Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Correcting for Endogeneity in Models with Bunching

Carolina Caetano, Gregorio Caetano, and Eric Nielsen

2020-080

Please cite this paper as:
Caetano, Carolina, Gregorio Caetano, and Eric Nielsen (2020). "Correcting for Endogeneity in Models with Bunching," Finance and Economics Discussion Series 2020-080. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2020.080.

Correcting for Endogeneity in Models with Bunching[*]

Carolina Caetano
University of Georgia

Gregorio Caetano
University of Georgia

Eric Nielsen
Federal Reserve Board

August 2020

Abstract

We show that in models with endogeneity, bunching at the lower or upper boundary of the distribution of the treatment variable may be used to build a correction for endogeneity. We derive the asymptotic distribution of the parameters of the corrected model, provide an estimator of the standard errors, and prove the consistency of the bootstrap. An empirical application reveals that time spent watching television, corrected for endogeneity, has roughly no net effect on cognitive skills and a significant negative net effect on non-cognitive skills in children. Codes: C2, C21, C24

1 Introduction

When the treatment variable is constrained to be above (or below) a certain threshold, we often encounter bunching of observations at the threshold itself. In this paper, we show that this situation can be leveraged to build a correction for endogeneity. In some models, ranging from linear to some types of nonseparable, nonparametric structures, the correction consists of estimating a nuisance parameter and adding it to the model. In the linear case, for example, this translates to adding a generated control to the original regression. In the linear model and on some of the generalizations, the entire approach can be implemented with off-the-shelf, packaged software.

This type of bunching is often observed in variables constrained to be non-negative, as in the case of demand or inputs to production. Examples include behavioral variables like the consumption of vitamin supplements, cigarettes, alcohol, and coffee;[1] financial variables such as credit card debt, credit access, expenditure on ads, and bequests;[2] variables quantifying different uses of time such as exercising, working, doing homework, volunteering, and using

[*] We thank seminar participants at various conferences and institutions. The analysis and conclusions set forth here are those of the authors and do not indicate concurrence by other members of the research staff, the Board of Governors, or the Federal Reserve System.

[1] Meta-analyses of studies estimating the effects of these variables on health outcomes include Shinton and Beevers (1989); Fawzi et al. (1993); Corrao et al. (2000); Hernán et al. (2002); Reynolds et al. (2003); Noordzij et al. (2005); Bischoff-Ferrari et al. (2005); Oken et al. (2008); Richardson et al. (2013).

[2] See, e.g., Joulfaian and Wilhelm (1994); Peek et al. (2003); Ekici and Dunn (2010); Bertrand et al. (2010); Brown et al. (2010); Melzer (2011); Kim and Ruhm (2012); Carman (2013); Boserup et al. (2016); Erixson (2017); Elinder et al. (2018).

social media;[3] and even count data such as the number of children, the frequency of doctor visits, and the crime rate.[4]

More generally, one can apply the proposed method whenever Caetano (2015)'s test of exogeneity can be applied at the boundary. It is a natural solution whenever exogeneity is rejected by that test, or whenever the researcher wants to obtain point estimates that are robust under a selection-on-unobservables assumption. This test has been applied in a variety of settings in economics, political science, and finance.[5]

The approach of adding a generated covariate to account for endogeneity is well known, starting with the ubiquitous use of control functions, which require instrumental variables. Heckman (1979)'s correction is just as well known, and was developed for models with selection in the dependent variable. In contrast, our correction approach does not require instrumental variables, and there may not be a missing data problem.

This paper is related to the growing literature on the use of bunching points to infer causal parameters. Saez (2010) first leveraged bunching points in the context of the estimation of labor supply elasticities. Other seminal references in that field are Chetty et al. (2011) and Kleven and Waseem (2013). Kleven (2016) surveys this large literature. Some recent theoretical advancements include Blomquist et al. (2019) and Bertanha et al. (2020).

The asymptotic theory of our estimator uses the results in Chen et al. (2003) for extremum estimators with possibly infinite dimensional nuisance parameters. Several forms of stochastic equicontinuity conditions are required to establish that the objective functionals (squares of residuals using the estimated regressor) belong to Donsker classes, and for this we also use results in Pakes and Pollard (1989) and Andrews (1994).

The paper is laid out as follows. Section 2 proposes the correction strategy in the linear model and then discusses generalizations into nonlinear, semiparametric, and some types of nonparametric models. Section 3 establishes the asymptotic normality of the coefficients of the corrected model in the linear case, provides an estimator of the asymptotic variance, and proves the consistency of the bootstrap. Section 4 details several strategies for the identification and estimation of the correction term. Section 5 uses the correction approach to estimate the effect of time spent watching television (TV) on children's cognitive and non-cognitive skills. The application also showcases how the main identifying assumptions may be tested or argued. Another application of our correction method can be seen in Caetano et al. (2020). Section 6 concludes, and the proofs are gathered in the Appendix. An online

[3] See, e.g., Luoh and Herzog (2002); Baum II (2003); James-Burdumy (2005); Ruhm (2004, 2008); Eren and Henderson (2011); Chatterji et al. (2013); Ermisch and Francesconi (2013); Bhutani et al. (2013); Holt et al. (2013); Bettinger et al. (2014); Boulianne (2015).

[4] See, e.g., McDuffie et al. (1996); Black et al. (2005); Cohen (2008); Black et al. (2010).

[5] See, e.g., Rozenas et al. (2017), Erhardt (2017), Pang (2017), Bleemer (2018a), Bleemer (2018b), Ferreira et al. (2018), Lavetti and Schmutte (2018), Caetano and Maheshri (2018), De Vito et al. (2019), Caetano et al. (2019) and Caetano et al. (2020).

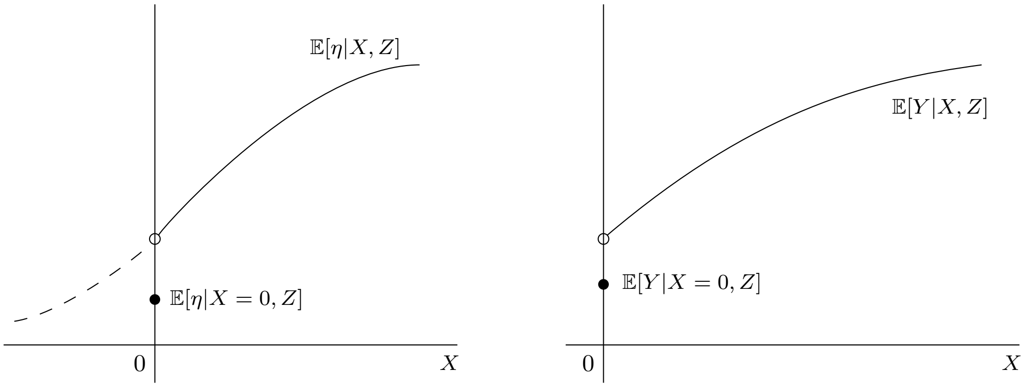
appendix discusses a real-data Monte Carlo study and additional implementation strategies.[6]

2 Correction Approach

The main idea behind our correction approach can be understood intuitively in the figure below. We are interested in estimating the effect of the treatment X on the outcome Y. Let X be constrained to be non-negative, and let the unobservable variable η be correlated with X conditional on controls Z. For concreteness, suppose that X represents a choice.

As depicted in the left panel of Figure 1, observations with similar, positive values of X tend to have similar η. However, those at the threshold point $X = 0$ are particularly selected. Because of the constraint on X, many chose the corner solution because they wanted to choose a negative amount but could not (dashed part of the left panel). Therefore, the expected η among those who chose $X = 0$ may be quite different from the expected η among those who chose a small but positive X.

Figure 1: Discontinuity in the Outcome Reflects Variation in Unobservables at the Boundary



(a) Unobservables

(b) Outcome

If η influences Y, then we have a problem of endogeneity. This means that, even holding Z constant, any variation in the outcome will reflect both changes in the treatment X and changes in the confounder η at the same time, as in the right panel of Figure 1. However, when we compare outcomes at $X = 0$ and X immediately above zero, the difference in X is negligible but the expected difference in η is large. Therefore, exactly at this location, the discontinuity in the outcome Y at $X = 0$ reflects changes in η without contamination from changes in X.

Caetano (2015)'s test looks at the discontinuity $\mathbb{E}[Y|X = 0, Z] - \lim_{x\downarrow 0} \mathbb{E}[Y|X = x, Z]$ to determine whether η is included in the outcome equation, and thus whether there is endogeneity. If we impose some structure on the model (e.g. linearity, partial linearity or some types of

[6] The online appendix is available at https://bit.ly/3gCigdZ.

nonseparable and nonparametric structures), this same discontinuity can be leveraged further to reveal information about the treatment effects. Section 2.1 formalizes the approach in the context of a linear model. Section 2.2 then discusses possible generalizations.

Technical note: in this paper, all equations and results involving random variables should be read as holding almost surely. $\mathbb{P}$ denotes the probability, and details about implied probability spaces and conditional sigma-algebras should be self-evident and are thus omitted. The expectation $\mathbb{E}$ is assumed to exist wherever written. The support of the distribution of any random variable R is denoted $\text{supp}(R)$. For brevity, we sometimes say "the support of R" to mean the support of the distribution of R. To keep identification arguments simple we will omit rank conditions when they are obvious.

2.1 Linear Model

Suppose that the structural equation is

$$Y = \beta X + Z'\gamma + \delta\eta + \varepsilon, \text{ where } \mathbb{E}[\varepsilon|X, Z, \eta] = 0. \tag{1}$$

X is a scalar variable and we are interested in its effect on Y, β. The vector of controls, Z, may include a constant term. We observe Y, X and Z, and we do not observe η and ε. Suppose that there exists a latent, unobservable variable X^* which depends on both Z and η:

$$X^* = Z'\pi + \eta. \tag{2}$$

The actual treatment, X, is equal to X^* subject to the (binding) constraint

$$X = \max\{0, X^*\}, \qquad \text{with} \qquad \mathbb{P}(X^* < 0) > 0. \tag{3}$$

Note that this is not a censored model in the typical sense. The outcome Y is a function of the actual treatment X, which is observed, not the latent variable X^*.

As suggested in the discussion of Figure 1, an intuitive framework in which this model can be understood is that of choice and utility maximization. Suppose for example that the utility is written as a function of X, Z and η, and that X is the choice which maximizes utility conditional on the constraint (3). X^* is the desired choice without this constraint. The model's key requirement is that some observations are at a "corner solution:" their desired choice in the unconstrained optimization would have been different from their actual choice in the constrained optimization. Note that this framework may help justify the model, but it is not necessary. The specific motivation for the three equations above is irrelevant for the validity of the approach.

Given equations (1), (2) and (3),

$$\mathbb{E}[Y|X, Z] = (\beta + \delta)X + Z'(\gamma - \pi\delta) + \delta\mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0). \tag{4}$$

If $\delta = 0$, X is exogenous, and thus β is identifiable as in the standard linear model. If $\delta \neq 0$, X is endogenous. Then, whenever $\mathbb{E}[X^*|X^* \leq 0, Z] < 0$ (which happens with positive probability because of (3)), the outcome will be discontinuous at $X = 0$.

Let us rewrite equation (4) as

$$\mathbb{E}[Y|X, Z] = \beta X + Z'(\gamma - \pi\delta) + \delta(X + \mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)). \tag{5}$$

If we can identify $\mathbb{E}[X^*|X^* \leq 0, Z]$ for all $z \in \text{supp}(Z|X = 0)$, then we can eliminate the endogeneity bias and identify β by adding the term $X + \mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)$ to the regression.

Correcting for endogeneity thus depends on the identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$. In essence, our approach transforms the problem of endogeneity into a problem of out-of-sample prediction. Because X^* is observed whenever $X^* > 0$, we can use the observed empirical distribution of $X^*|Z$ for $X^* > 0$ to predict this expectation. Although the out-of-sample nature of this prediction creates a great deal of difficulty, the fact that it is a prediction problem, rather than a causal identification problem, opens up a multitude of data-driven strategies that can be tailored to the particular empirical application. To allow for such flexibility, in Section 3, we provide the asymptotic distribution, variance estimator, and consistency of the bootstrap for any estimator of $\mathbb{E}[X^*|X^* \leq 0, Z]$ satisfying higher-level conditions. In Section 4, we propose several options for the identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$. Finally, in Section 5 we provide guidance about how to test the linearity assumption as well as the assumptions needed to identify the expectation.

Remark 2.1. *Can a correction be built without bunching? Consider the simplest alternative approach: specify $X = Z'\pi + \eta$, thereby removing any bunching structure from the model. In this case, $\mathbb{E}[Y|X, Z] = (\beta + \delta)X + Z'(\gamma - \pi\delta)$. It is impossible to separate β and δ in this equation. Even if we could somehow identify π (for example, by supposing the strong exclusion restriction that $\mathbb{E}[\eta|Z] = 0$), this would still be insufficient to identify δ, and thus β.*

It is possible to separate δ if there exists some form of nonlinearity in the relationship between X, Z and η. The simplest nonlinear specification is $X = g(Z) + \eta$. In this case, $\mathbb{E}[Y|X, Z] = (\beta + \delta)X + Z'\gamma - \delta g(Z)$. In order to identify δ in this equation, g must be nonlinear and identifiable. To identify g, we would need to suppose that $\mathbb{E}[\eta|Z] = 0$, or that there exists an instrument for Z in the first stage equation.

In our model, $\mathbb{E}[\eta|X, Z] = X - Z'\pi + \mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)$ is discontinuous, which allows us to identify δ without identifying the parameters governing the relationship between X and Z. Other sharp features such as bunching at interior points or kinks may also allow the construction of similar corrections without the need for requirements of independence of η and Z.

Remark 2.2. *The sign of δ is identified even if $\mathbb{E}[X^*|X^* \leq 0, Z]$ is not identified. To see this, note that if equations (1), (2) and (3) hold*

$$\mathbb{E}[\mathbb{E}[Y|X=0,Z] - \lim_{x\downarrow 0}\mathbb{E}[Y|X=x,Z]|X=0] = \delta\mathbb{E}[X^*|X^* \leq 0]. \tag{6}$$

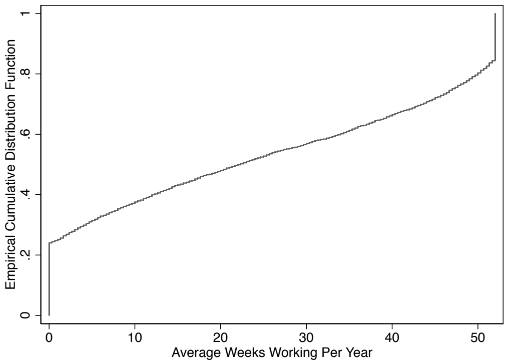
Since $\mathbb{E}[X^*|X^* \leq 0] < 0$ (by (3)), the sign of δ is opposite to the sign of the expected discontinuity. We show in Section 4 how knowing the sign of δ can be useful for partial identification. Note that this result also provides the basis for a test of exogeneity in the spirit of Caetano (2015), since $\delta\mathbb{E}[X^*|X^* \leq 0] = 0$ if and only if $\delta = 0$.

Implementation of this approach is simple, and we discuss it in Appendix A. In the online appendix (Section 2.1), we estimate the sign of δ in our application using this method. In particular, we strongly reject exogeneity.

Remark 2.3. *(More than one bunching point) Additional bunching points may be used in conjunction with our approach. This affords both a correction and a test of the underlying assumptions of the correction method in the same regression. To do this, calculate the correction using the bunching point at one end of the support and then apply Caetano (2015)'s exogeneity test on the other bunching point (or apply Caetano and Maheshri (2018)'s method if there are multiple additional bunching points).*

For example, consider the problem of estimating the effect of maternal labor supply on children's skills (e.g., James-Burdumy (2005)). Figure 2 shows the empirical c.d.f. of the treatment X = average number of weeks per year in which the mother worked in the three years following her child's birth. There are clearly two bunching points, one at $X = 0$, and another at $X = 52$.

Figure 2: Evidence of Bunching, Maternal Labor Supply



Note: This figure shows the empirical c.d.f. for the average number of weeks per year in which the mother worked in the three years following her child's birth. Source: National Longitudinal Study of Youth, 1979 cohort, sample of mothers whose children were born from 1979 to 2002.

A simple implementation of this test consists of running a regression of Y on X, Z, $X + \hat{\mathbb{E}}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)$, and $\mathbf{1}(X = 52)$. Then, the test of the exogeneity assumption *after the correction is applied is equivalent to a t-test of whether the coefficient of* $\mathbf{1}(X = 52)$ *is equal to zero. The asymptotic variance of the estimator of the coefficient of* $\mathbf{1}(X = 52)$*, the variance estimator, and the consistency of the bootstrap critical values are established in Section 3 (see footnote 8).*

2.2 Extensions

Linearity is not a fundamental requirement for this type of correction. We discuss several possible generalizations here, though the list of examples below is by no means exhaustive. Readers who are not interested in these extensions can skip directly to Section 3 without missing any notation or concept important to the understanding of the rest of the paper.

2.2.1 Linear Correlated Random Coefficients

Suppose that equations (1), (2) and (3) hold, but $\beta = \alpha_0 X + Z'\alpha$. This is equivalent to Garen (1984)'s model (see e.g. Chay and Greenstone (2005)), except that we do not exclude Z from the structural equation, and we do not require that Z and η are independent. Then,

$$\mathbb{E}[Y|X,Z] = \alpha_0 X^2 + XZ'\alpha + Z'(\gamma - \pi\delta) + \delta(X + \mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)).$$

If $\mathbb{E}[X^*|X^* \leq 0, Z]$ is identified, α_0 and α are identified, and thus we can identify the treatment effects. Our estimation results also cover this model (see footnote 8 in Section 3).

