome, CFNSV motivated this specification by the random coefficient model

$$Y = X_1^* + X_2^* D, \qquad \mathbb{E}[X_1^*|D,V] = p_1(V)'\gamma_1, \quad \mathbb{E}[X_2^*|D,V] = p_2(V)'\gamma_2$$

where the conditional independence $X_l^* \perp\!\!\!\perp D|V$, $l = 1, 2$ follows under the hypothesis of Theorem 1. CFNSV pointed out the above model allows for heterogeneous responses to treatment, which is an important feature of many empirical settings, while keeping the model parsimonious. With this setup, (3) holds with the identity link function, $p(D,V) = (p_1(V)', Dp_2(V)')'$, and $\gamma_0 = (\gamma_1', \gamma_2')'$. For the case of a binary outcome, one may consider the outcome equation

$$Y = \mathbb{1}\{\gamma_1 D - X^* \geq 0\}$$

and using Theorem 1, $\mathbb{E}[Y|D,V] = F(\gamma_1 D|V)$ where F is the conditional distribution of X^* given V. This is a semiparametric model studied by Blundell and Powell (2004). To further simplify, we may assume that the conditional distribution of X^* given V is a location family i.e., $F(\cdot|V) = \psi(\cdot - p_2(V)'\gamma_2)$, which gives rise to the specification (3) with $p(D,V) = (D, p_2(V)')'$ and $\gamma_0 = (\gamma_1, \gamma_2')'$.

Under (3), identification of the ASF holds if the parameter γ_0 is identified. In turn, the coefficients γ_0 is identified if the matrix $\mathbb{E}[p(D,V)p(D,V)']$ is non-singular. This full-column rank condition is weaker than the support invariance condition, and CFNSV and Newey and Stouli (2019) provided sufficient conditions for the non-singularity of $\mathbb{E}[p(D,V)p(D,V)']$. In Section 3.1,

I develop a flexible parametric estimation method based on the modelling approach discussed here.

2.3 Examples of excluded variables

Assumption 3 and the rank condition discussed in the previous section constitute the requirements for excluded variables. Intuitively, excluded variables need to satisfy (i) Z has no impact on the outcome conditional on (X^*, D) and (ii) Z is correlated with either X^* or D. An additional proxy variable satisfies these requirements since a noisy measurement is often assumed to be independent of other variables conditional on the unobserved, true measurement X^* and by definition, it should be correlated with X^*.

Another example of excluded variables is an IV for the treatment variable. Suppose $D = \varphi(Z, \eta)$ where φ is a non-stochastic function and η is some unobserved heterogeneity. By the exclusion restriction, $(Y(d), X^*, \eta) \perp\!\!\!\perp Z$, and for Assumption 2, we may impose $Y(d) \perp\!\!\!\perp \eta | X^*$. Then, $Y(d) \perp\!\!\!\perp (D, Z) | X^*$ follows, which implies the required conditional independence. The other requirement on excluded variables is implied by the IV relevance condition, and thus, IV is a valid excluded variable.

Repeated measurements and IVs are familiar objects, and they provide easy-to-understand examples of excluded variables. But, there are other candidates for excluded variables. Consider the wage example discussed in the beginning of Section 2. There, one may take educational attainment of worker's parents and/or household characteristics during worker's childhood as excluded variables. The exclusion restriction (i.e., $Y(d) \perp\!\!\!\perp Z | D, X^*$) is plausible as potential wage levels are unlikely to be affected by parents' education or early-stage family environments once you control for ability and own education as well as other observed characteristics. For the relevance condition (i.e., Z correlated with X^* and/or D), parents' education and household characteristics during childhood are likely to influence the probability of college attendance. Note that educational attainment of worker's parents and household characteristics during worker's childhood may not satisfy the IV exclusion restriction as they are potentially correlated with worker's unobserved ability through human capital formation. This example demonstrates that there exist empirically relevant excluded variables other than proxies and IVs.

2.4 Connection with existing results

Group-level correlated random effect models Theorem 1 has a close connection with the identification results of Altonji and Matzkin (2005) (AM henceforth). To explain, it is helpful to use the notation

$$Y(d) = \mathscr{Y}(d, \varepsilon), \qquad D \perp\!\!\!\perp \varepsilon | X^*$$

where $\mathscr{Y}$ is an unknown non-stochastic function and ε is an unobserved heterogeneity. AM based their identification results on finding pairs $(d_1, z_1), (d_2, z_2)$ such that

$$f_{\varepsilon|DZ}(\cdot|d_1, z_1) = f_{\varepsilon|DZ}(\cdot|d_2, z_2) \tag{4}$$

where $f_{\varepsilon|DZ}$ is the conditional density of ε given (D, Z) (see Equation (1.4) in AM). In proving Theorem 1, I show that (4) is implied by

$$f_{X|DZ}(\cdot|d_1, z_1) = f_{X|DZ}(\cdot|d_2, z_2).$$

Thus, my identification strategy uses the proxy distribution to find pairs $(d_1, z_1), (d_2, z_2)$ such that AM's exchangeability condition holds. In this sense, my paper provides a framework in which the exchangeability condition (4) follows from model primitives.

Arkhangelsky and Imbens (2019) also provided a framework that implies a version of exhcnageability condition, and they presented additional identification results. They focused on settings where observational units belong to groups and there exists a group-level unobserved heterogeneity. My model does not have an explicit group structure, and what my identification strategy does is to form groups based on the value of $V = \{f_{X|DZ}(x|D, Z) : x \in \mathcal{X}\}$. That is, two observations belong to the same group if they have the same value of V. Similar to the setup in Arkhangelsky and Imbens, the treatment assignment becomes exogenous within groups, and treatment effects can be identified using the group structure.