Other models $\beta = g_1(X, Z; \alpha_1)$, where the function g_1 is known up to the finite parameter vector α_1, may be identified analogously. If $\mathbb{E}[Y|X,Z]$ is linear in parameters, estimation is covered by the results in Section 3. Otherwise, this is a special case of the next model.

2.2.2 Nonlinear Correlated Random Effects

Suppose that the model is

$$Y = g_1(X, Z; \alpha_1)X + g_2(Z; \alpha_2)\eta + \varepsilon, \quad \mathbb{E}[\varepsilon|X, Z, \eta] = 0,$$

$$X^* = h_1(Z; \kappa_1) + h_2(Z; \kappa_2)\eta,$$

and equation (3), where g_1, g_2, h_1 and h_2 are functions which are known up to the finite parameter vectors $\alpha_1, \alpha_2, \kappa_1$ and κ_2, and $h_2(Z; \kappa_2) \neq 0$. We are interested in identifying g_1. Then,

$$\mathbb{E}[Y|X,Z] = g_1(X, Z; \alpha_1)X - \frac{g_2(Z; \alpha_2)h_1(Z; \kappa_1)}{h_2(Z; \kappa_2)} + \frac{g_2(Z; \alpha_2)}{h_2(Z; \kappa_2)}(X + \mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)).$$

This expression can usually be simplified a great deal when the functions are specified in an application. Identification in this model can be established with the method of moments.

All the combinations of parameters necessary for the identification of the treatment effects of X on Y are identified (see an explicit identification argument for a more general model in Section 2.2.4.) In fact, depending on the specific functional form of g_1, some (or all) of the elements of α_1 may be identified. Estimation can be done with nonlinear regression or generalized method of moments.

2.2.3 Partially Linear Model

The previous case requires that all the functions be specified. We now show that it is possible to build a correction in semiparametric models. Suppose that the model is

$$Y = \beta X + g(Z) + \delta\eta + \varepsilon, \quad \mathbb{E}[\varepsilon|X, Z, \eta] = 0,$$

$$X^* = h(Z) + \eta,$$

and equation (3), where g and h are not known. Then,

$$\mathbb{E}[Y|X, Z] = (\beta + \delta)X + (g(Z) - \delta h(Z)) + \delta\mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0).$$

We can follow Robinson (1988)'s strategy for partially linear models:

$$\begin{aligned} Y - \mathbb{E}[Y|Z] &= \beta(X - \mathbb{E}[X|Z]) \\ &\qquad + \delta\left(X - \mathbb{E}[X|Z] + \mathbb{E}[X^*|X^* \leq 0, Z][\mathbf{1}(X = 0) - \mathbb{P}(X = 0|Z)]\right) + \varepsilon. \end{aligned}$$

If the terms $X - \mathbb{E}[X|Z]$ and $\mathbb{E}[X^*|X^* \leq 0, Z][\mathbf{1}(X = 0) - \mathbb{P}(X = 0|Z)]$ are linearly independent, β and δ are identifiable. There are alternative identification methods for β in the equation above – see Härdle et al. (2000) for a detailed treatment of partially linear models. Estimation follows any of the methods available for partially linear models but substituting $\mathbb{E}[X^*|X^* \leq 0, Z]$ for its estimate.

2.2.4 A Nonparametric Nonseparable Model

The following case gives an example of a nonparametric, nonseparable structure in which a correction can be built. Suppose that

$$Y = g_1(X, Z, \varepsilon) + g_2(Z, \varepsilon)\eta, \quad \varepsilon \perp\!\!\!\perp X, Z, \eta$$

$$X^* = h(Z) + \eta$$

and equation (3), where g_1, g_2 and h are unknown functions. In this model, we are interested in the identification of the expected partial effect of X on Y everywhere, $\mathbb{E}\left[\frac{\partial g_1(X,Z,\varepsilon)}{\partial X}\middle|X, Z\right]$. It may be possible to weaken the structure further if the desired quantity is less ambitious, such as the average treatment effect $\mathbb{E}\left[\frac{\partial g_1(X,Z,\varepsilon)}{\partial X}\right]$ or the average structural function $a(x) = \mathbb{E}[g_1(x, Z, \varepsilon)]$. Given the model,

$$\mathbb{E}[Y|X, Z] = \mathbb{E}[g_1(X, Z, \varepsilon)] + \mathbb{E}[g_2(Z, \varepsilon)]X + \mathbb{E}[g_2(Z, \varepsilon)]\mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0).$$

Assuming the regularity conditions that guarantee the interchangeability of the order of the derivatives, expectations, and limits, then for $X > 0$,

$$\frac{\partial}{\partial X}\mathbb{E}[Y|X, Z] = \mathbb{E}\left[\frac{\partial}{\partial X}g_1(X, Z, \varepsilon)\middle| X, Z\right] + \mathbb{E}[g_2(Z, \varepsilon)]. \tag{7}$$

At the same time, note that

$$\mathbb{E}[Y|X = 0, Z] - \lim_{x\downarrow 0}\mathbb{E}[Y|X = x, Z] = \mathbb{E}[g_2(Z, \varepsilon)]\mathbb{E}[X^*|X^* \leq 0, Z]. \tag{8}$$

If $\mathbb{E}[X^*|X^* \leq 0, Z = z]$ is identifiable and strictly negative (that is, if $\mathbb{P}(X^* < 0|Z = z) > 0$), we can identify $\mathbb{E}[g_2(z, \varepsilon)]$ from equation (8), and thus we can identify the partial derivatives $\mathbb{E}\left[\frac{\partial}{\partial X}g_1(X, z, \varepsilon)\middle| X, Z = z\right]$ from equation (7).

Estimation in this example is much more involved than in the previous cases, requiring the nonparametric estimation of $\mathbb{E}[Y|X = 0, Z]$, the limit $\lim_{x\downarrow 0}\mathbb{E}[Y|X = x, Z]$ as well as the first derivatives of $\mathbb{E}[Y|X, Z]$ with respect to X.

2.2.5 Probit Model with Endogeneity

All of the previous cases admit simple correction strategies which are based on the separation of η and X in the structural equation. Building a correction without this separability is much harder, and will usually require the identification of other moments of $X^*|Z$ or perhaps of the entire distribution. We now give an example of how this may be done in a probit model with endogeneity.

Let the model be

$$Y = \mathbf{1}(\beta X + Z'\gamma + \delta\eta + \varepsilon \geq 0), \text{ with } \varepsilon|X, Z, \eta \sim \mathcal{N}(0, \sigma),$$

where equations (2) and (3) hold. Let Φ be the c.d.f. of the standard normal distribution. Then for $X > 0$,

$$\mathbb{P}(Y = 1|X, Z) = 1 - \Phi\left(-X(\beta + \delta)/\sigma - Z'(\gamma - \pi\delta)/\sigma\right),$$

which allows us to identify and estimate $(\beta + \delta)/\sigma$ and $(\gamma - \pi\delta)/\sigma$ by maximum likelihood using only observations with $X > 0$. At the same time, for $X = 0$,

$$\mathbb{P}(Y = 1|X = 0, Z) = 1 - \frac{1}{\mathbb{P}(X = 0|Z)}\int_{-\infty}^{0}\Phi\left(-Z'(\gamma - \pi\delta)/\sigma - x\delta/\sigma\right)\mathbb{P}(X^* \leq dx|Z).$$

If $\mathbb{P}(X^* \leq x|Z)$ is identified for all $x \leq 0$, the integral above may be calculated.[7] This allows us to identify $(\gamma - \pi\delta)/\sigma$ and δ/σ by maximum likelihood using only observations such that $X = 0$. Subtracting δ/σ from the previously identified $(\beta + \delta)/\sigma$ then allows us to identify β/σ.

Remark 2.4. *It is not possible to correct for endogeneity in all types of nonseparable models. Our approach allows us a glimpse of how η affects Y separately from the effect of X only at the bunching point. It is only at that location that we can guarantee that the treatment will not vary while the unobservables will. Consider the following nonparametric nonseparable model:*

$$Y = g(X, \eta) + \varepsilon, \quad \mathbb{E}[\varepsilon|X, \eta] = 0,$$

where g is not known. Here, no matter how restrictive is the equation that generates X^*, the bunching only allows us to learn something about $\partial g(0, \eta)/\partial\eta$. This is not sufficient information to allow us to learn anything about $g(X, \eta)$ for $X > 0$. Some form of regularity, be it some type of separability between η and X or a semiparametric structure, will be necessary for the identification of the treatment effects.

3 Asymptotic Theory for Estimation of the Corrected Model

We provide estimation results in the linear model.[8] Estimation follows equation (5) and consists of an OLS regression of Y onto X, Z and the estimated correction term, $X + \hat{\mathbb{E}}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)$. The coefficient of X is $\hat{\beta}$, and the coefficient of the correction term is $\hat{\delta}$.

We provide general asymptotic results that can be adapted to different estimators of $\mathbb{E}[X^*|X^* \leq 0, Z]$, including estimators not discussed in this paper, provided they are uniformly consistent at a minimum $n^{1/4}$ rate. Moreover, there is no impediment to using different identification and estimation methods to obtain $\hat{\mathbb{E}}[X^*|X^* \leq 0, Z]$ for different values of Z.

Assumption 1. *Denote $\psi_0 = \mathbb{E}[X^*|X^* \leq 0, Z = \cdot]$, $\hat{\psi} = \hat{\mathbb{E}}[X^*|X^* \leq 0, Z = \cdot]$, and $\dim(Z)$ equal to the number of elements in Z. Suppose that*

(i) The observations $\{(Y_i, X_i, Z_i')'\}_{i=1}^n$ are independent.

(ii) Denote $W = (X, Z', X + \psi_0(Z)\mathbf{1}(X = 0))'$, where W_i is an observation of this vector and W_{ij} is the j-th element of W_i. There exist constants $\alpha > 0$ and $\Delta < \infty$ such

[7] This may have to be done numerically. However, if $X^*|Z$ belongs to a very simple distributional family, such as uniform, or if instead of probit we have a more tractable model such as a logit, a closed form expression may be obtained.

[8] Results for some extensions are proven identically. For the case with more than one bunching point (Remark 2.3 in Section 2.1), just substitute the first coordinate in W in all assumptions and proofs by $(X, \mathbf{1}(X = \bar{x}))'$, where $\bar{x}$ is the second bunching point, and β by $(\beta, \alpha_{\bar{x}})'$, where $\alpha_{\bar{x}}$ is the coefficient of $\mathbf{1}(X = \bar{x})$. For the linear correlated random coefficients model (Section 2.2.1), simply substitute the first coordinate in W in all assumptions and proofs by $(X^2, XZ')'$, and β by $(\alpha_0, \alpha')'$.

that (a) $\mathbb{E}[|\varepsilon_i^2|^{1+\alpha}] < \Delta$, $\mathbb{E}[|\eta_i^2|^{1+\alpha}] < \Delta$, $\mathbb{E}[|W_{ij}W_{ik}|^{1+\alpha}] < \Delta$, $\mathbb{E}[|\eta_i^2 W_{ij}W_{ik}|^{1+\alpha}] < \Delta$ *and* $\mathbb{E}[|W_{ij}^2 W_{ik}W_{il}|^{1+\alpha}] < \Delta$ *for all* $i = 1, \ldots, n$ *and* $j, k, l = 1, \ldots, \dim(Z) + 2$*; (b)* $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[W_i W_i']$, $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_i^2 W_i W_i']$, *and* $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Var(\eta|X = 0, Z_i)W_i W_i']$ *are non-singular for* n *sufficiently large, with determinants bounded away from zero. Moreover, for each pair* $j, s = 1, \ldots, \dim(Z) + 2$*, either* $W_{ij}W_{is}$ *is constant for all* i*, or* $\frac{1}{n}\sum_{i=1}^n Var(W_{ij}W_{is}) \geq \alpha$ *for* n *sufficiently large.*

(iii) Let Θ *be a compact subset of* $\mathbb{R}^{\dim(Z)+2}$*, and* $\theta_0 := (\beta, (\gamma - \pi\delta)', \delta)' \in int(\Theta)$*.*

(iv) $\hat{\psi}, \psi_0 \in \mathcal{H}$*, a normed linear space with* $\int_0^\infty \sqrt{log N_{[]}(\varepsilon, \mathcal{H}, ||\cdot||_{\mathcal{H}})} d\varepsilon < \infty$ *(this implies that* $\mathcal{H}$ *is a* $\mathbb{P}$*-Donsker class).*

(v) Let $\mathcal{Z} = supp(Z)$*, then* $\sup_{z\in\mathcal{Z}} \left|\hat{\psi}(z) - \psi_0(z)]\right| = o_p(n^{-1/4})$*.*

(vi) $\sqrt{n}\mathbb{E}[(\hat{\psi}(Z_i) - \psi_0(Z_i))\mathbf{1}(X_i{=}0, X_j{=}0)W_i W_j'] \to_d \mathcal{N}(0, \Omega)$*.*

Theorem 3.1. *If equations* (1)*,* (2) *and* (3)*, and Assumption 1 hold, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d \mathcal{N}\left(0, \Sigma + \delta^2 \mathbb{E}[WW']^{-1}\Omega\mathbb{E}[WW']^{-1}\right),$$

where Σ *is the asymptotic covariance of a regression of* Y *onto* X*,* Z*, and* $X + \mathbb{E}[X^*|X^* \leq 0, Z]\mathbf{1}(X = 0)$ *(i.e., it is the asymptotic covariance if the true rather than the estimated expectation had been used in the regression).*

The proof of this theorem is in Appendix B.1. It applies Theorem 2 in Chen et al. (2003). Assumption 1(iv) requires that the uniform entropy integral (as defined in Van der Vaart and Wellner (1996), Chapter 2.1) is finite. This is used to establish the stochastic equicontinuity condition 2.5' in Chen et al. (2003) using Lemma 2.17 in Pakes and Pollard (1989). The same condition is also an indirect requirement of Assumption 2.6 in Chen et al. (2003) when we are allowing for an arbitrary estimator $\hat{\psi}(z)$. When a specific estimator is defined, Assumption 1(iv) often holds under more standard primitive conditions. In practice, this condition requires that the expectation and its estimator are well behaved functions. Many of the function classes adopted in economics satisfy this condition. For example, if the expectation assumes a parametric form, or if it is Lipschitz-continuous on the parameter vector, or smooth, or belongs to the Sobolev class, or is of bounded variation, the condition is satisfied.

Assumption 1(vi) is used to verify Assumption 2.6 in Chen et al. (2003). Note that the expectation is taken with respect to Z_i and X_i, conditional on the data that generated $\hat{\psi}$. Given Assumptions 1 (iv) and (v), Assumption 1(vi) may be substituted by $\frac{1}{\sqrt{n}}\sum_{i=1}^n(\hat{\psi}(Z_i) - \psi_0(Z_i))\mathbf{1}(X_i{=}0, X_j{=}0)W_i W_j' \to_d \mathcal{N}(0, \Omega)$ (this follows from Lemma 19.24 in Van der Vaart

(1998)). When the expectation estimator converges weakly at the $\sqrt{n}$ rate, Assumption 1(vi) does not need to be verified because it always holds (see Remark 3.1).

Next, we present an estimator of the asymptotic variance. Let $\hat{\mathbf{w}}$ be the matrix of regressors, with rows equal to $\hat{W}_i := (X_i, Z_i', X_i + \hat{\psi}(Z_i)\mathbf{1}(X_i = 0))'$. Let $\hat{\mathcal{D}} = Diag\{(Y_i - \hat{W}_i'\hat{\theta})^2\}_{i=1}^n$ be the diagonal matrix of the square of the residuals. Finally, let $\hat{\mathcal{V}}$ be a matrix with row i equal to $(\hat{C}_{i1}\mathbf{1}(X_1 = 0, X_i = 0), \ldots, \hat{C}_{in}\mathbf{1}(X_n = 0, X_i = 0))'$, where the $\hat{C}_{ij} = \hat{C}(Z_i, Z_j)$ are defined in Assumption 2 below. Then, an estimator of the asymptotic variance of $\hat{\theta}$ is

$$\hat{\mathbb{V}}_\theta = \left(\frac{\hat{\mathbf{w}}'\hat{\mathbf{w}}}{n}\right)^{-1}\left(\frac{\hat{\mathbf{w}}'\hat{\mathcal{D}}\hat{\mathbf{w}}}{n}\right)\left(\frac{\hat{\mathbf{w}}'\hat{\mathbf{w}}}{n}\right)^{-1} + \hat{\delta}^2\left(\frac{\hat{\mathbf{w}}'\hat{\mathbf{w}}}{n}\right)^{-1}\left(\frac{\hat{\mathbf{w}}'\hat{\mathcal{V}}\hat{\mathbf{w}}}{n^2}\right)\left(\frac{\hat{\mathbf{w}}'\hat{\mathbf{w}}}{n}\right)^{-1}.$$

Note that the first term is simply the Eicker-White covariance estimator in a regression of Y onto X, Z and $X + \hat{\psi}(Z)\mathbf{1}(X = 0)$. The second term is the penalty resulting from the fact that we are using an estimate instead of the true value $\psi_0(Z)$.

Assumption 2. *Suppose that*

(i) $\Omega = \mathbb{E}[C_{ij}\mathbf{1}(X_i\!=\!0, X_j\!=\!0)W_iW_j']$, *for some variables* $C_{ij} = C(Z_i, Z_j)$,

(ii) $\mathbb{E}[||C_{ij}W_iW_j'||]$, $\mathbb{E}[||C_{ij}W_i||]$ *and* $\mathbb{E}[|C_{ij}|]$ *are bounded,*

(iii) $\sup_{z,\tilde{z}\in\mathcal{Z}}\left|\hat{C}(z,\tilde{z}) - C(z,\tilde{z})\right| = o_p(1)$.