Non-classical measurement error models An alternative to my identification strategy is the operator diagonalization technique developed by Hu and Schennach (2008). Their method first identifies the joint distribution of (Y, D, X^*) using two proxies for X^* and then use the identified

distribution to compute the ASF (under Assumption 2). On one hand, the operator diagonalization technique identifies a larger class of parameters than my method does, including distributional features of unobserved heterogeneity. On the other hand, my method is less stringent on data requirements as I do not require a second measurement of X^*, although when available, I can take advantage of it as an excluded variable. On a related point, the operator diagonalization technique imposes the rank condition on both proxies while my method only needs one proxy to satisfy the rank condition. Another distinction is that my method handles measurement errors in covariates only, whereas the operator diagonalization technique applies to general measurement error problems. Since there is a trade-off between the strength of results and identifying restrictions, the two approaches are complementary.

2.5 Proof sketch of the conditional independence result

I sketch the identification argument using a simple setting. I focus on the case where D, X, X^* are all discrete. Specifically, the supports of X and X^* are $\mathcal{X} = \{x_1, \ldots, x_L\}$ and $\mathcal{X}^* = \{x_1^*, \ldots, x_L^*\}$ for some L. Note that in this special case, Assumption 5 reduces to the full-column rank of the matrix

$$\mathbf{\Pi} = \begin{bmatrix} \Pr[X = x_1 | X^* = x_1^*] & \ldots & \Pr[X = x_1 | X^* = x_L^*] \\ \vdots & \ddots & \\ \Pr[X = x_L | X^* = x_1^*] & \ldots & \Pr[X = x_L | X^* = x_L^*] \end{bmatrix}.$$

Define

$$V = \begin{bmatrix} \Pr[X = x_1 | D, Z] & \ldots & \Pr[X = x_L | D, Z] \end{bmatrix}'$$

which is the conditional distribution of the proxy variable X. Now I show that V is a valid control function in the sense that $Y(d) \perp\!\!\!\perp D | V$. To verify this claim, it suffices to show $X^* \perp\!\!\!\perp D | V$ since Assumptions 2 and 3 imply $\Pr[Y(d) \leq y | D, V] = \mathbb{E}[\Pr[Y(d) \leq y | X^*] | D, V]$ and $X^* \perp\!\!\!\perp D | V$ implies the desired result.

The law of total probabilities and Assumption 1 imply

$$\begin{aligned}\Pr[X = x|D, Z] &= \sum_{l=1}^{L} \Pr[X = x|X^* = x_l^*, D, Z]\Pr[X^* = x_l^*|D, Z] \\ &= \sum_{l=1}^{L} \Pr[X = x|X^* = x_l^*]\Pr[X^* = x_l^*|D, Z],\end{aligned}$$

and stacking this equation for different values of $x \in \{x_1, \ldots, x_L\}$,

$$V = \mathbf{\Pi} U, \qquad U = \big[\Pr[X^* = x_1^*|D, Z] \ \ldots \ \Pr[X^* = x_L^*|D, Z]\big]'$$

By the full-rank condition,

$$U = \big(\mathbf{\Pi}'\mathbf{\Pi}\big)^{-1}\mathbf{\Pi}'V. \tag{5}$$

This equality indicates that if two groups of workers have the same conditional distribution of test scores, then they also have the same conditional distribution of unobserved ability since $\mathbf{\Pi}$ is non-stochastic. This in turn implies that the conditional proxy distribution has the balancing property. To substantiate this last claim, for any $l \in \{1, \ldots, L\}$,

$$\begin{aligned}\Pr[X^* = x_l|D, V] &= \mathbb{E}\big[\Pr[X^* = x_l|D, Z]\big|D, V\big] \\ &= \mathbb{E}[e_l'U|D, V] \\ &= \mathbb{E}[e_l'\big(\mathbf{\Pi}'\mathbf{\Pi}\big)^{-1}\mathbf{\Pi}'V|D, V] \\ &= \mathbb{E}[e_l'\big(\mathbf{\Pi}'\mathbf{\Pi}\big)^{-1}\mathbf{\Pi}'V|V] \\ &= \mathbb{E}[\Pr[X^* = x_l|D, Z]|V] \\ &= \Pr[X^* = x_l|V]\end{aligned}$$

where $e_l \in \mathbb{R}^L$ is the unit vector whose lth element is unity, the first equality holds as V is a function of (D, Z), the third equality follows from (5), the fourth equality is by $\mathbf{\Pi}$ being non-random, and the fifth equality applies (5) again. The conclusion $\Pr[X^* = x_l|D, V] = \Pr[X^* = x_l|V]$ establishes the desired result $X^* \perp\!\!\!\perp D|V$.

3 Estimation

In applications, it is important to include other covariates that are free of measurement errors. Denote such variables by W. For convenience, I redefine $Z := (Z', W')'$, $X := (X', \text{vech}(XX')')'$, and $V := (\mathbb{E}[X|D, Z]', W')'$. Also, for a generic random vector A, write k_a for the dimension of A. In the estimation procedure, I assume $X \perp\!\!\!\perp (W, D, Z)|X^*$ and Theorem 1 holds conditional on W with appropriate modifications in the proof. If researchers are only willing to impose the weaker condition $X \perp\!\!\!\perp (D, Z)|W, X^*$, then the estimation procedure should treat W as part of X.

3.1 Flexible parametric approach

In this section, I consider a flexible parametric estimation procedure based on the discussion in Section 2.2. I assume that the outcome equation is specified by (3). For the control function V, let

$$\mathbb{E}[X|D, Z] = Q(D, Z)\delta_0,$$

where $Q : \mathcal{Z}\mathcal{D} \to \mathbb{R}^{k_x \times k_q}$ is a matrix-valued transformation of (D, Z) and $\delta_0 \in \mathbb{R}^{k_q}$ is the parameter to be estimated. As a baseline, one may use $Q(D, Z) = I \otimes q(D, Z)$ with I being the identity matrix, $q(D, Z) = (1, D', Z')$, and $\otimes$ denoting the Kronecker product. Researchers can include higher-order polynomial terms to enhance flexibility. Note that this is a reduced form equation, and as long as the model has good predictive power, the procedure is expected to work reasonably well.