Theorem 3.2. *If equations* (1), (2), *and* (3), *and Assumptions 1 and 2 hold, then*

$$\hat{\mathbb{V}}_\theta \to_p \Sigma + \delta^2\mathbb{E}[WW']^{-1}\Omega\mathbb{E}[WW']^{-1}.$$

The proof of this theorem is in Appendix B.2. It uses a specific Strong Law of Large numbers for U-Statistics when the data is independent but not identically distributed. Letting $\hat{\mathbb{V}}_\beta$ be the first element in the matrix $\hat{\mathbb{V}}_\theta$, the standard error of $\hat{\beta}$ is $\sqrt{\hat{\mathbb{V}}_\beta/n}$. The first term in $\hat{\mathbb{V}}_\beta$ can be obtained directly from packaged software as the square of the Eicker-White standard errors of $\hat{\beta}$ on a regression of Y onto X, Z and $X + \hat{\psi}(Z)\mathbf{1}(X = 0)$. The second term in $\hat{\mathbb{V}}_\beta$ is the first element in the second matrix in $\hat{\mathbb{V}}_\theta$ divided by n.

When the expectation estimator converges weakly at the $\sqrt{n}$ rate, Assumption 2(i) does not need to be verified because it always holds (see Remark 3.1). Assumption 2(iii) requires consistent estimators of the C_{ij}, which are usually asymptotic covariances of the $\hat{\psi}(Z_i)$ for two different values of Z_i. Note that if the $\hat{\psi}(Z_i)$ are independent, then $C_{ij} = 0$ for all $i \neq j$, and thus the asymptotic variance is not affected by the fact that the expectation is estimated.

The following remarks discuss simplifications of the previous theorems for two important special cases.

Remark 3.1. *($\sqrt{n}$-rate of convergence) Suppose that $\hat{\psi}$ converges at the parametric rate to a Brownian Bridge, so that for all $z \in \mathcal{Z}$, there exists a normal random variable χ_z such that $\sqrt{n}(\hat{\psi}(z) - \psi_0(z)) \to_d \chi_z$. Then, Assumptions 1(vi) and 2(i) always hold, and $C_{ij} = Cov(\chi_{Z_i}, \chi_{Z_j})$. The proof of this statement follows from the Functional Delta Method (Theorem 3.9.5 in Van der Vaart and Wellner (1996)) and is presented in Appendix B.3.*

Remark 3.2. *(Z with finite support) If $supp(Z) = \{z_1, \ldots, z_L\}$, and if $\hat{\psi}(z_l)$ uses only observations such that $Z_i = z_l$, then the assumptions of Theorems 3.1 and 3.2 may be simplified. Specifically, replace Assumptions 1 (iii)-(vi) with*

(iii') Define $p_{0,l} = \mathbb{P}(X = 0, Z = z_l)$. Then $p_{0,l} > 0$ for at least one $l = 1, \ldots, L$.

(iv') $\sqrt{n}(\hat{\psi}(z_l) - \psi_0(z_l)) \to_d \mathcal{N}(0, V_{z_l})$ for all $l = 1, \ldots, L$ such that $p_{0,l} > 0$.

Define $W_{0,z_l} = (0, z_l', \psi_0(z_l))'$. The asymptotic variance can be written as

$$\Sigma + \delta^2 \mathbb{E}[WW']^{-1} \left(\sum_{l=1}^{L} p_{0,l}^2 V_{z_l} W_{0,z_l} W_{0,z_l}' \right) \mathbb{E}[WW']^{-1}.$$

The term Σ can be estimated as in the general case, and the second term can be estimated by replacing δ with $\hat{\delta}$, V_{z_l} with an estimator $\hat{V}_{z_l}$ and the other terms with sample equivalents. To establish the consistency of this variance estimator, Assumption 2 may be replaced with $\hat{V}_{z_l} \to_p V_{z_l}$ for all l such that $p_{0,l} > 0$.

Now we establish that the ordinary nonparametric bootstrap can consistently estimate the distribution of $\sqrt{n}(\theta - \theta_0)$ for an $i.i.d.$ sample for a wide class of estimators $\hat{\psi}(Z)$.

Assumption 3. *Suppose that*

(i) The observations $\{(Y_i, X_i, Z_i')'\}_{i=1}^n$ are i.i.d.

(ii) $\mathcal{H}$ has a bounded envelope function.

(iii) $n^{1/4} \sup_{z \in \mathcal{Z}} \left| \hat{\psi}(z) - \psi_0(z) \right| = o_{a.s.}(1)$

(iv) $\frac{1}{\sqrt{n}} \sum_{i=1}^n (\hat{\psi}(Z_i) - \psi_0(Z_i))^r \mathbf{1}(X_i = 0) R_i = o_{a.s.}(1)$, for $r = 1$ and $R_i = \varepsilon_i, \eta_i$, and W_i; and for $r = 2$ and $R_i = 1$.[9]

(v) Denote the bootstrap sample quantities with a "b," let $\hat{\psi}^b = \hat{\mathbb{E}}^b[X^* | X^* \leq 0, Z = \cdot]$, and let $o_{p^b}(\cdot)$ and $O_{p^b}(\cdot)$ denote the $o_p(\cdot)$ and $O_p(\cdot)$ notation for the $\mathbb{P}^b$-probability. Then,

(a) $n^{1/4} \sup_{z \in \mathcal{Z}} \left| \hat{\psi}^b(z) - \hat{\psi}(z) \right| = o_{p^b}(1)$.

(b) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\psi}^b(Z_i) - \hat{\psi}(Z_i))^r \mathbf{1}(X_i = 0)R_i = o_{p^b}(1)$, *for* $r = 1$ *and* $R_i = \varepsilon_i, \eta_i$, *and* W_i; *and for* $r = 2$ *and* $R_i = 1$.[9]

(c) $\sqrt{n}\mathbb{E}\big[[(\hat{\psi}^b(Z) - \hat{\psi}(Z)) - (\hat{\psi}(Z) - \psi_0(Z))]\mathbf{1}(X = 0)R\big] = o_{p^b}(1)$, *for* $R = Z$ *and* $R = \psi_0(Z)$.

Theorem 3.3. *If equations* (1), (2) *and* (3), *and Assumptions 1 and 3 hold, then*

$$\sqrt{n}(\hat{\theta}^b - \hat{\theta}) \to_d \mathcal{N}(0, \Sigma + \delta^2\mathbb{E}[WW']^{-1}\Omega\mathbb{E}[WW']^{-1}) \text{ in } \mathbb{P}^b\text{-probability.}$$

The proof of this theorem is in Appendix B.4, and follows from Theorems 3 and B in Chen et al. (2003). Theorem B makes a requirement of almost sure stochastic equicontinuity (Assumption 2.5' a.s. in that paper). Its direct translation to our context is expressed in footnote 9, which is a difficult condition to establish.[10] Instead, we bypass the need for almost sure stochastic equicontinuity and prove that, in our context and given the other assumptions of Theorem 3.3, it may be substituted by two weaker conditions, Assumptions 3 (iv) and (vb).

4 Identifying $\mathbb{E}[X^*|X^* \leq 0, Z]$

The identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$ is a prediction exercise. In this sense it is a simpler problem than causal identification. The difficulty is that the prediction is out-of-sample, as we only observe the distribution of X^* for $X^* > 0$ and the bunching at $X = 0$. Nevertheless, there is a lot of information which, together with assumptions of varying degrees of generality, can be leveraged into partial or point identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$, and therefore of β.

In this section, we focus on providing guidance to practitioners and exposing the reader to many options that are available. We formalize much of the analysis below, but a comprehensive treatment of the material in this section is beyond the scope of this paper.

The following sections discuss three avenues of investigation into the identification of the expectation. In Section 4.1, we examine opportunities for partial identification. In Section 4.2, we discuss identification inside of parametric classes of distributions, where the parameters

[9] Assumptions 3 (iv) and (vb) may be substituted by the almost sure stochastic equicontinuity condition: for all positive sequences $\tau_n = o(1)$ (it's sufficient to prove this for decreasing sequences $\delta_n = o(n^{1/2-\alpha})$ for some $\alpha > 0$),

$$\sup_{||\psi-\psi_0||_{\mathcal{H}}\leq\tau_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(\psi(Z_i) - \psi_0(Z_i))V_i - \mathbb{E}[(\psi(Z_i) - \psi_0(Z_i))V_i]]\right| = o_{a.s.}(1).$$

[10] Pötscher and Prucha (1994) and Jenish and Prucha (2009) discuss almost sure stochastic equicontinuity for establishing Uniform Laws of Large Numbers. The primitives explored in those papers are not sufficient to establish almost sure stochastic equicontinuity with a $\sqrt{n}$ denominator as required by Chen et al. (2003). We did not find other references of primitives of almost sure stochastic equicontinuity, and a general treatment of this condition in the lines of Pakes and Pollard (1989) is an open question.

may be specified parametrically or nonparametrically. Finally, in Section 4.3, we discuss how to discretize Z and how to leverage the resulting discretized controls to improve both the identification and the estimation of the expectation.

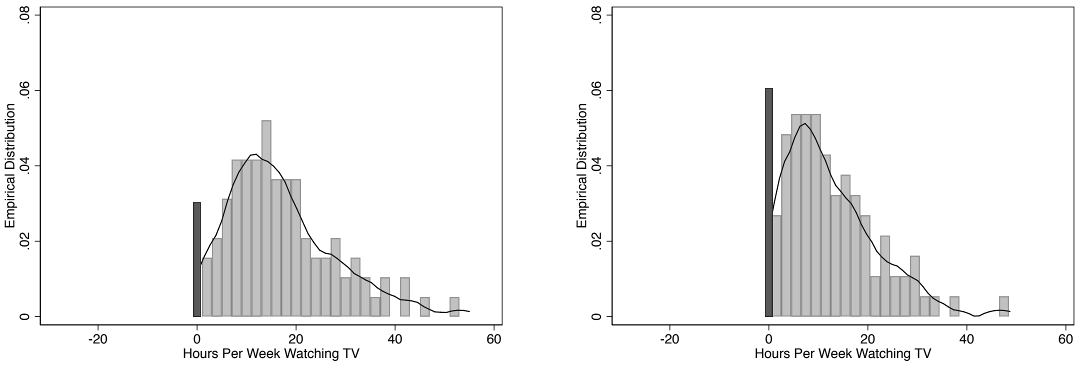
Note that to predict $\mathbb{E}[X^*|X^* \leq 0, Z]$, we only need data on X and Z. A researcher may wish to predict the expectation using an entirely different dataset. There is no impediment to doing so, as the theorems in Section 3 do not make any restrictions on the data used to estimate the expectation.[11]

For simplicity, we use the notation $F_{R|Z}(r)$ to denote $\mathbb{P}(R \leq r|Z)$, for $R = X, X^*$ and η. $F_{R|Z}^{-1}(q)$ denotes the quantile q of $F_{R|Z}$. When they exist, $f_{R|Z}$ is the density of $F_{R|Z}$, and $f'_{R|Z}$ is the derivative of $f_{R|Z}$.

4.1 Partial Identification

Consider Figure 3, which shows the distribution of the treatment variable X in our application (hours per week watching TV) for two different values of the controls. We leave details about the data to Section 5. For now, we focus on the shape of the distribution. The main pieces of information we have are the amount of bunching at zero and the height and slope of the density as it reaches zero from the positive side.

Figure 3: Conditional Empirical Distributions of X



Note: Each panel restricts the sample to a different value of the controls. The kernel density plot and histogram of X in the positive side are shown. The darker bar represents the proportion of observations at $X = 0$. Both the kernel bandwidth and the histogram bin width are equal to 2.

We begin by establishing an upper bound on $\mathbb{E}[X^*|X^* \leq 0, Z]$ that is fairly agnostic about the shape of the distribution of $X^*|Z$ below zero.

[11] In fact, in some applications, other datasets may allow the observation of the entire distribution of X^*, thus enabling the direct identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$. For example, minimum schooling laws, minimum wages, and minimum working age change over time and by state, and 401K minimum contributions are mandatory in some jobs and not in others, which may provide an opportunity to identify $\mathbb{E}[X^*|X^* \leq 0, Z]$ from similar observations that are unconstrained.

Proposition 4.1. *Suppose that the right derivative of $F_{X|Z}(x)$ at zero exists, and denote it $f_{X|Z}(0)_+$. Suppose also that in $(-\infty, 0)$, $f_{X^*|Z}(x)$ exists and $f_{X^*|Z}(x) \leq f_{X|Z}(0)_+$. Then,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] \leq -F_{X|Z}(0)^2/f_{X|Z}(0)_+. \tag{9}$$

This proposition states that, if the density for $X^* < 0$ is no higher than the density at $X^* = 0$, then the uniform density which integrates to $F_{X|Z}(0)$ in the negative side, $f_{X|Z}(0)_+\mathbf{1}(x \geq -F_{X|Z}(0)/f_{X|Z}(0)_+)$ in $(-\infty, 0]$, yields the highest possible value of the expectation. The proposition allows $f_{X^*|Z}(x)$ to be discontinuous anywhere, including at zero, and to not be monotonic. Its proof is a trivial application of the lemma in Appendix B.5.

Substituting an estimate of the upper bound in equation (9) on the main regression (equation (5)) yields an estimator of an asymptotic upper bound for β if $\delta > 0$ and of an asymptotic lower bound for β if $\delta < 0$. Since the sign of δ can be identified (Remark 2.2), this bound may be used strategically to obtain conservative conclusions.

Figure 3 reveals a clear bell shape on the positive side. Note that in the left panel of Figure 3, where bunching is smaller, there is even some evidence of an inflection point in the left tail of the empirical distribution of $X^*|Z$. In order to obtain more precise information about β, the temptation is to assume a normal, logistic, or at the very least a symmetric structure. The concern with such an approach is that there may be a steep descent to zero on the side we do not observe, as in the case of a beta or a gamma distribution.

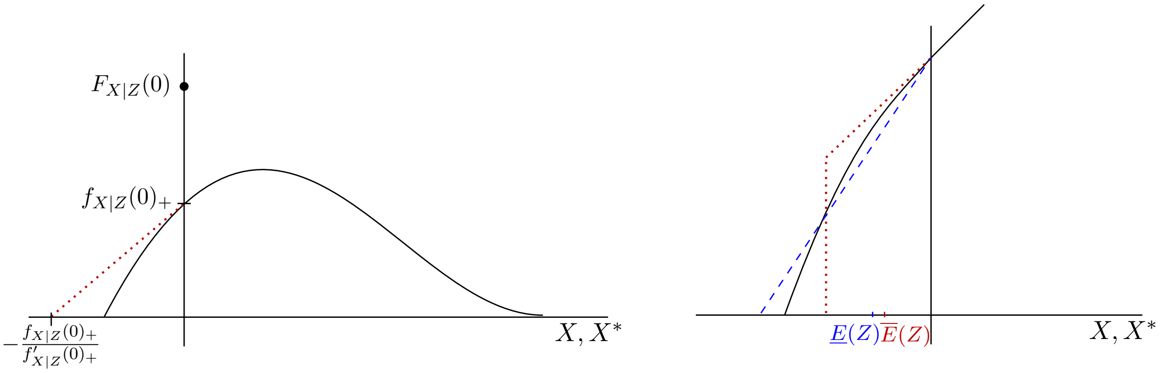
A visual inspection of Figure 3 suggests that the probability of bunching is large enough to rule out a steep descent in both cases. The following proposition offers a condition which allows us to test the concavity of the density on the negative side, which is a common manifestation in distributions with steep descent.

Proposition 4.2. *Suppose that $f_{X^*|Z}(x)$ exists in $(-\infty, 0]$, and is differentiable at $x = 0$. Denote by $f'_{X|Z}(0)_+$ the right derivative of $f_{X|Z}(x)$ at zero. Then $f'_{X^*|Z}(0) = f'_{X|Z}(0)_+$ is identifiable. Moreover, if $f'_{X^*|Z}(0) > 0$ and $f_{X^*|Z}(x)$ is concave in supp$(X^*|X^* < 0, Z)$, then*

$$f_{X|Z}(0)_+^2 - 2f'_{X|Z}(0)_+ \cdot F_{X|Z}(0) \geq 0. \tag{10}$$

This proposition is a corollary of Proposition 4.3. The rationale behind equation (10) can be understood in the left panel in Figure 4. The dotted red line is $f_{X|Z}(0)_+ + f'_{X|Z}(0)_+ \cdot x$. If $f_{X^*|Z}(x)$ is concave in $(-\infty, 0]$, the area below the dotted line must not be smaller than $F_{X|Z}(0)$. Confirming the visual inspection, our estimation of the quantities above rules out concavity in both panels of Figure 3.

Figure 4: Concave Distribution of $X^*|Z$ below zero



(a) Testing Concavity

(b) Bounds Under Concavity

Note: The left panel shows the distribution of $X^*|Z$. On the positive side, it is equal to the distribution of $X|Z$. The right panel shows a zoomed-in version around the negative side of the distribution of $X^*|Z$.

Under concavity, it is possible to derive upper and lower bounds of the expectation.

Proposition 4.3. *Suppose that $f_{X^*|Z}(x)$ exists in $(-\infty, 0]$, and is differentiable at $x = 0$, $f_{X^*|Z}(0) > 0$ and $f'_{X|Z}(0)_+$ is defined as in Proposition 4.2. If, moreover, $f_{X^*|Z}(x)$ is concave in supp$(X^*|X^* < 0, Z)$, then*

$$\underline{E}(Z) \leq \mathbb{E}[X^*|X^* \leq 0, Z] \leq \overline{E}(Z), \text{ where}$$

$$\underline{E}(Z) = -2a/3 \text{ and } \overline{E}(Z) = \begin{cases} -a/2, \text{ if } f'_{X|Z}(0)_+ = 0 \\ -a/3(b(3-b) + b^{1/2}(b-2)^{3/2}), \text{ if } f'_{X|Z}(0)_+ \neq 0, \end{cases}$$

$$\text{where } a = \frac{F_{X|Z}(0)}{f_{X|Z}(0)_+} \text{ and } b = \frac{f_{X|Z}(0)_+^2}{F_{X|Z}(0) f'_{X|Z}(0)_+}.$$

The bounds can be seen in the right panel of Figure 4. Under concavity, the most negative value of the expectation, $\underline{E}(Z)$, is the one obtained under the linear density which integrates to $F_{X|Z}(0)$, illustrated by the dashed blue line.[12] The least negative value of the expectation, $\overline{E}(Z)$, is obtained under the density with derivative $f'_{X|Z}(0)_+$ which integrates to $F_{X|Z}(0)$, illustrated by the red dotted line.[13] Note that the bounds also hold when $f'_{X|Z}(0)_+ < 0$. The proof of this proposition follows from the lemma in Appendix B.5, since the density's concavity and continuity at 0 implies the necessary stochastic dominance relationships between those distributions.

[12] This density is equal to $[f_{X|Z}(0)_+ + (f_{X|Z}(0)_+^2/2F_{X|Z}(0)) \cdot x]\mathbf{1}(x \geq -2a)$ in $(-\infty, 0]$.

[13] If $f'_{X|Z}(0)_+ = 0$, this density is equal to $f_{X|Z}(0)_+\mathbf{1}(x \geq -a)$. Otherwise, this density is

$$[f_{X|Z}(0)_+ + f'_{X|Z}(0)_+ \cdot x]\mathbf{1}(x \geq -a(b - \sqrt{b(b-2)})) \text{ in } (-\infty, 0].$$

Substituting an estimate of the upper bound in equation (9) on the main regression (equation (5)), and then repeating the process using the lower bound instead yields estimates of the implied bounds on β.

Similar arguments allow us to determine that if $f_{X^*|Z}(x)$ is convex in $\text{supp}(X^*|X^* < 0, Z)$, then: (a) if $f'_{X|Z}(0)_+ > 0$, then $\underline{E}(Z)$ is an upper bound of $\mathbb{E}[X^*|X^* \leq 0, Z]$. Thus, substituting an estimator of $\underline{E}(Z)$ in the regression yields a $\hat{\beta}$ estimator which is an asymptotic lower bound of β if $\delta > 0$, and an asymptotic upper bound on β if $\delta < 0$. (b) If $f'_{X|Z}(0)_+ \leq 0$, then $\overline{E}(Z)$ is a lower bound of $\mathbb{E}[X^*|X^* \leq 0, Z]$. Thus, substituting an estimator of $\overline{E}(Z)$ in the regression yields the opposite conclusions as (a).

We discuss how the bounds in this section may be estimated in Appendix A.

4.2 Identification Through Families of Distributions

To point-identify $\mathbb{E}[X^*|X^* \leq 0, Z]$, a natural approach is to suppose that X^* belongs to some parametric family of distributions. It is possible to do this in varying degrees of flexibility, as we show below. To help clarify how the assumptions made to identify $\mathbb{E}[X^*|X^* \leq 0, Z]$ impact the original model, all assumptions henceforth are written with respect to η. Importantly, these assumptions are testable (see Remark 4.1).

4.2.1 Parametric Methods

Model 4.2.1. *(Tobit) $\eta|Z \sim \mathcal{N}(Z'\mu, \sigma^2)$. In this case,*

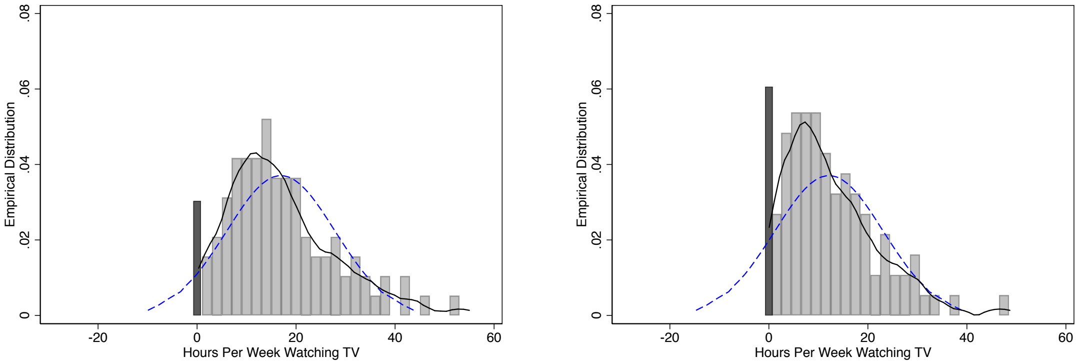
$$\mathbb{E}[X^*|X^* \leq 0, Z] = Z'(\pi + \mu) - \sigma\lambda(-Z'(\pi + \mu)/\sigma),$$

where $\lambda(\cdot)$ is the inverse Mill's ratio. Note that $X^*|Z \sim \mathcal{N}(Z'(\pi + \mu), \sigma^2)$ together with equation (3) satisfy the conditions of the Tobit model. Thus, both $\pi + \mu$ and σ can be identified as in Tobin (1958), and therefore the expectation is also identified. We never identify π and μ separately, nor is doing so necessary.

Appendix A shows how this model may be estimated. The Tobit model assumes homoskedasticity. Turning to the application in Section 5, the dashed blue line in Figure 5 demonstrates the homoskedastic normal fit for two different values of the controls. These panels demonstrate that homoskedasticity is clearly not a good assumption in this application. The variance of the treatment variable X (TV time) for the left panel is clearly higher than the corresponding variance for the right panel, yet the fitted normal distribution (dashed blue curve) does not reflect this.

In the online appendix, we show how to identify $\mathbb{E}[X^*|X^* \leq 0, Z]$ in the logistic, exponential and uniform distribution families by maximum likelihood. Other distribution families can

Figure 5: Homoskedastic Tobit Fit



Note: This figure adds to Figure 3 the fitted distribution from the homoskedastic Tobit model (Model 4.2.1), shown as the dashed blue curve.

be identified similarly. The homoskedasticity in Model 4.2.1 can also be relaxed by assuming that $\sigma(Z)$ has a parametric functional form.

4.2.2 Semiparametric Methods

The parameters in the previous models may be nonparametrically identified.

Model 4.2.2. *(Semiparametric Tobit)* $\eta|Z \sim \mathcal{N}(\mu(Z), \sigma^2(Z))$. *In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = \sigma(Z)\left(\frac{Z'\pi + \mu(Z)}{\sigma(Z)} - \lambda\left(-\frac{Z'\pi + \mu(Z)}{\sigma(Z)}\right)\right),$$

where $\lambda(\cdot)$ *is the inverse Mill's ratio. We can identify* $(Z'\pi + \mu(Z))/\sigma(Z) = -\Phi^{-1}(F_{X|Z}(0))$, *where* Φ *is the c.d.f. of the standard normal distribution. We can also identify* $\sigma(Z) = -\mathbb{E}[X|X > 0]/(\Phi^{-1}(F_{X|Z}(0)) - \lambda(-\Phi^{-1}(F_{X|Z}(0)))$. *Therefore,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = -\mathbb{E}[X|X > 0, Z]\left(\frac{\Phi^{-1}(F_{X|Z}(0)) + \lambda(\Phi^{-1}(F_{X|Z}(0))}{-\Phi^{-1}(F_{X|Z}(0)) + \lambda(-\Phi^{-1}(F_{X|Z}(0)))}\right).$$

The complicated expression above is just a function of $F_{X|Z}(0)$ and $\mathbb{E}[X|X > 0, Z]$, which are both identifiable. In Appendix A, we show how these quantities may be estimated.

In the online appendix, we study identification in the cases in which $\eta|Z$ has a logistic, exponential or uniform distribution, while allowing all the parameters of those distributions to be fully nonparametric functions of Z. In all these cases, we derive the formula of $\mathbb{E}[X^*|X^* \leq 0, Z]$ as a function of $F_{X|Z}(0)$ and $\mathbb{E}[X|X > 0, Z]$. Identification using other distribution families can be achieved analogously.

4.3 Discrete/Discretized Z

In Section 4.3.1 we showcase the advantages that a finite support Z affords in our context, both in the identification and in the estimation of $\mathbb{E}[X^*|X^* \leq 0, Z]$. Unfortunately, it is rare that Z has a finite support in practice. In Section 4.3.2, we show how Z may be discretized to leverage the advantages of finite support in cases with arbitrary Z.

4.3.1 Methods for Z with Finite Support

In this section, suppose that for all $z \in \text{supp}(Z)$, $\mathbb{P}(Z = z) > 0$.

Semiparametric Models with Finite Support Z

We begin by showing that if $\text{supp}(Z)$ is a finite set, then the identification in the semiparametric models of Section 4.2.2 may be achieved by simpler methods which yield better estimators. Consider first the semiparametric Tobit model.

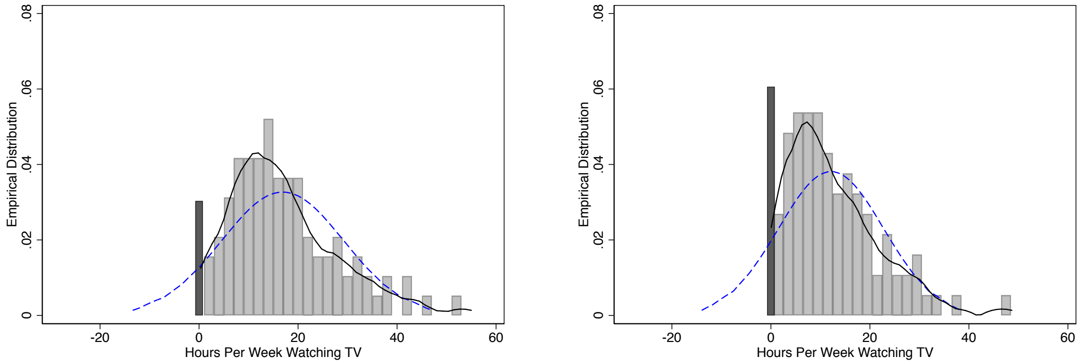
Model 4.3.1. *(Semiparametric Tobit, discrete case) Suppose that Model 4.2.2 holds. Let $\alpha_z = z'\pi + \mu(z)$ and $\sigma_z = \sigma(z)$. This implies that $X^*|Z = z \sim \mathcal{N}(\alpha_z, \sigma_z^2)$, where both the mean and variance depend arbitrarily on the value z. For a given z, this is a simple Tobit model with constant mean and variance. This means that we can identify α_z and σ_z as in (Tobin (1958)).*

We show how this model is estimated in Appendix A. Note that this approach also has computational advantages. We only need to estimate a 2-dimensional Tobit model for each value of Z in $\text{supp}(Z)$. This is generally faster than estimating a $(\dim(Z) + 1)$-dimensional Tobit model, as in the homoskedastic Tobit case (Model 4.2.1).

Figure 6 is identical to Figure 5, but it shows the semiparametric Tobit fit with the estimation method implied by Model 4.3.1. The fit in Figure 6 is better than the homoskedastic Tobit fit in Figure 5. However, the upper tails of the distributions of $X|Z$ appear to be too heavy for a normal fit. In each panel, in an effort to match the heavier tail, the fitted distribution ends up missing the location and height of the peak observed in the raw data. If the upper tails are any indication of what is happening in the lower tails, this suggests that this approach may underestimate the magnitude of $\mathbb{E}[X^*|X^* \leq 0, Z]$, thus overestimating the magnitude of δ. This is consistent with what we find in our empirical results in Section 5 (compare column (iv) to our preferred results in column (v) of Table 1).

Depending on the application, other distribution families may be more appropriate. In the online appendix, we provide the likelihood functions for the semiparametric logistic, exponential and uniform cases when Z has finite support.

Figure 6: Semiparametric Tobit Fit



Note: This figure adds to Figure 3 the fitted distribution from the semiparametric Tobit model (Model 4.3.1), shown as the dashed blue curve.

Nonparametric Methods: Symmetry

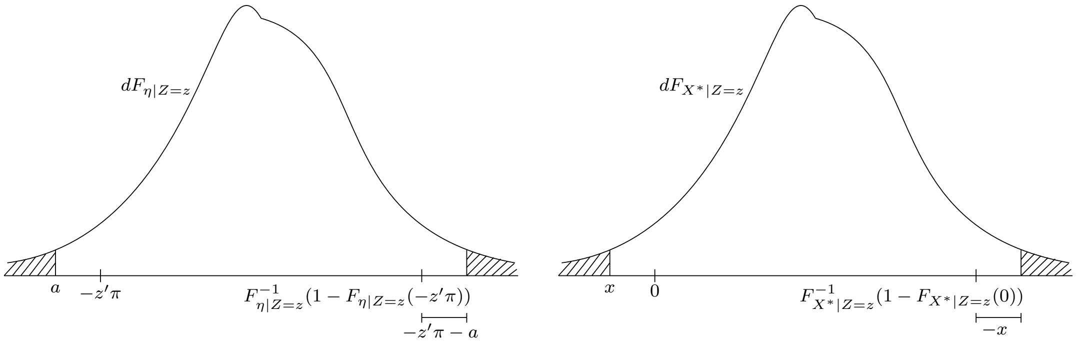
A finite-support Z makes it possible to gain more than semiparametric simplicity. If $\mathbb{P}(X = 0|Z = z) \leq 0.5$, we can drop the assumption that $\eta|Z$ belongs to a known distribution family. Instead, we can use pieces of the distribution of $X|Z = z$ reflected to the negative side.

Model 4.3.2. *(Conditional Tail Symmetry) Suppose that $F_{X|Z=z}(0) \leq 0.5$. Assume that the tails of $F_{\eta|Z=z}$ below $-z'\pi$ and above the corresponding location on the positive side are symmetric, so the dashed areas on each of the plots in Figure 7 perfectly mirror each other. Specifically, if $a \leq -z'\pi$, we suppose that*

$$F_{\eta|Z=z}(a) = 1 - F_{\eta|Z=z}(F^{-1}_{\eta|Z=z}(1 - F_{\eta|Z=z}(-z'\pi)) - a - z'\pi).$$

Note that, as is clear from the figure, we do not assume symmetry in the "middle" of the distribution (between $-z'\pi$ and $z'\pi$). Also, the mean and variance of $\eta|Z = z$ can assume any value.

Figure 7: Symmetric Density in the Tails


$dF_{\eta|Z=z}$
a
$-z'\pi$
$F^{-1}_{\eta|Z=z}(1 - F_{\eta|Z=z}(-z'\pi))$
$-z'\pi - a$
$dF_{X^*|Z=z}$
x
0
$F^{-1}_{X^*|Z=z}(1 - F_{X^*|Z=z}(0))$
$-x$

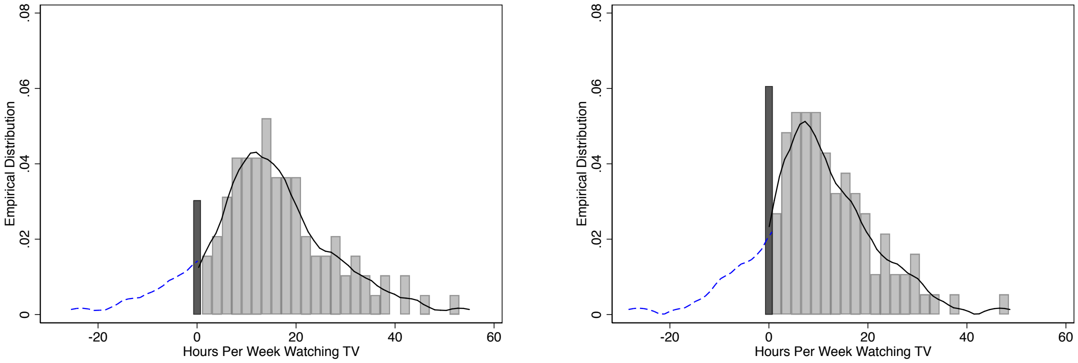
For all $x < 0$ it follows that $F_{X^*|Z=z}(x) = 1 - F_{X^*|Z=z}(F^{-1}_{X^*|Z=z}(1 - F_{X^*|Z=z}(0)) - x) = 1 - F_{X|Z=z}(F^{-1}_{X|Z=z}(1 - F_{X|Z=z}(0)) - x)$. Therefore, if we calculate the expectation, we can identify the conditional expectation via a change of variables as

$$\mathbb{E}[X^*|X^* \leq 0, Z = z] = F^{-1}_{X|Z=z}(1 - F_{X|Z=z}(0)) - \mathbb{E}[X|X \geq F^{-1}_{X|Z=z}(1 - F_{X|Z=z}(0)), Z = z].$$

We show how this model is estimated in Appendix A.

Figure 8 is identical to Figures 5 and 6 but for the tail symmetry fit. Note that the fitted plots under tail symmetry imply a heavier tail than the fitted plots under normality in those figures.

Figure 8: Conditional Tail Symmetry Fit



Note: This figure adds to Figure 3 the fitted distribution from the conditional tail symmetry model (Model 4.3.2), shown as the dashed blue curve.

In the online appendix we discuss identification and estimation under full symmetry. Full symmetry is easy to test, and it implies tail symmetry. The next remark discusses how the model may be tested.