For implementation, first estimate δ_0 by least squares and form $\widehat{V}_i = (Q(D_i, Z_i)\widehat{\delta}_n, W_i')'$. Then, estimate γ_0 in $\mathbb{E}[Y|D, V] = \psi(p(D, V)'\gamma_0)$ by (non-linear) regression of Y on $p(D, \widehat{V})$. Finally, the estimator for the average structural function is formed by

$$\widehat{\beta}_n(d) = \frac{1}{n}\sum_{i=1}^{n} \psi\big(p(d, \widehat{V}_i)'\widehat{\gamma}_n\big).$$

For inference, it is useful to have a closed-form variance estimator. Let $\dot{\psi}, \ddot{\psi}$ be the first and

derivative of ψ, respectively,

$$\widehat{\psi}_{ni} = \dot{\psi}(p(D_i, \widehat{V}_i)'\widehat{\gamma}_n), \quad \widehat{\Gamma}_2 = \frac{1}{n}\sum_{i=1}^{n} |\widehat{\psi}_{ni}|^2 p(D_i, \widehat{V}_i) p(D_i, \widehat{V}_i)',$$

$$\widehat{\varepsilon}_i = Y_i - \psi(p(D_i, \widehat{V}_i)'\widehat{\gamma}_n), \quad Q_i = Q(D_i, Z_i), \quad \widehat{\zeta}_i = X_i - Q_i\widehat{\delta}_n, \quad \widehat{\Gamma}_1 = \frac{1}{n}\sum_{i=1}^{n} Q_i'Q_i,$$

$$\widehat{\Gamma}_3 = \frac{1}{n}\sum_{i=1}^{n} \Big[\partial p(D_i, \widehat{V}_i) Q_i \widehat{\psi}_{ni}\widehat{\varepsilon}_i + \{\widehat{\varepsilon}_i \ddot{\psi}(p(D_i, \widehat{V}_i)'\widehat{\gamma}_n) - |\widehat{\psi}_{ni}|^2\} p(D_i, \widehat{V}_i)\widehat{\gamma}_n' \partial p(D_i, \widehat{V}_i) Q_i\Big],$$

$$\widehat{c}_1(d) = \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}(p(d, \widehat{V}_i)'\widehat{\gamma}_n) p(d, \widehat{V}_i)', \quad \widehat{c}_2(d) = \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}(p(d, \widehat{V}_i)'\widehat{\gamma}_n)\widehat{\gamma}_n' \partial p(d, \widehat{V}_i) Q_i,$$

and ∂p be the derivative of p with respect to the estimated elements of V. Then,

$$\frac{1}{n}\sum_{i=1}^{n} \left[\psi(p(d, \widehat{V}_i)'\widehat{\gamma}_n) - \widehat{\beta}_n(d) + \widehat{c}_1(d)\widehat{\Gamma}_2^{-1} p(D_i, \widehat{V}_i)\widehat{\psi}_{ni}\widehat{\varepsilon}_i + \{\widehat{c}_1(d)\widehat{\Gamma}_2^{-1}\widehat{\Gamma}_3 + \widehat{c}_2(d)\}\widehat{\Gamma}_1^{-1} Q_i'\widehat{\zeta}_i\right]^2$$

is an estimator for the asymptotic variance of $\sqrt{n}(\widehat{\beta}(d) - \beta(d))$. Note that the variance estimator does not require additional nuisance parameter estimation.

Since the asymptotic distributional theory for $\widehat{\beta}_n(d)$ is well-established (e.g., Newey and McFadden, 1994), I relegate the discussion of the asymptotic theory to the supplemental appendix. Under the assumptions stated there, the ASF estimator $\widehat{\beta}_n(d)$ is asymptotically normal and the variance estimator is consistent.

3.2 Nonparametric estimation

In this section, I consider a kernel-based nonparametric estimator for the ASF. Unlike the flexible parametric procedure, I do not use the specification in (3) and maintain the nonparametric specification of $\mathbb{E}[Y|D, V]$. I maintain Assumption 6 and use the location-scale family model of the conditional proxy distribution. The nonparametric estimator is useful when excluded variables have a large support so that the common support condition is plausible. Here, I focus on the discrete treatment variable.

In nonparametric estimation, small denominators may be problematic. To handle this issue, I introduce a trimming variable $T \geq 0$ and redefine the parameter of interest in relation to the

trimming:

$$\beta(d) = \mathbb{E}[Y(d)T]/\mathbb{E}[T].$$

For the first stage, let $L : \mathbb{R}^{k_z} \to \mathbb{R}$ be a kernel function and $q_1 \in \mathbb{Z}_{\geq 0}$ be the order of local polynomial regression. Then, for estimation of the control function,

$$\widehat{\delta}_{ln}(d,z) = \arg\min_{\delta} \sum_{i=1}^{n} \Big(X_{li} - r_{q_1}\Big(\frac{Z_i - z}{b_n}\Big)'\delta \Big)^2 \mathbb{1}\{D_i = d\} L\Big(\frac{Z_i - z}{b_n}\Big), \qquad l \in \{1, \ldots, k_x\}$$

where b_n is a sequence of vanishing bandwidths and r_{q_1} is an appropriately defined $\sum_{l=0}^{q_1} \binom{l+k_z-1}{k_z-1}$-dimensional vector. Specifically, for any $t \in \mathbb{Z}_{\geq 0}$ and $z \in \mathbb{R}^{k_z}$,

$$r_t(z) = \Big[\begin{array}{cccc} 1 & [z]^{1\prime} & \ldots & [z]^{t\prime} \end{array} \Big]', \qquad [z]^{\ell} = \Big[\begin{array}{cccc} z_1^{\ell} & z_1^{\ell-1} z_2 & \ldots & z_{k_z}^{\ell} \end{array} \Big]'.$$

Then, we take $\widehat{V} = (e_1'\widehat{\delta}_{1n}(D,Z), \ldots, e_1'\widehat{\delta}_{k_x n}(D,Z), W')'$ where e_1 is the vector whose first element is unity and remaining elements are zero. For the second-stage estimation, let $K : \mathbb{R}^{k_v} \to \mathbb{R}$ be another kernel, $q_2 \in \mathbb{Z}_{\geq 0}$ be the order of local polynomial regression, and h_n be a sequence of bandwidths. Then, define

$$\widehat{\gamma}_n(d,v) = \arg\min_{\gamma} \sum_{i=1}^{n} \Big(Y_i - r_{q_2}\Big(\frac{\widehat{V}_i - v}{h_n}\Big)'\gamma \Big)^2 \mathbb{1}\{D_i = d\} K\Big(\frac{\widehat{V}_i - v}{h_n}\Big)$$

and $\widehat{m}_n(d,v) = e_1'\widehat{\gamma}_n(d,v)$ to be an estimate of $\mathbb{E}[Y|D=d, V=v]$. Finally, the estimator of the ASF is formed by

$$\widehat{\beta}_n(d_0) = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_n(d_0, \widehat{V}_i) T_i \Big/ \frac{1}{n}\sum_{i=1}^{n} T_i$$

where $d_0 \in \mathcal{D}$ is the treatment level of interest.

To analyze the asymptotic properties of this estimator, I impose the following assumptions. The first set of conditions concerns properties of the kernel functions.

Assumption 7.

(i) The kernel function K is even and supported on $[-1,1]^{k_v}$. Also, it is differentiable and the derivatives are Lipschitz continuous.

(ii) The kernel function L is even, bounded, and supported on $[-1,1]^{k_z}$.

These restrictions on kernel functions are standard. Smoothness on K is required in order to handle generated regressors. The next set of assumptions imposes regularity conditions on the data generating process. To describe them, define

$$m_0(D,V)=\mathbb{E}[Y|D,V], \qquad \varepsilon = Y - m_0(D,V), \qquad \zeta = X - \mathbb{E}[X|D,Z], \qquad \rho(D,Z)=\mathbb{E}[\varepsilon|D,Z],$$

$$\tau(V)=\mathbb{E}[T|V], \qquad \tau(D,Z)=\mathbb{E}[T|D,Z], \qquad \pi_d=\mathbb{P}[D=d], \qquad \pi_d(V)=\frac{f_{V|D}(V|d)\pi_d}{f_V(V)}.$$

Assumption 8. *Let $\mathcal{V}_0$ be a compact subset of the support of V such that $\tau(v)$ vanishes outside $\mathcal{V}_0$. Also define $\mathcal{V}_0^\eta = \{v : \inf_{\tilde{v}\in\mathcal{V}_0} \|v-\tilde{v}\| \le \eta\}$ for some fixed $\eta > 0$ where $\|\cdot\|$ denotes the usual Euclidean norm. The following conditions hold for each $d \in \mathcal{D}$.*

(i) The observation $\{(Y_i, D_i, Z_i, X_i, T_i) : i = 1, \dots, n\}$ is a random sample.

(ii) The conditional distribution of V given $D = d$ has a bounded, continuous Lebesgue density, which is continuously differentiable in its first k_x arguments and bounded away from zero on $\mathcal{V}_0^\eta$. The conditional distribution of Z given $D = d$ has a Lebesgue density. The support is a bounded rectangle, and the density is continuous and bounded away from zero on its support.

(iii) The function $m_0(d, v)$ is (q_2+1)-times differentiable in v on $\mathcal{V}_0^\eta$ with bounded derivatives. Each element of $\mathbb{E}[X|D = d, Z = z]$ is $(q_1 + 1)$-times differentiable in z with bounded derivatives on the support of Z. $\rho(d, z)$ and $\tau(d, z)$ are continuous in z, $f_V(v)$ and $\tau(v)$ are continuously differentiable, and $\pi_d(v)$ is bounded away from zero on $\mathcal{V}_0^\eta$. Also, $\mathbb{E}[T] > 0$ and $\mathbb{P}[0 \le T \le C] = 1$ for some $C > 0$.

(iv) The regression errors ε, ζ satisfy $\mathbb{E}[|\varepsilon|^s] + \sup_v \mathbb{E}[|\varepsilon|^s|D = d, V = v]f_{V|D}(v|d) < \infty$, $\mathbb{E}[\|\zeta\|^s] + \sup_z \mathbb{E}[\|\zeta\|^s|D = d, Z = z]f_{Z|D}(z|d) < \infty$ for some $s \ge 4$.

The conditions are mostly standard in the multi-step semiparametric estimation literature. Condition (ii) imposes that the control function V has the Lebesgue density bounded away from zero on the region where the trimming variable is positive. Via this restriction, the trimming addresses the small denominator issue. Also, the assumption imposes that the support of Z is a rectangle and the density is bounded away from zero on the support. The following is the formal result on the asymptotic properties of the ASF estimator $\widehat{\beta}_n$.

Theorem 2. *Assumptions 7 and 8 hold,* $\frac{n^{2/s}\log n}{nb_n^{k_z/2}h_n^{k_v}} + nh_n^{2(q_2+1)} + nb_n^{2(q_1+1)} + \frac{(\log n)^2}{nh_n^{3\vee 2k_v}} = o(1)$, *and* $\frac{(\log n)^2}{nb_n^{2k_z}} + nb_n^{4(q_1+1)} = o(h_n^4)$. *Then,*

$$\sqrt{n}\big(\widehat{\beta}_n(d_0) - \beta(d_0)\big)\mathbb{E}[T] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{m_0(d_0,V_i)T_i - \beta(d_0)\mathbb{E}[T] + \frac{J_i(d_0)\tau(V_i)}{\pi_{d_0}(V_i)}\varepsilon_i + \delta_i(d_0)\zeta_i\Big\} + o_{\mathbb{P}}(1)$$

where $J_i(d) = \mathbb{1}\{D_i = d\}$,

$$\delta_i(d) = \frac{\rho(d,Z_i)J_i(d)}{\pi_d(V_i)^2}\big[\pi_d(V_i)\partial\tau(V_i) - \tau(V_i)\partial\pi_d(V_i)\big] + \Big[\tau(D_i,Z_i) - \frac{J_i(d)\tau(V_i)}{\pi_d(V_i)}\Big]\partial m_0(d,V_i),$$

and ∂ *denotes the differentiation operator with respect to the first* k_x *arguments in* V.