Remark 4.1. *(Testing the Distributional Assumption) The distribution of $X^*|Z$ is observed when $X^* > 0$. Therefore, one may test the model assumptions by comparing the empirical distribution of $X|Z$ and the estimated distribution implied by the model for $X > 0$. The well known two-sample Kolmogorov-Smirnov test is an option, but there are more powerful alternatives, such as the test developed by Goldman and Kaplan (2018).*[14]

In our application, we test whether the distribution of $X^*|Z$ is symmetric by comparing the empirical distribution of $X|Z$ in $(0, med(X|Z))$ and the mirror of the empirical distribution of $X|Z$ in $(med(X|Z), F^{-1}_{X|Z}(1-F_{X|Z}(0)))$ using both Kolmogorov-Smirnov and Goldman and Kaplan (2018) tests. We also test the null hypothesis that the mean of $X|Z$ in $(0, med(X|Z))$

[14] This test is designed to have better power to detect deviations from the null at the extremities of distributions. This may be a concern in our setting when including the upper tail in the comparison of the distributions.

is the same as the mean of $X|Z$ *in the mirror image of* $(med(X|Z), F^{-1}_{X|Z}(1 - F_{X|Z}(0)))$. *For each of these three tests, we fail to reject the null hypothesis for all clusters even at the 10% level of significance, using the Bonferroni critical values to avoid concerns about multiple testing.*

4.3.2 Discretizing Z: Hierarchical Clustering

When Z does not have a finite support, it may be "discretized" using a dimensionality reduction technique. Here we focus on clustering techniques because, in order to use the methods in the previous section, our goal is to reduce the size of the support, not necessarily the number of elements in Z. For a given number of elements in the support, clustering methods aim to minimize the loss of information.

Let $\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K\}$ be a finite partition of supp(Z) into sets, which we call clusters, and let $\hat{C}_K = (\mathbf{1}(Z \in \hat{\mathcal{C}}_1), \ldots, \mathbf{1}(Z \in \hat{\mathcal{C}}_K))'$ be the cluster indicators. In the estimation of the expectation, we propose substituting Z with $\hat{C}_K$, which has finite support. The estimator $\hat{\mathbb{E}}[X^*|X^* \leq 0, Z] = \hat{\mathbb{E}}[X^*|X^* \leq 0, \hat{C}_K]$ is thus a two-step procedure in which first Z is discretized and then one of the methods in the previous section is applied.

In general, if $\mathbb{E}[X^*|X^* \leq 0, Z]$ is continuous, the ability of this estimator to approximate the expectation depends on how much information about Z is given by the cluster indicator vector $\hat{C}_K$. Thus, it is desirable to choose a clustering method that minimizes the within-cluster variation in the values of Z. All unsupervised clustering methods in the statistical learning literature could in principle be used (e.g. k-means, k-medoids, self-organizing maps, and spectral – see Hastie et al. (2009)).[15] Below, we show results using hierarchical clustering for its simplicity, but similar results were also obtained with other unsupervised clustering methods.[16]

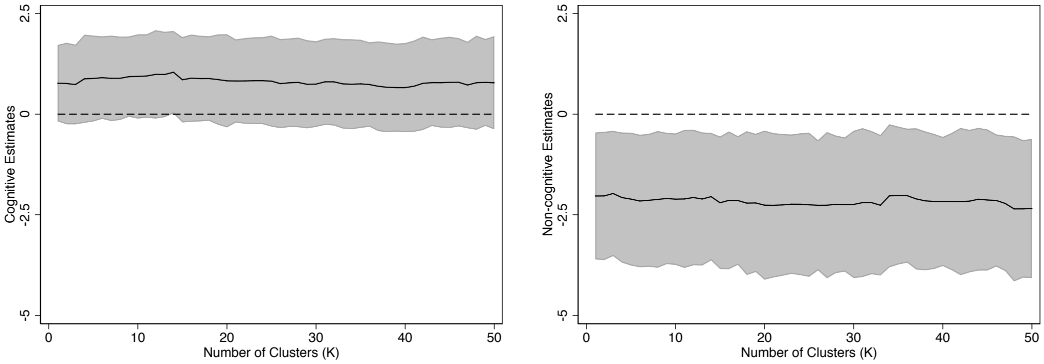
Figure 9 shows how the estimates of β in our application change as we increase the number of clusters K. As K increases, the role of the discretization assumption diminishes. If our results were an artifact of the discretization, we should expect β estimates to change as K increases. Thus, if the estimates of β remain close to constant as the number of clusters increases, as we find in Figure 9, this raises our confidence in the assumption that the clusters adequately capture differences in the conditional distributions $\eta|Z$.

There is a growing literature in economics using clustering techniques in panel settings (e.g. Bonhomme and Manresa (2015); Bonhomme et al. (2017)). In this paper, clusters are used only to improve the estimation of the expectation, not to control for unobserved heterogeneity. Clusters may in principle be used in two ways, both to improve the identification of the

[15] We can also envision supervised methods which leverage different values of Z for which there is less bunching to "train" predictions of $\mathbb{E}[X^*|X^* \leq 0, Z]$ for Z's with more bunching.

[16] Hierarchical clustering requires the choice of a dissimilarity measure and a linkage method. The reported results use the Gower measure and Ward's linkage, but we also obtained similar results with other choices.

Figure 9: Uncorrected and Corrected Cognitive and Non-cognitive β Estimates



Note: Shaded areas depict 90% confidence intervals for the corrected estimates using the tail symmetry method (model 4.3.2) and different total numbers of clusters K. All standard errors are bootstrapped using 1,000 iterations.

expectation and to control for remaining confounders that may be present in spite of the correction. For instance, Caetano et al. (2020) provides a robustness check where cluster fixed effects are also included in the corrected regression in order to allow control variables to enter the outcome equation more flexibly.

5 Application: The Effect of TV on Children's Skills

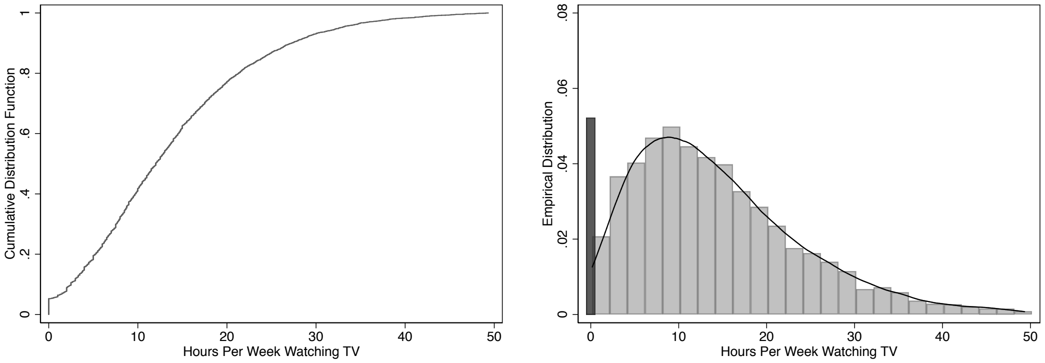
In this section, we apply our method to estimate the effect of time spent watching TV on children's skills[17] using the 1997, 2002 and 2007 Waves of the Child Development Supplement from the Panel Study of Income Dynamics (CDS-PSID). A second application estimating the effect of enrichment activities on children's skills can be found in Caetano et al. (2020). As both applications use the same sample, we refer the reader to that paper for details about the data, specification of controls, and definition of skills.

Our analysis sample has a total of 4,330 observations. We would like to estimate β in equation (1), where Y is either the cognitive or non-cognitive skills of the child, X is the number of hours the child spent watching TV in a typical week, and Z is a vector of controls which includes a constant as well as characteristics of the child, family, and environment.

Figure 10 shows the unconditional c.d.f (left panel) and the empirical distribution (right panel) of X. About 5% of the sample is bunched at $X = 0$. Examples of conditional versions of the right panel are shown in Figure 3.

[17] See for instance Zavodny (2006), Gentzkow and Shapiro (2008) and Munasib and Bhattacharya (2010) for recent empirical papers in this literature.

Figure 10: Unconditional Distribution of X



Note: The left panel shows the cumulative distribution function of X. The right panel shows the kernel density estimate along with the histogram for $X > 0$ (bandwidth equals to 2). The darker bar is the proportion of observations with $X = 0$.

5.1 Main Results

Table 1 presents the main results of the estimation of β in equation (1) both with and without our endogeneity correction. Column (i) shows the raw associations between TV time and skills without controls. Time spent watching TV is negatively correlated with both cognitive and non-cognitive skills. Column (ii) adds controls but does not use the endogeneity correction. After controlling for observables, the estimates of β are closer to zero, but the cognitive estimate is still negative and highly significant.

In columns (iii)-(v), we show estimates of the corrected regression (equation (5)), where $\mathbb{E}[X^*|X^* \leq 0, Z]$ is estimated using different models. Column (iii) shows the results under the homoskedastic Tobit assumption (Model 4.2.1). The β estimates are positive and insignificant for cognitive skills, but negative and significant for non-cognitive skills. In column (iv), we relax the assumption of homoskedasticity while keeping the assumption of normality (Model 4.3.1) and find estimates that are very close to the homoskedastic case. Finally, we relax the assumption of normality in column (v) and assume only that the conditional η distributions are symmetric in the tails (Model 4.3.2). This assumption yields estimates that are a bit closer to zero, yet the non-cognitive estimate remains significant at 5%. Statistically, all three corrections yield similar results. Table 1 also displays the estimates of δ for all correction methods. All the estimates are significant at 5% and are negative for cognitive skills and positive for non-cognitive skills.

Note that the bootstrapped standard errors of $\hat{\beta}$ in the corrected models (columns (iii)-(v) in Table 1) are much larger than the corresponding standard errors in the uncorrected model (column (ii) in Table 1). The Eicker-White standard errors of the corrected model (Σ, see Theorem 3.1) gravitate around 95% of the bootstrapped standard errors for all specifications, so the penalty due to the estimation of the expectation turns out to not be important in

Table 1: Main Empirical Results

		(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Homosk. Tobit	(iv) Semipar. Tobit	(v) Tail Symmetry
Cognitive	β	-0.57**	-0.44**	1.54	1.52	0.94
		(0.13)	(0.10)	(0.98)	(0.98)	(0.68)
	δ			-1.92**	-1.90**	-1.31**
				(0.93)	(0.93)	(0.63)
Non-Cognitive	β	-0.28*	-0.10	-3.05**	-3.07**	-2.11**
		(0.14)	(0.14)	(1.44)	(1.42)	(1.00)
	δ			2.85**	2.87**	1.91**
				(1.40)	(1.38)	(0.95)

Note: N=4,330. Results are reported in terms of percentage points of the standard deviation of the outcome variable. For example, the results in the last column suggest that an increase of one hour per week watching TV leads to a reduction in non-cognitive skills of 2.11 percentage points of one standard deviation. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). Columns (iii), (iv) and (v) show results for Models 4.2.1, 4.3.1 and 4.3.2, respectively. The corrected specifications use 10 clusters. See Figure 9 for a reproduction of the results in column (v) for different numbers of clusters. ** p<0.05, * p<0.1.

explaining the larger standard errors. Rather, the standard errors are larger because much of the raw variation in X is contaminated by variation from confounders, and thus in the uncorrected models X is predicting a large part of the error.

5.2 Supporting Evidence for Main Identifying Assumptions

This section provides a road map of the types of sensitivity analyses that we discussed throughout the paper to validate the correction approach. To keep it brief, we organize the discussion as a list of checks for each identifying assumption, and simply refer the reader to the relevant discussion in the text for details.

5.2.1 Linearity Assumption in Equations (1) and (2)

- Relax this assumption using a more flexible model specification, such as the models discussed in Section 2.2.

- Check the predicted residuals of the regression of Y on X and Z for $X > 0$. If equations (1) and (2) hold, the non-parametric fit of these residuals should be close to zero everywhere for $X > 0$, which is what we find in the application (see Section 2.1 in the online appendix).

- Estimate β for truncated samples ($X \leq \bar{x}$), then plot $\hat{\beta}$ for many values of $\bar{x}$. Results should be stable, which is what we find in the application (see Section 2.2 in the online appendix).
- If there is more than one bunching point in the support of X, implement the exogeneity test in Remark 2.3.

5.2.2 Assumption for Identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$

- Report results under different assumptions, as in Table 1.
- Run Monte Carlo simulations using the application data comparing the results of different methods under different distributional assumptions (see Section 3 in the online appendix).
- Apply the tests and calculate bounds in Propositions 4.1, 4.2, and 4.3.
- Visual checks can be done by value of Z, if possible, or by groups of values of Z otherwise. See Figures 5, 6 and 8, which show the fit for two of the clusters used for the main results (Table 1).
- Test whether the empirical distribution and the fitted distribution are the same for $X^* > 0$ (see Remark 4.1).
- Knowing the sign of δ without any assumption about the expectation (Remark 2.2) allows one to choose identification strategies for $\mathbb{E}[X^*|X^* \leq 0, Z]$ that are conservative in the context of the application. For example, if $\delta > 0$ and we want to make the point that $\beta < 0$, it is preferable to err towards an overestimation of the magnitude of $\mathbb{E}[X^*|X^* \leq 0, Z]$ so that the remaining bias after correction is positive. In this case, if $\hat{\beta}$ is still negative, we could be confident in the conclusion that β is negative even if the correction is imperfect. This is in part why we report as our main results the tail symmetry estimates (column (v) of Table 1).
- If using clusters, examine how results change with the number of clusters used in the estimation of the expectation (Figure 9 in Section 4.3.2).

Of course, some of these checks can detect violations from both assumptions jointly (e.g., the last item from each section).

6 Concluding Remarks

This paper shows how to leverage bunching at the lower (or upper) extremum of the distribution of the treatment variable to transform a problem of endogeneity into a problem of out-of-sample prediction. We examine several models in which this type of correction can be built. In a linear model, the correction consists of a generated regressor which is added to the original regression. We study the asymptotic behavior of the estimated coefficients of the corrected regression. We consider several ways in which the out-of sample prediction might be done. Finally, we apply our correction to an empirical problem and showcase how the underlying assumptions of the method may be tested or argued.

The method developed in this paper opens up several paths for new research. Here we highlight a few: (1) Throughout the paper, we proposed several tests of the underlying assumptions. Although all are based on existing tests, the consequences of the use of estimated nuisance parameters should be studied. (2) The correction strategy for the probit with endogeneity (Section 2.2.5) indicates that this type of approach may be developed for some widely used models in the structural literature, such as discrete choice models. (3) Discretizing Z before estimating the expectation proved to be a useful approach in this application and in Caetano et al. (2020). The advantages/drawbacks of the use of clusters in this context need to be investigated further. (4) The interaction of the correction with existing methods is promising. We mentioned the potential combination of this method with Caetano (2015)'s test in Remark 2.3. The interaction of this approach with instrumental variables methods may also prove valuable. Our preliminary analyses indicate that the validity requirements of an IV may be substantially weakened when the correction term is added to the regression.

A Estimators

- Remark 2.2: This can be implemented in two steps. (1) Regress Y on X and Z using only observations with $X > 0$. Record the estimate of the coefficient of Z, $\hat{\alpha}_Z$. (2) Calculate the average of the residuals at $X = 0$, $\left(\sum_{i=1}^n \mathbf{1}(X_i = 0)\right)^{-1} \sum_{i=1}^n (Y_i - Z_i'\hat{\alpha}_Z)\mathbf{1}(X_i = 0)$. This is an estimator of $\delta\mathbb{E}[X^*|X^* \leq 0]$.

- Bounds from Section 4.1 for discrete/discretized Z: for each value Z assumes, say z, restrict the sample only to observations such that $Z = z$. Then, (1) calculate $\hat{F}_{X|Z=z}(0) = \left(\sum_{i=1}^n \mathbf{1}(Z_i = z)\right)^{-1} \sum_{i=1}^n \mathbf{1}(X_i = 0, Z_i = z)$, and (2) apply the method in Cattaneo et al. (2019) to estimate both the density and the derivative terms.

- Model 4.2.1: Estimation can be done with a Tobit regression of X onto Z with censoring below zero.

- Model 4.2.2: Note that $\mathbb{E}[\mathbf{1}(X = 0)|Z = z] = F_{X|Z=z}(0)$ and can be estimated as a nonparametric regression of $\mathbf{1}(X = 0)$ onto Z at z, and $\mathbb{E}[X|X > 0, Z = z] = (1 - F_{X|Z=z}(0))^{-1}\mathbb{E}[X|Z = z]$, and $\mathbb{E}[X|Z = z]$ can be estimated as a nonparametric regression of X onto Z at z. A standard Nadaraya-Watson kernel regression could be used, for example. Let $K(u)$ be a kernel function ($\int_{-\infty}^{\infty} K(u) = 1$, and suppose if convenient that $K(u) \geq 0$ and $K(u)\mathbf{1}(|u| > 1) = 0$). Let $k_n(Z_i - z) = K((Z_i - z)/h_n)/\sum_{i=1}^n K((Z_i - z)/h_n)$, for some sequence $h_n \to 0$, $nh_n \to \infty$. Then

$$\hat{F}_{X|Z=z}(0) = \sum_{i=1}^n \mathbf{1}(X_i = 0)k_n(Z_i - z)$$

and

$$\hat{\mathbb{E}}[X|X > 0, Z = z] = (1 - \hat{F}_{X|Z=z}(0))^{-1} \sum_{i=1}^n X_i k_n(Z_i - z).$$

Conditions for uniform convergence of such estimators can be verified in the existing literature. See, for example, Andrews (1995) and Hansen (2008). One could use different estimators, for example local polynomials, see e.g. Masry (1996) or series estimators, see e.g. Song (2008).

- Model 4.3.1: For each value Z assumes, say z, run a Tobit regression of X onto a constant with censoring below zero using only observations such that $Z = z$.