The theorem characterizes the form of the influence function for $\widehat{\beta}_n(d)$. To understand the result, split the influence function into two parts: $m_0(d,V)T - \beta(d)\mathbb{E}[T] + \frac{J(d)\tau(V)}{\pi_d(V)}\varepsilon$ and $\delta(d)\zeta$. The first part coincides with the influence function of the infeasible estimator that uses the true $\mathbb{E}[X|D,Z]$ instead of its estimated counterpart. Thus, the second part of the influence function captures the contribution of the first-stage estimation error to the asymptotic distribution. This "correction term" coincides with the formula derived by Hahn and Ridder (2013) who used Newey (1994)'s path-derivative method.[4] As Hahn and Ridder only provided high-level assumptions, Theorem 2 is a new result in the literature by providing one set of primitive sufficient conditions to characterize the asymptotic distribution of the three-step estimator.

The hypothesis of Theorem 2 includes conditions on the two bandwidth sequences. The second set of the assumptions (i.e., $(\log n)^2/nb_n^{2k_z} + nb_n^{4(q_1+1)} = o(h_n^4)$) ensures that $\max_{1\le i\le n}\|\widehat{V}_i - V_i\|^2 = o_{\mathbb{P}}(h_n^2\sqrt{n})$. Up to h_n^2, this is the standard "faster-than-$n^{1/4}$" rate restriction on preliminary non-parametric estimators. The presence of h_n^2 comes from Taylor expansion of the kernel function. In order to illustrate how the restrictions on the bandwidths affect the choice of local polynomial orders q_1 and q_2, suppose $k_x = 2$, $k_v = 2$, $k_z = 2$, and $s = 4$: one proxy, two excluded variables, no additional covariate W, and finite fourth moments of Y and X. Letting $h_n = O(n^{-c_1})$ and $b_n = O(n^{-c_2})$ for some $c_1, c_2 > 0$, the assumptions require $c_1 \in (\frac{1}{2q_2+2}, \frac{1}{4})$ and $c_2 \in (\frac{1}{2q_1+2}, \frac{1-4c_1}{4})$. From this, we see that the second-stage estimation (i.e., estimation of $\mathbb{E}[Y|D,V]$) requires quadratic

[4] See their Theorem 7. To be precise, one needs to modify their formula because D is also included in the first-stage estimation in my setting. With this change, the correction term $\delta(d)$ coincides with the one in Hahn and Ridder.

or higher-order local polynomial regression. This type of bias-reduction requirement is typical in multi-step semiparametric estimation problems: for instance, Powell et al. (1989) required the use of a higher-order kernel as soon as the dimension of covariate is greater than one. Also, the first-stage estimation requires local polynomial regression of order greater than 1.

Assumption 8 does not impose exponentially thin tails of the outcome variable, which would be necessary if I were to apply exponential bounds on tail probabilities from empirical process theory. Instead, I exploited the U-statistics structure directly, and consequently, I did not need to invoke asymptotic equicontinuity type arguments.

4 Empirical application

With the results developed in this paper, I estimate treatment effects of Catholic high school attendance on college enrollment using the National Longitudinal Study of 1972 dataset. Altonji et al. (2005) (AET henceforth) studied this question, and they concluded that, consistent with the analysis of the preceding studies, instruments used in the literature may fail to satisfy the identification conditions. I use an alternative set of identifying assumptions and provide a point estimate of the causal effect in a setting where methods based on IV may not be appropriate. Specifically, the key identifying assumption of my approach is (i) conditional on one-dimensional unobserved student's academic ability, attending a Catholic high school is exogenous with respect to the decision to attend college and (ii) math test score in the 12th grade is a good proxy for the unobserved academic ability.

I use the specification discussed in Section 3.1. Specifically,

$$\mathbb{E}[Y|D,W,V] = \Lambda(\beta_0 + \beta_1 D + \beta_2' W + \beta_3 V)$$

where Y is the indicator for enrollment in four-year college, D is the indicator of Catholic high school attendance, W is a vector of additional controls without measurement errors, $V = \mathbb{E}[X|D,Z]$, and Λ is the logistic CDF.[5] In this setting, X^* denotes unobserved academic ability, X denotes math test score, and Z is reading test score and categorical variables for distance to the closest Catholic

[5]For the multi-collinearity issue, I use only the conditional mean of X as the control function. My approach is still distinct from the classical 2SLS approach since the model is non-linear.

high school. AET and other studies examined whether distance to the closest Catholic high school could be used as an IV, but they concluded that the IV identification strategy may not be credible. In my approach, the distance variable does not need to satisfy the IV exclusion restriction as I only need the conditional independence condition to hold given the unobserved ability. For selecting covariates, I followed AET. Table 1 lists the variables used in analysis. For the proxy variable, I use math test score conducted in the 12th grade. I assume that conditional on other covariates, attending a Catholic high school affects the math test score only through the unobserved ability (see Figure 1).

Table 1 presents means and standard deviations for the entire sample, the control group, and the treated group. In the sample, about 7% of the students attended Catholic high schools, and we see that there exist some notable differences across the two groups in terms of observed characteristics. For instance, the treated group has lower fractions of racial minorities, more educated parents, and a lower probability of being in the low socio-economic status category. To assess the covariate distributions across the treated and control, I plotted histograms of the estimated propensity scores in Figures 2 and 3. From Figure 2, we see that there is a spike around zero for the propensity score in the control group. Figure 3 shows the histograms of the treated and control where I zoom in for the control group histogram. We see that the treated group does not have many observations near zero. Examining the propensity score distribution indicates that the overlap condition for ATE ($0 < \Pr[D = 1|V] < 1$) may be violated, and thus I focus on the average treatment effect on the treated (ATT).