- Model 4.3.2: Substitute quantities in the expectation formula by sample equivalents.

B Proofs

B.1 Proof of Theorem 3.1

For any function $\psi \in \mathcal{H}$, and parameter $\theta \in \Theta$, define $M(\theta, \psi) = \mathbb{E}[W(Y - W_\psi'\theta)]$, where $W_\psi = (X, Z', \psi(Z)\mathbf{1}(X = 0))'$, and note that $M(\theta_0, \psi_0) = 0$. Since it will be used several times, note that $W_\psi - W = (0, 0', (\psi(Z) - \psi_0(Z))\mathbf{1}(X = 0))' = (\psi(Z) - \psi_0(Z))\mathbf{1}(X = 0)e_\delta$, where $e_\delta = (0, \ldots, 0, 1)'$ is the last $(\dim(Z) + 2) \times 1$ canonical vector. Define $M_n(\theta, \psi) = \frac{1}{n}\sum_{i=1}^n W_{i\psi}(Y_i - W_{i\psi}'\theta)$. Since this is a just-identified problem, $\hat{\theta}$ is chosen exactly as the solution to $\min_\theta ||M_n(\theta, \hat{\psi})||$. Denote $\theta = (\theta_X, \theta_Z', \theta_E)'$. Define $\epsilon = Y - \mathbb{E}[Y|X, Z] = \varepsilon + \delta(\eta - \mathbb{E}[\eta|X, Z])\mathbf{1}(X = 0)$. Finally, Chen et al. (2003) also define a matrix W, which in our case should be understood as the identity matrix (i.e. whenever W appears in Chen et al. (2003), substitute it for the identity matrix).

We show the asymptotic normality, using Theorem 2 in Chen et al. (2003).

- $\hat{\theta} - \theta_0 = o_p(1)$: We prove this directly.

$$\hat{\theta} - \theta_0 = \left(\frac{1}{n}\sum_{i=1}^n W_{i\hat{\psi}} W_{i\hat{\psi}}'\right)^{-1} \frac{1}{n}\sum_{i=1}^n W_{i\hat{\psi}}(\epsilon_i - (W_{i\hat{\psi}} - W_i)'\theta_0)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} W_{i\hat{\psi}}W_{i\hat{\psi}}'\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} W_i\epsilon_i + \frac{1}{n}\sum_{i=1}^{n}(W_{i\hat{\psi}} - W_i)\epsilon_i - \frac{1}{n}\sum_{i=1}^{n} W_i(W_{i\hat{\psi}} - W_i)'\theta_0 + \right.$$
$$\left. -\frac{1}{n}\sum_{i=1}^{n}(W_{i\hat{\psi}} - W_i)(W_{i\hat{\psi}} - W_i)'\theta_0\right]$$

The term $\frac{1}{n}\sum_{i=1}^{n} W_i\epsilon_i = o_{a.s.}(1)$ by Assumption 1(ii) and Brunk-Chung's Law of Large Numbers (Chow and Teicher (1997), Theorem 10.1.3 for $r = 1$).

The term $\left|\frac{1}{n}\sum_{i=1}^{n}(W_{i\hat{\psi}} - W_i)\epsilon_i\right| \leq ||\hat{\psi} - \psi_0||\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|$. By Assumption 1(v), the first term is $o_p(1)$. The second term is $O_{a.s.}(1)$ by Assumption 1(ii) and Brunk-Chung's Law of Large Numbers.

Remaining terms are shown to be $o_p(1)$ analogously. For the last term, in Appendix B.2 we show that $\frac{1}{n}\sum_{i=1}^{n} W_{i\hat{\psi}}W_{i\hat{\psi}}' = \frac{1}{n}\sum_{i=1}^{n} W_iW_i'$ plus a matrix with terms that can be seen there. The terms inside the matrix are all similar to the terms above, in that they combine $(\hat{\psi}(Z_i) - \psi_0(Z_i))$ or the square of it, and other independent variables, and thus it can be shown that they are all $o_p(1)$ analogously to what we just did. In Appendix B.2 we also show that $\frac{1}{n}\sum_{i=1}^{n} W_iW_i' \rightarrow_{a.s.} \mathbb{E}[WW']$, which is full rank by Assumption 1(ii), which completes the proof.

- Assumption 2.1 is trivially satisfied.
- Assumption 2.2: $\Gamma_1(\theta, \psi_0) = \mathbb{E}[WW']$ by Assumption 1(ii) and the Dominated Convergence Theorem, and is constant in θ. The requirements are thus satisfied by Assumption 1(ii).
- Assumption 2.3: First, we calculate the derivative $\Gamma_2(\theta, \psi_0)[\psi - \psi_0]$. Let $\psi_t = \psi + t(\psi - \psi_0)$. By Assumption 1(ii) and the Dominated Convergence Theorem,

$$\begin{aligned}\Gamma_2(\theta, \psi_0)[\psi - \psi_0] &= \mathbb{E}\left[\lim_{t\to 0}\frac{1}{t}\left(-W(W_{\psi_t} - W)'\theta + (W_{\psi_t} - W)(Y - W_{\psi_t}'\theta)\right)\right] \\ &= \mathbb{E}\left[(\psi(Z) - \psi_0(Z))\mathbf{1}(X = 0)(0, -\theta_E Z', -W'(\theta - \theta_0) - \theta_E\psi_0(Z))'\right],\end{aligned}$$

which exists in all directions $[\psi - \psi_0] \in \mathcal{H}$. Next,

$$\begin{aligned}||M(\theta, \psi) - M(\theta, \psi_0) - \Gamma_2(\theta, \psi_0)[\psi - \psi_0]|| &= \left|\left|\mathbb{E}\left[-\theta_E(\psi(Z) - \psi_0(Z))^2\mathbf{1}(X = 0)e_\delta\right]\right|\right| \\ &\leq |\theta_E| \cdot ||\psi - \psi_0||_{\mathcal{H}}^2\end{aligned}$$

and for $\tau_n = o(1)$, and $||\theta - \theta_0|| \leq \tau_n$, and $||\psi - \psi_0|| \leq \tau_n$,

$$\begin{aligned}&||\Gamma_2(\theta, \psi_0)[\psi - \psi_0] - \Gamma_2(\theta_0, \psi_0)[\psi - \psi_0]|| = \\ &\qquad \left|\left|\mathbb{E}\left[-(\psi(Z) - \psi_0(Z))\mathbf{1}(X = 0)(0, (\theta_E - \delta)Z', W'(\theta - \theta_0) + (\theta_E - \delta)\psi_0(Z))'\right]\right|\right| \\ &\qquad\qquad \leq \sup_{||\psi - \psi_0|| \leq \tau_n} ||\psi - \psi_0||\mathbb{E}[||W||]\left(|\theta_E - \delta| + ||\theta - \theta_0||\right) \leq 2\Delta \cdot \tau_n^2,\end{aligned}$$

where the first inequality uses the triangle and Cauchy-Schwarz's inequalities, and the second inequality is true by Assumption 1(ii).

- Assumption 2.4 is true by Assumption 1(v).
- Assumption 2.5 (we prove the stronger condition 2.5'): unfortunately, we cannot take advantage of Theorem 3 in Chen et al. (2003) because we would like to allow the data to be independent but not identically distributed. We cannot apply the result they mention on Remark 3(iii) either, because the last element of our function m is not monotonic in ψ. We must therefore prove the stochastic equicontinuity condition directly.

Let $||\psi - \psi_0|| \leq \tau_n$ and $||\theta - \theta_0|| \leq \tau_n$, with $\tau_n = o(1)$. Then,

$$\begin{aligned}
&\sqrt{n}||M_n(\theta,\psi) - M(\theta,\psi) - M_n(\theta_0,\psi_0)|| \\
&\quad \leq \left|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_iW_i' - \mathbb{E}[W_iW_i'])(\theta-\theta_0)\right|\right| + \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)\epsilon_i\right| \\
&+(\tau_n+\theta_E)\left|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(\psi(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)W_i' - \mathbb{E}[(\psi(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)W_i']\right]\right|\right| \\
&\quad +\theta_E\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(\psi(Z_i)-\psi_0(Z_i))^2\mathbf{1}(X_i=0) - \mathbb{E}[(\psi(Z_i)-\psi_0(Z_i))^2\mathbf{1}(X_i=0)]\right]\right|
\end{aligned}$$

The convergence of the $\sup_{||\theta-\theta_0||\leq\delta_n}$ of the first term is established in Andrews (1994), equation (2.4), and for that we need only to show that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(W_iW_i' - \mathbb{E}[W_iW_i']) = O_p(1)$. Some of the terms $\sum_{i=1}^{n} W_{ij}W_{is}$ are constant, for example $W_{i1}W_{i,\dim(Z)+2} = X_i\psi_0(Z_i)\mathbf{1}(X_i = 0) = 0$. For the terms that are not constant, we show that Liapounov's condition is satisfied. By Assumption 1(ii),

$$\frac{\sum_{i=1}^{n}\mathbb{E}\left[|W_{ij}W_{is} - \mathbb{E}[W_{ij}W_{is}]|^{2+\alpha}\right]}{\left(\sum_{i=1}^{n}Var(W_{ij}W_{is})\right)^{1+\frac{\alpha}{2}}} \leq \frac{\sum_{i=1}^{n}\mathbb{E}\left[|W_{ij}W_{is}|^{2+\alpha}\right]}{(n\alpha)^{1+\frac{\alpha}{2}}} \leq \frac{\Delta}{n^{\frac{\alpha}{2}}\alpha^{1+\frac{\alpha}{2}}} = o(1).$$

For the remaining three terms, we note that Lemma 2.17 in Pakes and Pollard (1989) can be directly extended from their case $f(\cdot,\theta)$ to our case $f(\cdot,\psi)$, and can be proven in the same way as theirs, as pointed out by Chen et al. (2003) in the proof of their Lemma 1.

Next, we show that $f(X_i, Z_i, \epsilon_i, \psi) = (\psi(Z_i) - \psi_0(Z_i))\mathbf{1}(X_i = 0)\epsilon_i$ satisfies the conditions of Lemma 2.17 in Pakes and Pollard (1989). To see this, note that

$$|f(X_i, Z_i, \epsilon_i, \psi_1) - f(X_i, Z_i, \epsilon_i, \psi_2)| \leq b(X_i, Z_i, \epsilon_i)||\psi_1 - \psi_2||,$$

where $b(X_i, Z_i, \epsilon_i) = \mathbf{1}(X_i = 0)\epsilon_i$, and thus f is Lipschitz continuous in ψ. Assumption 1(ii) guarantees that $\mathbb{E}[b(X_i, Z_i, \epsilon_i)^{2+\alpha}] \leq \Delta$ for some $\alpha > 0$. Therefore, f is $L_2(P)$-continuous in ψ. Finally, an identical argument to Chen et al. (2003)'s proof of Theorem 3, item (i) establishes that Hölder continuity of f in ψ combined with the finite uniform entropy of ψ

(Assumption 1(iv)) imply that f belongs to an Euclidean class. A direct application of Lemma 2.17 concludes that $\sup_{||\psi-\psi_0||\leq\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)\epsilon_i\right|=o_p(1)$.

For the last two terms, we can show the local uniform continuity in probability analogously. Note that the last term is not Lipschitz, but instead Hölder continuous with constant equal to 2, which is treated identically.

- Assumption 2.6:

$$\sqrt{n}(M_n(\theta_0,\psi_0)+\Gamma_2(\theta_0,\psi_0)[\hat{\psi}-\psi_0]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_i\epsilon_i-\delta\sqrt{n}\mathbb{E}[(\hat{\psi}(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)W_i],$$

Both terms are uncorrelated (because ϵ_i is mean independent of all X_j and Z_j). White (1980) establishes (by Assumption 1(ii)) that the first term converges to a normally distributed random variable with zero mean and variance equal to the middle term in Eicker-White's covariance matrix. The second term converges to $\mathcal{N}(0,\Omega)$ by Assumption 1(vi). □

B.2 Proof of Theorem 3.2

We use the notation defined in the beginning of the previous section. The convergence in probability of the first term of $\hat{\mathbb{V}}_\theta$ to Σ is a consequence of Assumption 1(ii) and is established in White (1980). For the second term, $\hat{\delta}\rightarrow_p\delta$ proceeds from Theorem 3.1.

Next, we establish that $\left(\frac{\hat{\mathbf{w}}'\hat{\mathbf{w}}}{n}\right)^{-1}\rightarrow_p\mathbb{E}[WW']^{-1}$. We can decompose

$$\left(\frac{\hat{\mathbf{w}}'\hat{\mathbf{w}}}{n}\right)=\left(\frac{\mathbf{w}'\mathbf{w}}{n}\right)+\left(\frac{(\hat{\mathbf{w}}-\mathbf{w})'\mathbf{w}}{n}\right)+\left(\frac{\mathbf{w}'(\hat{\mathbf{w}}-\mathbf{w})}{n}\right)+\left(\frac{(\hat{\mathbf{w}}-\mathbf{w})'(\hat{\mathbf{w}}-\mathbf{w})}{n}\right)=\left(\frac{\mathbf{w}'\mathbf{w}}{n}\right)$$
$$+\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & \frac{1}{n}\sum_{i=1}^{n}Z_i(\hat{\psi}(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)\\ 0 & \frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}(Z_i)-\psi_0(Z_i))Z_i'\mathbf{1}(X_i=0) & \frac{3}{n}\sum_{i=1}^{n}(\hat{\psi}(Z_i)-\psi_0(Z_i))^2\mathbf{1}(X_i=0)\end{pmatrix}.$$

By Markov's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Z_i(\hat{\psi}(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)\right|>\tau\right)\leq\sup_{\mathcal{Z}}|\hat{\psi}(z)-\psi_0(z)|\frac{1}{\tau n}\sum_{i=1}^{n}\mathbb{E}[||Z_i||]. \quad (11)$$

By Assumption 1(v), the first term in (11) is $o_p(1)$. By Assumption 1(ii), the second term in (11) is bounded. The other terms in the matrix are shown to be $o_p(1)$ analogously.

The term $\frac{\mathbf{w}'\mathbf{w}}{n}\rightarrow_{a.s.}\mathbb{E}[WW']$ by Assumption 1(ii) and Brunk-Chung's Strong Law of Large Numbers (see Chow and Teicher (1997), Theorem 10.1.3, for $r=1$).

Next, we decompose

$$\frac{\hat{\mathbf{w}}'\hat{\mathcal{V}}\hat{\mathbf{w}}}{n^2}-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}W_iW_j'\mathbf{1}(X_i=0,X_j=0)=$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{C}_{ij}-C_{ij})(\hat{W}_i-W_i)(\hat{W}_j-W_j)'\mathbf{1}(X_i=0,X_j=0)\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}(\hat{W}_i-W_i)(\hat{W}_j-W_j)'\mathbf{1}(X_i=0,X_j=0)\\
&\quad+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{C}_{ij}-C_{ij})(\hat{W}_i-W_i)W_j'\mathbf{1}(X_i=0,X_j=0)\\
&\quad+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}(\hat{W}_i-W_i)W_j'\mathbf{1}(X_i=0,X_j=0)\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{C}_{ij}-C_{ij})W_iW_j'\mathbf{1}(X_i=0,X_j=0)
\end{aligned}
$$

Note that the sums $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|C_{ij}|\mathbf{1}(X_i=0,X_j=0)$, $\frac{1}{n}\sum_{i=1}^{n}||W_j||\mathbf{1}(X_i=0,X_j=0)$, $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||C_{ij}W_j||\mathbf{1}(X_i=0,X_j=0)$ and $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||W_iW_j'||\mathbf{1}(X_i=0,X_j=0)$ are all von Mises statistics, corresponding to U-statistics with kernel $h((X_i,Z_i')',(X_j,Z_j')')=|C_{ij}|\mathbf{1}(X_i=0,X_j=0)$ in the first case, and analogously for the others. By Assumption 2(ii) and Assumption 1(ii), the U-statistic converges *a.s.* to the kernel mean (see Theorem 3.1.1 in Korolyuk and Borovskich (2013).) Since the U-statistic converges *a.s.* to a finite constant, the von Mises statistic also converges *a.s.* to that constant.

The decomposition is therefore $o_p(1)$ by the triangle inequality and Assumption 2(iii) and because $\sup_{i=1,\dots,n}||\hat{W}_i-W_i||\leq\sup_{i=1,\dots,n}|\hat{\psi}(Z_i)-\psi_0(Z_i)|=o_p(1)$ as a consequence of Assumption 1(v).

Finally, we show that $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}W_iW_j'\mathbf{1}(X_i=0,X_j=0)\rightarrow_{a.s.}\Omega$. This is a von Mises statistic corresponding to the U-statistic with kernel function $h((X_i,Z_i')',(X_j,Z_j')')=C_{ij}W_iW_j'\mathbf{1}(X_i=0,X_j=0)$. By Assumption 2(ii), the U-statistic converges *a.s.* to its mean, Ω (Theorem 3.1.1 in Korolyuk and Borovskich (2013) again). Since the U-statistic converges *a.s.* to a finite constant, the von Mises statistic does as well. □

B.3 Proof of Remark 3.1

We show that when $\sqrt{n}(\hat{\psi}-\psi_0)$ converges in distribution to a Brownian Bridge, Assumption 1(vi) holds, and $\Omega=\mathbb{E}[C_{ij}\mathbf{1}(X_i\!=\!0,X_j\!=\!0)W_iW_j']$, for $C_{ij}=Cov(\chi_{Z_i},\chi_{Z_j})$.

Define

$$\varphi(T_n)=\int T_n(z)\mathbf{1}(x=0)(x,z',\psi_0(z)\mathbf{1}(x=0))'\mathbb{P}(dx,dz),$$

then, $\sqrt{n}\mathbb{E}[(\hat{\psi}(Z_i)-\psi_0(Z_i))\mathbf{1}(X_i=0)W_i]=\sqrt{n}(\varphi(\hat{\psi})-\varphi(\psi_0))$. The Hadamard derivative of φ at ψ_0 is

$$\varphi_{\psi_0}'(h)=\int h(z)\mathbf{1}(x=0)(x,z',\psi_0(z)\mathbf{1}(x=0))'\mathbb{P}(dx,dz)$$

Therefore, by Assumption 1(iv) and the Functional Delta Method,

$$\sqrt{n}(\varphi(\hat{\psi}) - \varphi(\psi_0)) \to_d \int \chi_z \mathbf{1}(x=0)(x, z', \psi_0(z)\mathbf{1}(x=0))'\mathbb{P}(dx, dz).$$

Denote the limit random variable as A, and $w(x,z) = (x, z', \psi_0(z)\mathbf{1}(x=0))'$, then

$$A \sim \mathcal{N}\left(0, \iint C(z,\tilde{z})\mathbf{1}(x=0,\tilde{x}=0)w(x,z)\tilde{w}(\tilde{x},\tilde{z})'\mathbb{P}(dx,dz)\mathbb{P}(d\tilde{x},d\tilde{z})\right).$$

(See e.g. Van der Vaart (1998) Example 22.11 for a similar calculation). □

B.4 Proof of Theorem 3.3

We show that the assumptions in Theorem B in Chen et al. (2003) hold. We use the notation defined in the beginning of Section B.1. Note that, in our case, $m(Y_i, X_i, Z_i, \theta, \psi(Z_i)) = W_{i\psi}(Y_i - W'_{i\psi}\theta)$.

- First, we show that $\hat{\theta} - \theta_0 = o_{a.s.}(n^{-1/4})$. We prove this directly. The decomposition is the same as in the first item in Appendix B.1:

$$\hat{\theta} - \theta_0 = \left(\frac{1}{n}\sum_{i=1}^n W_{i\hat{\psi}}W'_{i\hat{\psi}}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n W_i\epsilon_i + \frac{1}{n}\sum_{i=1}^n (W_{i\hat{\psi}} - W_i)\epsilon_i - \frac{1}{n}\sum_{i=1}^n W_i(W_{i\hat{\psi}} - W_i)'\theta_0 + \right.$$
$$\left. - \frac{1}{n}\sum_{i=1}^n (W_{i\hat{\psi}} - W_i)(W_{i\hat{\psi}} - W_i)'\theta_0\right].$$

The term $\frac{1}{n}\sum_{i=1}^n W_i\epsilon_i = o_{a.s.}(n^{-1/2+\alpha})$ for some $\alpha > 0$ by Assumption 1(ii) and Marcinkiewicz-Zygmund Strong Law of Large Numbers (Chow and Teicher (1997) Theorem 5.3.2).

The term $\left|\frac{1}{n}\sum_{i=1}^n (W_{i\hat{\psi}} - W_i)\epsilon_i\right| \leq ||\hat{\psi} - \psi_0||\frac{1}{n}\sum_{i=1}^n |\epsilon_i|$. By Assumption 3(iii), the first term is $o_{a.s.}(n^{-1/4})$. The second term is $O_{a.s.}(1)$ by the Strong Law of Large Numbers.

The remaining two terms inside the brackets are shown to be $o_{a.s.}(n^{-1/4})$ analogously. For the last term, in Section B.2 we showed that $\frac{1}{n}\sum_{i=1}^n W_{i\hat{\psi}}W'_{i\hat{\psi}} = \frac{1}{n}\sum_{i=1}^n W_iW'_i$ plus a matrix with terms which can be seen there. The terms inside the matrix are all similar to the terms above, in that they combine $(\hat{\psi}(Z_i) - \psi_0(Z_i))$, or the square of it, and other *i.i.d.* variables, and thus it can be shown that they are all $o_{a.s.}(n^{-1/4})$ analogously to what we just did. In the previous section, we also showed that $\frac{1}{n}\sum_{i=1}^n W_iW'_i \to_{a.s.} \mathbb{E}[WW']$, which is full rank by Assumption 1(ii) and completes the proof.

- Assumption 2.1 holds a.s. trivially.
- Assumption 2.4 holds a.s. by Assumption 3(v).
- Assumption 2.5' holds a.s.: this assumption is used to show that $||\nu_n(\hat{\theta}^*, \hat{h}^*) - \nu_n(\hat{\theta}, \hat{h})|| = o_{p*}(1)$. Instead of using the almost sure stochastic equicontinuity, we establish this result directly.

By the triangle inequality, the term is bounded above by $||\nu_n(\hat{\theta}^*, \hat{h}^*) - \nu_n(\theta_0, h_0)|| + ||\nu_n(\hat{\theta}, \hat{h}) - \nu_n(\theta_0, h_0)||$, and in our case, each of those terms is bounded above by four quantities identical to the bounds in the proof of Assumption 2.5 in Section B.1, except that ψ and θ are substituted by $\hat{\psi}^b$ and $\hat{\theta}^b$, and $\hat{\psi}$ and $\hat{\theta}$ respectively.

The first quantity is bounded above by $\left|\left|\frac{1}{n^{1/2+c}} \sum_{i=1}^n W_i W_i' - \mathbb{E}[W_i W_i']\right|\right| n^c ||\theta - \theta_0||$ for some $0 < c \leq 1/4$ and $\theta = \hat{\theta}^b$ or $\theta = \hat{\theta}$, respectively. By Assumption 1(ii) and Marcinkiewicz-Zygmund Law of Large Numbers (Chow and Teicher (1997) Theorem 5.3.2), the first term is $o_{a.s.}(1)$. We also proved (first point in this Section) that $||\hat{\theta} - \theta_0|| = o_{a.s.}(n^{-1/4})$. Given Assumptions 3 (va) and (vc), all the assumptions of Theorem 3.1 hold with ψ^b in place of $\hat{\psi}$ and $\hat{\psi}$ in place of ψ_0, and changing the probability measure from $\mathbb{P}$ to $\mathbb{P}^b$. Therefore, by the proof of Theorem 2 in Chen et al. (2003) (as mentioned in the proof of their Theorem B), $||\hat{\theta}^b - \hat{\theta}|| = O_{p^b}(n^{-1/2})$. Therefore $||\hat{\theta}^b - \theta_0|| = O_{p^b}(n^{-1/2})$, which completes the proof.

The result for the remaining quantities is implied directly by Assumptions 3 (iv) and (vb).

- Assumption 2.6 does not have “in probability" in its statement and thus holds by the proof in Section B.1.

- Assumption 2.2 with ψ_0 replaced by $\psi \in \mathcal{H}_\tau$: The derivative $\Gamma_1(\theta_0, \psi) = \mathbb{E}[W_\psi W_\psi']$ exists, is continuous everywhere in θ (it does not depend on θ), and is of full rank for τ sufficiently small by Assumption 1(ii).

Note also that $\Gamma_1(\theta_0, \psi)$ is continuous in ψ at $\theta = \theta_0$ and $\psi = \psi_0$, since

$$\begin{aligned} ||\Gamma_1(\theta_0, \psi) - \Gamma_1(\theta_0, \psi_0)|| &\leq \mathbb{E}[||(W_\psi - W)W_\psi' + W(W_\psi - W)'||] \\ &\leq ||\psi - \psi_0|| \mathbb{E}[||W_\psi|| + ||W||] \leq (2\Delta + \tau)||\psi - \psi_0||. \end{aligned}$$

- Assumption 2.3 with ψ_0 replaced by $\psi \in \mathcal{H}_{\tau_n}$:

$$\Gamma_2(\theta, \psi)[\tilde{\psi} - \psi] = \mathbb{E}\left[(\tilde{\psi}(Z) - \psi(Z))\mathbf{1}(X = 0)(0, -\theta_E Z', -W'(\theta - \theta_0) - \theta_E \psi(Z))'\right]$$

exists in all directions $[\tilde{\psi} - \psi] \in \mathcal{H}$. Moreover,

$$\begin{aligned} \left|\left|M(\theta, \tilde{\psi}) - M(\theta, \psi) - \Gamma_2(\theta, \psi)[\tilde{\psi} - \psi]\right|\right| &= \left|\left|\mathbb{E}\left[-\theta_E(\tilde{\psi}(Z) - \psi(Z))^2 \mathbf{1}(X = 0) e_\delta\right]\right|\right| \\ &\leq |\theta_E| \cdot ||\tilde{\psi} - \psi||^2_{\mathcal{H}}, \end{aligned}$$

and for $\tau_n = o(1)$, and $||\theta - \theta_0|| \leq \tau_n$, and $||\tilde{\psi} - \psi|| \leq \tau_n$

$$\begin{aligned} &\left|\left|\Gamma_2(\theta, \psi)[\tilde{\psi} - \psi] - \Gamma_2(\theta_0, \psi)[\tilde{\psi} - \psi]\right|\right| = \\ &\qquad \left|\left|\mathbb{E}\left[-(\tilde{\psi}(Z) - \psi(Z))\mathbf{1}(X = 0)(0, (\theta_E - \delta)Z', W'(\theta - \theta_0) + (\theta_E - \delta)\psi(Z))'\right]\right|\right| \\ &\qquad \leq \sup_{||\tilde{\psi} - \psi|| \leq \tau_n} ||\tilde{\psi} - \psi|| \mathbb{E}[||W_\psi||] \left(|\theta_E - \delta| + ||\theta - \theta_0||\right) \leq 2(\Delta + \tau_n)\tau_n^2, \end{aligned}$$

where the first inequality uses the triangle and Cauchy-Schwarz's inequalities, and the second inequality uses the triangle inequality again and Assumption 1(ii).

- Assumption 2.4B holds by Assumption 3(va).
- We show Assumption 2.5'B by applying Theorem 3 in Chen et al. (2003). In our case, $l = \dim(Z) + 2$ and $m(Y, X, Z, \theta, \psi) = m_c(Y, X, Z, \theta, \psi)$, and $m_{lc}(Y, X, Z, \theta, \psi) = 0$, which automatically satisfies condition 3.2. Condition 3.3 is satisfied by Assumptions 1 (iii) and (iv). We show condition 3.1:

$$\begin{aligned} |m_c(Y, X, Z, \theta_1, \psi_1) - m_c(Y, X, Z, \theta_2, \psi_2)| = \\ = |(W_{\psi_1} - W_{\psi_2})Y - (W_{\psi_1}W_{\psi_1} - W_{\psi_2}W_{\psi_2})\theta_2 - W_{\psi_1}W'_{\psi_1}(\theta_1 - \theta_2)| \\ \leq |\psi_1(Z) - \psi_2(Z)|(|Y| + ||\theta_1||(2||Z|| + |\psi_1(Z)| + |\psi_2(Z)|)) \\ + ||\theta_1 - \theta_2||(X^2 + 2|X| \cdot ||Z|| + ||Z||^2 + 2||Z|| \cdot |\psi_1(Z)| + \psi_1(Z)^2). \end{aligned}$$

There exists $\Delta > 0$ such that by Assumption 1(iii), $||\theta_1|| \leq \Delta$ and by Assumption 3(ii), $|\psi_1(Z)| \leq \Delta$. Therefore, $b(Y, X, Z) = \max\{|Y| + 2\Delta||Z|| + 2\Delta^2, X^2 + 2(|X| + \Delta) \cdot ||Z|| + ||Z^2|| + \Delta^2\}$. By Assumption 1(ii), $\mathbb{E}[b(Y, X, Z)] < \infty$. Therefore, the condition holds with $s_{1j} = s_j = 1$.

- Assumption 2.6B: let $M_n^b(\hat{\theta}, \hat{\psi})$ be equal to $M_n(\hat{\theta}, \hat{\psi})$ as defined in Section B.1, except that it is calculated with the bootstrap sample. As stated in Chen et al. (2003) (p. 1596), from Giné and Zinn (1990) we know that the $\mathbb{P}^b$-distribution of $\sqrt{n}(M_n^b(\hat{\theta}, \hat{\psi}) - M_n(\hat{\theta}, \hat{\psi}))$ approximates the distribution of $\sqrt{n}(M_n(\hat{\theta}, \hat{\psi}) - M(\hat{\theta}, \hat{\psi}))$, which is approximately the same as the distribution of $\sqrt{n}M_n(\theta_0, \psi_0)$ by condition 2.5' shown in Section B.1.

Next, we show that the the $\mathbb{P}^b$-distribution of $\sqrt{n}\Gamma_2(\hat{\theta}, \hat{\psi})[\hat{\psi}^b - \hat{\psi}]$ approximates the distribution of $\sqrt{n}\Gamma_2(\theta_0, \psi_0)[\hat{\psi} - \psi_0]$. Specifically, $\sqrt{n}(\Gamma_2(\hat{\theta}, \hat{\psi})[\hat{\psi}^b - \hat{\psi}] - \Gamma_2(\theta_0, \psi_0)[\hat{\psi} - \psi_0]) = (0, A_n, B_n)$, where

$$\begin{aligned} A_n = &- \delta\sqrt{n}\mathbb{E}[[(\hat{\psi}^b(Z) - \hat{\psi}(Z)) - (\hat{\psi}(Z) - \psi_0(Z))]\mathbf{1}(X = 0)Z'] \\ &- \sqrt{n}\mathbb{E}[(\hat{\theta}_E - \delta)(\hat{\psi}^b(Z) - \hat{\psi}(Z))\mathbf{1}(X = 0)Z'] \end{aligned}$$

$$\begin{aligned} B_n = &- \delta\sqrt{n}\mathbb{E}[[(\hat{\psi}^b(Z) - \hat{\psi}(Z)) - (\hat{\psi}(Z) - \psi_0(Z))]\mathbf{1}(X = 0)\psi_0(Z)] \\ &- \sqrt{n}\mathbb{E}[(\hat{\theta}_E - \delta)(\hat{\psi}^b(Z) - \hat{\psi}(Z))\mathbf{1}(X = 0)\psi_0(Z)] \\ &- \sqrt{n}\mathbb{E}[\hat{\theta}_E(\hat{\psi}^b(Z) - \hat{\psi}(Z))(\hat{\psi}(Z) - \psi_0(Z))\mathbf{1}(X = 0)]. \end{aligned}$$

By Assumption 3(vc), the first terms in A_n and B_n are $o_{p^b}(1)$. Now we discuss the second term in A_n. By Cauchy-Schwartz, its absolute value is bounded above by

$$\mathbb{E}[n(\hat{\theta}_E - \delta)^4]^{1/4}\mathbb{E}[||Z||^4]^{1/4}n^{1/4}\mathbb{E}[(\hat{\psi}^b(Z) - \hat{\psi}(Z))^2]^{1/2}. \tag{12}$$

By the first point in this Section, $n^{1/4}||\hat{\theta} - \theta_0|| = o_{a.s.}(1)$, and by Assumption 1(iii) and the Dominated Convergence Theorem, the first term in (12) converges to zero. The second term is bounded by Assumption 1(ii). Assumption 3(va) implies that the third term is $o_{p^b}(1)$. Therefore, the second term in A_n is $o_{p^b}(1)$. The second and third terms in B_n are also $o_{p^b}(1)$, and the proof is analogous, except that to bind the second term on B_n we use Assumption 3(ii), and to bind the third term on B_n we use Assumption 3(iii), the Continuous Mapping Theorem, and the Dominated Convergence Theorem.

Thus, $\sqrt{n}(M_n^b(\hat{\theta},\hat{\psi})-M_n(\hat{\theta},\hat{\psi})+\Gamma_2(\hat{\theta},\hat{\psi})[\hat{\psi}^b-\hat{\psi}])$ and $\sqrt{n}(M_n(\theta_0,\psi_0)+\Gamma_2(\theta_0,\psi_0)[\hat{\psi}-\psi_0])$ have the same asymptotic distribution. By the proof of Assumption 2.6 in Section B.1, this is the desired limit. □

B.5 Lemma: Establishing Stochastic Dominance in an Interval.

Let $0 < a \leq b < \infty$, and suppose that (i) f and g are two non-negative functions; (ii) $g(x) \geq f(x)$ for all $x \in [0,a]$; and (iii) $\int_0^a g(x)dx = \int_0^b f(x)dx = \Lambda(b) < \infty$. Then $\int_0^b xf(x)dx \geq \int_0^a xg(x)dx$.

Proof. Let $\int_0^x f(u)du = \Lambda(x)$, and $\int_0^x g(u)du = G(x)$. By Leibniz rule, item (iii), and the Mean Value Theorem for integrals,

$$\int_0^b xf(x)dx - \int_0^a g(x)dx = (b-a)(\Lambda(b)-\Lambda(c)) + \int_0^a (G(x)-\Lambda(x))dx,$$

where $c \in [a,b]$. By item (i), $\Lambda(b) \geq \Lambda(c)$, and by item (ii), the second term is non-negative. □

References

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72.

Andrews, D. W. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory*, pages 560–596.

Baum II, C. L. (2003). Does early maternal employment harm child development? An analysis of the potential benefits of leave taking. *Journal of Labor Economics*, 21(2):409–448.

Bertanha, M., McCallum, A. H., and Seegert, N. (2020). Better bunching, nicer notching. Working Paper.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., and Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, 125(1):263–306.

Bettinger, E., Hægeland, T., and Rege, M. (2014). Home with mom: The effects of stay-at-home parents on children's long-run educational outcomes. *Journal of Labor Economics*, 32(3):443–467.

Bhutani, S., Klempel, M. C., Kroeger, C. M., Aggour, E., Calvo, Y., Trepanowski, J. F., Hoddy, K. K., and Varady, K. A. (2013). Effect of exercising while fasting on eating behaviors and food intake. *Journal of the International Society of Sports Nutrition*, 10(1):50.