For comparison, I estimate the treatment effect by two other specifications. The first specification is the standard logit where the identifying assumption is the exogeneity of Catholic high school attendance and regressors do not include test scores. AET and other studies showed that ignoring selection into Catholic schooling induces upward bias on the treatment effect estimate. The second approach uses math and reading test scores as additional controls. This specification may partially correct for selection bias, but it suffers from the measurement error problem.

In Table 2, I present the ATT estimates. Using my control function method, the "naïve" logit, and the logit with test scores, the estimated treatment effects are 7.9, 13.8, and 8.8 percentage point increase in the probability of four-year college enrollment, respectively. For the "naïve" estimate, the effect of 13.8 percentage points may be too large to be reasonable given that in the sample, 28%

of students were enrolled in four-year college. This may be due to positive selection into Catholic high schools. Consistent with this hypothesis, the estimation methods that control for unobserved ability provide lower estimates of ATT. Now, consider the estimate based on logit with test scores, which suffers from measurement error bias. The logit estimate is larger than the control function estimate by 10%, and testing the null hypothesis of the same probability limit of the two estimators, the p-value is 0.047 (one-sided test). Therefore, this empirical application provides a suggestive evidence that the proposed control function method corrects for bias arising from both selection and measurement errors.

5 Conclusion

I developed a new identification strategy for causal effects such as the average structural functions by exploiting proxy variables for unobserved confounding factors and excluded variables. This new approach does not require an IV for treatment variables and it is applicable with one proxy. As illustrated through an empirical application, my approach provides an useful alternative to existing methods. For implementation, I proposed nonparametric and flexible parametric estimation methods and established asymptotic distribution results for these estimators.

Acknowledgement This paper is based on a portion of my job market paper, which was circulated under the title "Identification and Estimation of Group-Level Partial Effects". I am grateful to my advisor Matias Cattaneo for advice and encouragement. I would like to thank Stephane Bonhomme, Sebastian Calonico, Max Farrell, Yingjie Feng, Andreas Hagemann, Michael Jansson, Lutz Kilian, Xinwei Ma, Eric Renault, Rocío Titiunik, Gonzalo Vazquez-Bare, seminar participants at Brandeis, Bristol, Chicago, Florida, LSE, NC State, Northwestern, Peking (Probability and Statistics), Pittsburgh, Rice, UC Irvine, UC San Diego, UPenn, and Warwick, and conference participants at Microeconometrics Class of 2019, Causal Learning with Interactions, IAAE Conference 2021, and RES Conference 2021 for helpful comments.

6 Bibliography

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *Journal of Human Resources*, 40, 791–821.

Altonji, J. G. and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102.

Andrews, D. W. K. (2017): "Examples of L^2-Complete and Boundedly-Complete Distributions," *Journal of Econometrics*, 199, 213–220.

Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85, 693–734.

Arkhangelsky, D. and G. W. Imbens (2019): "The Role of the Propensity Score in Fixed Effect Models," Working Paper.

Battistin, E. and A. Chesher (2014): "Treatment Effect Estimation with Covariate Measurement Error," *Journal of Econometrics*, 178, 707–715.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

Blundell, R. W. and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnsovsky, Cambridge University Press, vol. 2, chap. 8, 321–357.

——— (2004): "Endogeneity in Semiparametric Binary Response Models," *Reviews of Economic Studies*, 71, 655–679.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739.

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli, and F. Vella (2020): "Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models," *Quantitative Economics*, 11.

Deaner, B. (2021): "Proxy Controls and Panel Data," Working Paper.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206.

Hahn, J. and G. Ridder (2013): "Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Econometrica*, 81, 351–340.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y., S. M. Schennach, and J.-L. Shiu (2017): "Injectivity of a Class of Integral Operators with Compactly Supported Kernels," *Journal of Econometrics*, 200, 48–58.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations models without Additivity," *Econometrica*, 77, 1481–1512.

Keane, M. P., P. E. Todd, and K. I. Wolpin (2011): "The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications," Elsevier, vol. 4 of *Handbook of Labor Economics*, 331–461.

Matzkin, R. L. (2007): "Nonparametric identification," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5307 – 5368.

Miao, W., Z. Geng, and E. J. Tchetgen Tchetgen (2018): "Identifying Causal Effects with Proxy Variables of an Unmeasured Confounder," *Biometrika*, 105, 987–993.

Newey, W. and S. Stouli (2019): "Control Variables, Discrete Instruments, and Identification of Structural Functions," Forthcoming in Journal of Economietrics.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Elsevier, vol. 4, 2111 – 2245.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

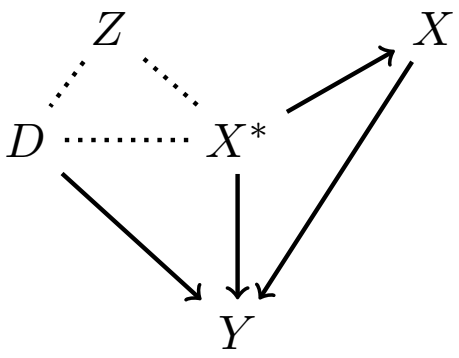
Sasaki, Y. (2015): "Heterogeneity and Selection in Dynamic Panel Data," *Journal of Econometrics*, 188, 236–249.

Schennach, S. M. (2020): "Mismeasured and Unobserved Variables," in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics*, 487–565.

Wilhelm, D. (2015): "Identification and Estimation of Nonparametric Panel Data Regression with Measurement Error," CEMMAP Working Paper.

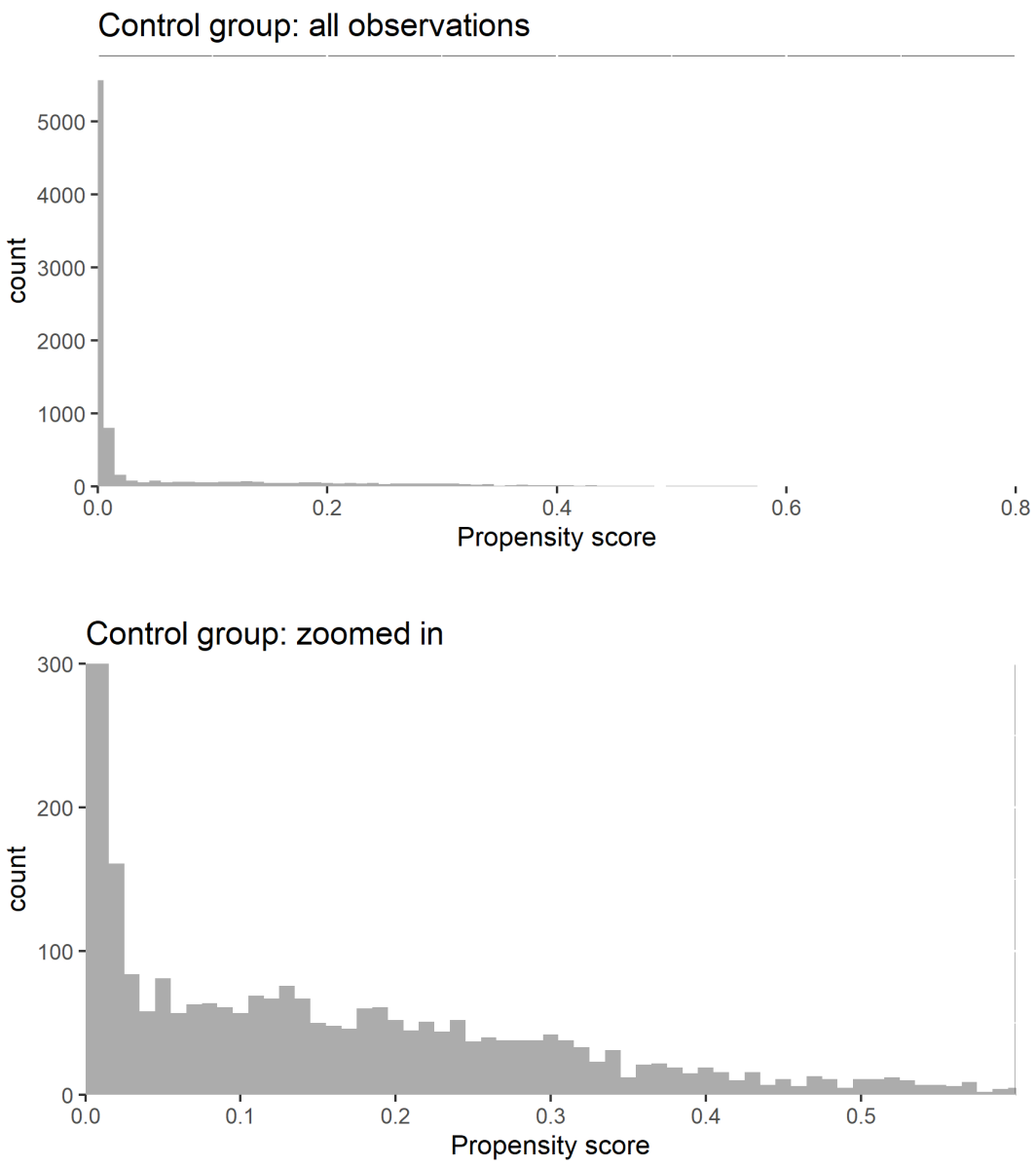
Wooldridge, J. M. (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445.

Figure 1: Independence Assumptions via DAG



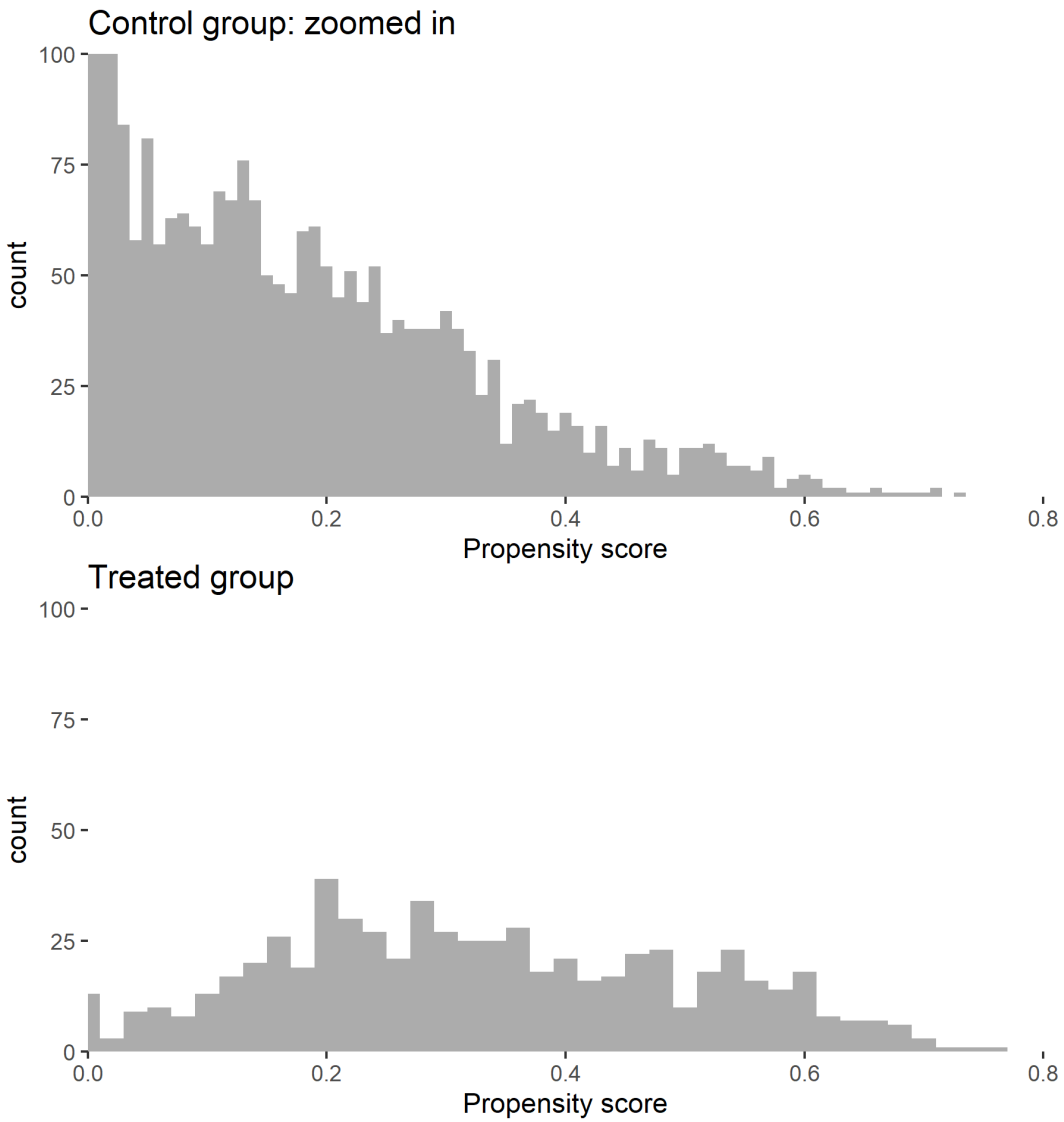
Notes. Arrows represent (potential) causal effects and dotted lines mean that two variables have a causal relationship with unspecified direction of effects.

Figure 2: Histograms of the Propensity Score for Control Group



Notes. The top plot is the histogram of the estimated propensity score for the control group and the bottom plot is the same graph with different vertical and horizontal ranges. The propensity score is estimated by logistic regression using variables in the covariates section on Table 1 and the estimated V variable.

Figure 3: Histograms of the Propensity Scores for Control and Treated Groups



Notes. The top plot is the histogram of the estimated propensity score for the control group and the bottom plot is the histogram for the treated group. The plot for the control group is zoomed in so that comparison is easier. The propensity score is estimated by logistic regression using variables in the covariates section on Table 1 and the estimated V variable.

Table 1: Summary statistics

	All: N=9142		Control: N=8498		Treated: N=644	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Treatment						
Catholic high school	0.07	(0.26)				
Outcome						
College enrollment	0.28	(0.45)	0.27	(0.44)	0.42	(0.49)
Covariates						
Female	0.50	(0.50)	0.50	(0.50)	0.47	(0.50)
Black	0.11	(0.31)	0.12	(0.32)	0.02	(0.14)
Hispanic	0.04	(0.19)	0.04	(0.20)	0.02	(0.12)
Father college degree	0.17	(0.38)	0.17	(0.38)	0.22	(0.41)
Mother college degree	0.11	(0.31)	0.10	(0.30)	0.14	(0.35)
Log family income	9.18	(0.74)	9.17	(0.74)	9.38	(0.58)
Father blue-collar work	0.30	(0.46)	0.30	(0.46)	0.27	(0.45)
SES low indicator	0.19	(0.39)	0.20	(0.40)	0.07	(0.26)
English at home	0.92	(0.28)	0.92	(0.28)	0.92	(0.27)
Newspaper at home	0.88	(0.33)	0.87	(0.33)	0.96	(0.19)
Mother works	0.58	(0.49)	0.58	(0.49)	0.55	(0.50)
Catholic	0.31	(0.46)	0.26	(0.44)	0.98	(0.15)
Urban	0.28	(0.45)	0.26	(0.44)	0.52	(0.50)
Suburban	0.23	(0.42)	0.23	(0.42)	0.26	(0.44)
Rural	0.19	(0.39)	0.21	(0.40)	0.02	(0.14)
Test scores						
Math	51.11	(9.86)	50.86	(9.90)	54.53	(8.54)
Reading	51.13	(9.81)	50.84	(9.85)	54.91	(8.44)
Dist. from Catholic HS						
Less than 1 mile	0.19	(0.39)	0.17	(0.38)	0.35	(0.48)
1-3 miles	0.19	(0.39)	0.18	(0.38)	0.32	(0.47)
3-6 miles	0.17	(0.37)	0.17	(0.37)	0.18	(0.38)
6-12 miles	0.11	(0.31)	0.11	(0.31)	0.06	(0.23)
12-20 miles	0.08	(0.27)	0.08	(0.27)	0.02	(0.14)

Notes. The table shows means and standard deviations for the entire sample, the control group, and the treated group. The sample size is 9142 for the entire sample, 8498 for the control, and 644 for the treated. The treatment variable is a binary variable that equals one if a student attended a Catholic high school. The outcome is a binary variable that equals one if a student was enrolled in a four-year college in 1973. For selecting covariates, I followed Altonji et al. (2005). Test scores were measured during the 12th grade. "Dist. from Catholic HS" denotes the distance to the closest Catholic high school based on the zip code information in the first follow-up survey. The information on Catholic high school locations was taken from Private School Universe Survey 1989-1990.

Table 2: ATT Estimates of Catholic Schooling on College Attendance

	Control function	logit (1)	logit (2)
ATT estimate	0.079	0.138	0.088
95% CI	[0.041 0.120]	[0.098 0.180]	[0.051 0.128]
Sample size		9142	
Mean outcome		0.280	

Notes. The table shows the average treatment effect on the treated (ATT) estimates of attending a Catholic high school on four-year college enrollment. The column "Control function" is based on the control function method developed in this paper. The column "logit (1)" is based on the logit regression using the treatment and "Covariates" variables in Table 1. The column "logit (2)" is based on the logistic regression using test scores as well as the regressors in "logit (1)". The row "95% CI" displays the 95% confidence interval, where quantiles were computed using the nonparametric bootstrap with 2000 iterations. The rows "Sample size" and "Mean outcome" are common across the three columns, and they show the sample size and the sample average of the outcome variable, respectively.