Bischoff-Ferrari, H. A., Willett, W. C., Wong, J. B., Giovannucci, E., Dietrich, T., and Dawson-Hughes, B. (2005). Fracture prevention with vitamin D supplementation: A meta-analysis of randomized controlled trials. *JAMA*, 293(18):2257–2264.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2005). The more the merrier? The effect of family size and birth order on children's education. *The Quarterly Journal of Economics*, 120(2):669–700.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2010). Small family, smart family? Family size and the IQ scores of young men. *Journal of Human Resources*, 45(1):33–58.

Bleemer, Z. (2018a). The effect of selective public research university enrollment: Evidence from California. Research & Occasional Paper Series: CSHE. 11.18. *Center for Studies in Higher Education.*

Bleemer, Z. (2018b). Top percent policies and the return to postsecondary selectivity. *Working Paper.*

Blomquist, N. S., Newey, W. K., Kumar, A., and Liang, C.-Y. (2019). On bunching and identification of the taxable income elasticity. *CENMAP Working Paper.*

Bonhomme, S., Lamadon, T., and Manresa, E. (2017). Discretizing Uunobserved Heterogeneity. Working Paper.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Boserup, S. H., Kopczuk, W., and Kreiner, C. T. (2016). The role of bequests in shaping wealth inequality: Evidence from Danish wealth records. *American Economic Review*, 106(5):656–61.

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, Communication & Society*, 18(5):524–538.

Brown, J. R., Coile, C. C., and Weisbenner, S. J. (2010). The effect of inheritance receipt on retirement. *The Review of Economics and Statistics*, 92(2):425–434.

Caetano, C. (2015). A test of exogeneity without instrumental variables in models with bunching. *Econometrica*, 83(4):1581–1600.

Caetano, C., Caetano, G., and Nielsen, E. (2020). Should children do more enrichment activities? Leveraging bunching to correct for endogeneity. *FEDS Working Paper No. 2020-036.*

Caetano, G., Kinsler, J., and Teng, H. (2019). Towards causal estimates of children's time allocation on skill development. *Journal of Applied Econometrics*, 34(4):588–605.

Caetano, G. and Maheshri, V. (2018). Identifying Dynamic Spillovers of Crime with a Causal Approach to Model Selection. *Quantitative Economics*, 9(1):343–394.

Carman, K. G. (2013). Inheritances, intergenerational transfers, and the accumulation of health. *American Economic Review*, 103(3):451–55.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, pages 1–7.

Chatterji, P., Markowitz, S., and Brooks-Gunn, J. (2013). Effects of early maternal employment on maternal health and well-being. *Journal of Population Economics*, 26(1):285–301.

Chay, K. Y. and Greenstone, M. (2005). Does air quality matter? Evidence from the housing market. *Journal of political Economy*, 113(2):376–424.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does your Kindergarten Classroom Affect your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory.* Springer - New York.

Cohen, M. A. (2008). The effect of crime on life satisfaction. *The Journal of Legal Studies*, 37(S2):S325–S353.

Corrao, G., Rubbiati, L., Bagnardi, V., Zambon, A., and Poikolainen, K. (2000). Alcohol and coronary heart disease: A meta-analysis. *Addiction*, 95(10):1505–1523.

De Vito, A., Jacob, M., and Müller, M. A. (2019). Avoiding taxes to fix the tax code. *Working Paper.*

Ekici, T. and Dunn, L. (2010). Credit card debt and consumption: Evidence from household-level data. *Applied Economics*, 42(4):455–462.

Elinder, M., Erixson, O., and Waldenström, D. (2018). Inheritance and wealth inequality: Evidence from population registers. *Journal of Public Economics*, 165:17 – 30.

Eren, O. and Henderson, D. J. (2011). Are we wasting our children's time by giving them more homework? *Economics of Education Review*, 30(5):950–961.

Erhardt, E. C. (2017). Microfinance beyond self-employment: Evidence for firms in Bulgaria. *Labour economics*, 47:75–95.

Erixson, O. (2017). Health responses to a wealth shock: Evidence from a Swedish tax reform. *The Journal of Population Economics*, 30:1281–1336.

Ermisch, J. and Francesconi, M. (2013). The effect of parental employment on child schooling. *Journal of Applied Econometrics*, 28(5):796–822.

Fawzi, W. W., Chalmers, T. C., Herrera, M. G., and Mosteller, F. (1993). Vitamin A supplementation and child mortality: A meta-analysis. *JAMA*, 269(7):898–903.

Ferreira, D., Ferreira, M. A., and Mariano, B. (2018). Creditor control rights and board independence. *The Journal of Finance*, 73(5):2385–2423.

Garen, J. (1984). The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica: Journal of the Econometric Society*, pages 1199–1218.

Gentzkow, M. and Shapiro, J. M. (2008). Preschool television viewing and adolescent test scores: Historical evidence from the Coleman study. *The Quarterly Journal of Economics*, 123(1):279–323.

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *The Annals of Probability*, pages 851–869.

Goldman, M. and Kaplan, D. M. (2018). Comparing distributions by multiple testing across quantiles or CDF values. *Journal of Econometrics*, 206(1):143–166.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, pages 726–748.

Härdle, W., Liang, H., and Gao, J. (2000). *Partially linear models*. Physica-Verlag Heidelberg.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Hernán, M. A., Takkouche, B., Caamaño-Isorna, F., and Gestal-Otero, J. J. (2002). A meta-analysis of coffee drinking, cigarette smoking, and the risk of parkinson's disease. *Annals of neurology*, 52(3):276–284.

Holt, K., Shehata, A., Strömbäck, J., and Ljungberg, E. (2013). Age and the effects of news media attention and social media use on political interest and participation: Do social media function as leveller? *European Journal of Communication*, 28(1):19–34.

James-Burdumy, S. (2005). The effect of maternal labor force participation on child development. *Journal of Labor Economics*, 23(1):177–211.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of econometrics*, 150(1):86–98.

Joulfaian, D. and Wilhelm, M. O. (1994). Inheritance and labor supply. *The Journal of Human Resources*, 29(4):1205–1234.

Kim, B. and Ruhm, C. J. (2012). Inheritances, health and death. *Health Economics*, 21(2):127–144.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8(1):435–464.

Kleven, H. J. and Waseem, M. (2013). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *The Quarterly Journal of Economics*, 128(2):669–723.

Korolyuk, V. S. and Borovskich, Y. V. (2013). *Theory of U-statistics*, volume 273. Springer Science & Business Media.

Lavetti, K. and Schmutte, I. M. (2018). Estimating compensating wage differentials with endogenous job mobility. *Working paper.*

Luoh, M.-C. and Herzog, A. R. (2002). Individual consequences of volunteer and paid work in old age: Health and mortality. *Journal of health and social behavior*, pages 490–509.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599.

McDuffie, R. S., Beck, A., Bischoff, K., Cross, J., and Orleans, M. (1996). Effect of frequency of prenatal care visits on perinatal outcome among low-risk women: A randomized controlled trial. *JAMA*, 275(11):847–851.

Melzer, B. T. (2011). The real costs of credit access: Evidence from the Payday lending market. *The Quarterly Journal of Economics*, 126(1):517–555.

Munasib, A. and Bhattacharya, S. (2010). Is the 'idiot's box' raising idiocy? Early and middle childhood television watching and child cognitive outcome. *Economics of Education Review*, 29(5):873 – 883.

Noordzij, M., Uiterwaal, C. S., Arends, L. R., Kok, F. J., Grobbee, D. E., and Geleijnse, J. M. (2005). Blood pressure response to chronic intake of coffee and caffeine: A meta-analysis of randomized controlled trials.

Oken, E., Levitan, E., and Gillman, M. (2008). Maternal smoking during pregnancy and child overweight: Systematic review and meta-analysis. *International Journal of Obesity*, 32(2):201–210.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057.

Pang, J. (2017). Do subways improve labor market outcomes for low-skilled workers. *Working Paper, Syracuse University.*

Peek, J., Rosengren, E. S., and Tootell, G. M. (2003). Identifying the macroeconomic effect of loan supply shocks. *Journal of Money, Credit and Banking*, pages 931–946.

Pötscher, B. M. and Prucha, I. R. (1994). Generic uniform convergence and equicontinuity concepts for random functions: An exploration of the basic structure. *Journal of Econometrics*, 60(1-2):23–63.

Reynolds, K., Lewis, B., Nolen, J. D. L., Kinney, G. L., Sathya, B., and He, J. (2003). Alcohol consumption and risk of stroke: A meta-analysis. *JAMA*, 289(5):579–588.

Richardson, T., Elliott, P., and Roberts, R. (2013). The relationship between personal unsecured debt and mental and physical health: A systematic review and meta-analysis. *Clinical Psychology Review*, 33(8):1148–1162.

Robinson, P. M. (1988). Root- N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Rozenas, A., Schutte, S., and Zhukov, Y. (2017). The political legacy of violence: The long-term impact of Stalin's repression in Ukraine. *The Journal of Politics*, 79(4):1147–1161.

Ruhm, C. J. (2004). Parental employment and child cognitive development. *The Journal of Human Resources*, 39(1):155–192.

Ruhm, C. J. (2008). Maternal employment and adolescent development. *Labour Economics*, 15(5):958 – 983.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, 2(3):180–212.

Shinton, R. and Beevers, G. (1989). Meta-analysis of relation between cigarette smoking and stroke. *BMJ*, 298(6676):789–794.

Song, K. (2008). Uniform convergence of series estimators over function spaces. *Econometric Theory*, pages 1463–1499.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 26(1):24–36.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Zavodny, M. (2006). Does watching television rot your mind? Estimates of the effect on test scores. *Economics of Education Review*, 25(5):565 – 573.

ONLINE APPENDIX
Correcting for Endogeneity in Models with Bunching

Carolina Caetano
University of Georgia

Gregorio Caetano
University of Georgia

Eric Nielsen
Federal Reserve Board

August 2020

This appendix supplements the paper by offering further analyses. In Section 1, we discuss alternative approaches for the identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$ that were mentioned but not shown in the paper. In Section 2, we present two empirical checks of the linearity assumptions in equations (1) and (2) in the paper which proved useful in the application in Section 5, as well as in Caetano et al. (2020). Finally, in Section 3, we present the results from a real-data Monte Carlo simulation.

1 Identifying $\mathbb{E}[X^*|X^* \leq 0, Z]$

1.1 Parametric Methods

We discuss how $\mathbb{E}[X^*|X^* \leq 0, Z]$ may be obtained in parametric models under some well-known distribution families. This section supplements the discussion in Section 4.2.1 in the paper.

Model 4.2.3 *(Logistic) $\eta|Z \sim$ Logistic$(Z'\kappa, \sigma)$ almost surely. In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = Z'(\pi + \kappa) - \sigma \log(1 + \exp(Z'(\pi + \kappa)/\sigma)).$$

The log-likelihood function is

$$\mathcal{L}(\pi + \kappa, \sigma) = -\sum_{i=1}^{n} \mathbf{1}(X_i = 0) \log(1 + \exp(Z_i'(\pi + \kappa)/\sigma)) + \mathbf{1}(X_i > 0)(X_i - Z_i'(\pi + \kappa))/\sigma$$
$$+ \mathbf{1}(X_i > 0) \log \sigma + 2 \cdot \mathbf{1}(X_i > 0) \log(1 + \exp(-X_i + Z_i'(\pi + \kappa))/\sigma).$$

Note that, similarly to the Tobit case, we only identify the sum $\pi + \kappa$ instead of π and κ separately, and this is sufficient to identify the expectation. Indeed, whenever η can be written as $Z'\kappa + \zeta$, where ζ is a distribution with fixed location, we will only identify $\pi + \kappa$, but this will be sufficient. This is true in the Tobit case, where $\zeta \sim \mathcal{N}(0, \sigma)$, and in the Logistic case, where $\zeta \sim$ Logistic$(0, \sigma)$. These are both symmetric distributions where the location is equal to the mean.

The uniform distribution is an example of a symmetric distribution with location determined by the extremum.

Model 4.2.4 *(Uniform) $\eta|Z \sim U[Z'\kappa, Z'\mu]$ almost surely. In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = \frac{1}{2}Z'(\pi + \kappa).$$

The log-likelihood function is

$$\mathcal{L}(\pi + \kappa, \mu - \kappa) = \sum_{i=1}^{n} \mathbf{1}(X_i = 0)\log(-Z_i'(\pi + \kappa)) - \log(Z_i'(\mu - \kappa)).$$

The last example is of an asymmetric distribution where the location is set by the lower limit of the support. The model implies that the distribution of X^* has support $[Z'(\pi+\kappa), \infty)$ with the higher concentration towards the lower values of X^*.

Model 4.2.5 *(Exponential) $\eta = Z'\kappa + \zeta$, where $\zeta|Z \sim Exp((Z'\mu)^{-1})$ almost surely. In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = Z'\mu + Z'(\pi + \kappa)\frac{(1 + \exp(Z'(\pi + \kappa)/Z'\mu))}{(1 - \exp(Z'(\pi + \kappa)/Z'\mu))}.$$

The log-likelihood function is

$$\mathcal{L}(\pi+\kappa, \sigma) = \sum_{i=1}^{n} \mathbf{1}(X_i = 0)\log(1-\exp(Z_i'(\pi+\kappa)/Z_i'\mu)) - \mathbf{1}(X_i > 0)(\log Z_i'\mu + (X_i - Z_i'(\pi+\kappa))/Z_i'\mu).$$

1.2 Semiparametric Methods

We discuss how $\mathbb{E}[X^*|X^* \leq 0, Z]$ may be obtained in semiparametric models in which the distribution family is known, but the parameters are identified nonparametrically. This section supplements the discussion in Section 4.2.2 in the paper.

Model 4.2.6 *(Semiparametric Logistic) $\eta|Z \sim Logistic(\kappa(Z), \sigma(Z))$ almost surely. In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = \sigma(Z)\left(\frac{Z'\pi + \kappa(Z)}{\sigma(Z)} - \log\left(1 + \exp\left(\frac{Z'\pi + \kappa(Z)}{\sigma(Z)}\right)\right)\right).$$

We can identify $1+\exp((Z'\pi+\kappa(Z))/\sigma(Z)) = (F_{X|Z}(0))^{-1}$, and $\sigma(Z) = -\mathbb{E}[X|X > 0, Z](1-F_{X|Z}(0))/\log(F_{X|Z}(0))$, and thus

$$\mathbb{E}[X^*|X^* \leq 0, Z] = -\mathbb{E}[X|X > 0, Z]\left(\frac{(1 - F_{X|Z}(0))\log(1 - F_{X|Z}(0))}{F_{X|Z}(0)\log F_{X|Z}(0)}\right)$$

Model 4.2.7 *(Semiparametric Uniform) $\eta|Z \sim U[\kappa(Z), \mu(Z)]$ almost surely. In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = \frac{1}{2}(Z'\pi + \kappa(Z)).$$

We can identify $Z'\pi + \kappa(Z)/(\mu(Z) - \kappa(Z)) = -F_{X|Z}(0)$, $Z'\pi + \mu(Z)/(\mu(Z) - \kappa(Z)) = 1 - F_{X|Z}(0)$, and $\mathbb{E}[X|X > 0, Z] = 1/2(Z'\pi + \mu(Z))$, thus

$$\mathbb{E}[X^*|X^* \leq 0, Z] = -\mathbb{E}[X|X > 0, Z]\frac{F_{X|Z}(0)}{1 - F_{X|Z}(0)}.$$

Model 4.2.8 *(Semiparametric Exponential) $\eta = \kappa(Z) + \zeta$, where $\zeta|Z \sim Exp(\mu(Z)^{-1})$ almost surely. In this case,*

$$\mathbb{E}[X^*|X^* \leq 0, Z] = \mu(Z)\left(1 + \frac{(Z'\pi + \kappa(Z))/\mu(Z)}{1 - \exp((Z'\pi + \kappa(Z))/\mu(Z))}\right).$$

We can identify $(Z'\pi + \kappa(Z))/\mu(Z) = \log(1 - F_{X|Z}(0))$ and $\mu(Z) = \mathbb{E}[X|X > 0, Z]$. Thus,

$$\mathbb{E}[X^*|X^* \leq 0, Z] = -\mathbb{E}[X|X > 0, Z]\left(\frac{-\log(1 - F_{X|Z}(0))}{F_{X|Z}(0)} - 1\right).$$

1.3 Semi- and Nonparametric Methods for Discrete/Discretized Z

We discuss how $\mathbb{E}[X^*|X^* \leq 0, Z]$ may be obtained in the semiparametric models of the previous section when Z is discrete or has been discretized. This section supplements the discussion in Section 4.3.1 in the paper. Throughout this section, assume that $\text{supp}(Z)$ is a finite set.

Model 4.3.3 *(Semiparametric Logistic, discrete case) Suppose that Model 4.2.6 holds. Let $\alpha_z = z'\pi + \mu(z)$ and $\sigma_z = \sigma(z)$. This implies that $X^*|Z = z \sim Logistic(\alpha_z, \sigma_z)$. In this case, the two parameters α_z and σ_z can be identified and estimated with the log-likelihood function*

$$\mathcal{L}(\alpha_z, \sigma_z) = -\sum_{i=1}^{n}\mathbf{1}(X_i = 0, Z_i = z)\log(1 + \exp(\alpha_z/\sigma_z)) + \mathbf{1}(X_i > 0, Z_i = z)(X_i - \alpha_z)/\sigma_z$$
$$+ \mathbf{1}(X_i > 0, Z_i = z)\log\sigma_z + 2 \cdot \mathbf{1}(X_i > 0)\log(1 + \exp(-(X_i - \alpha_z)/\sigma_z).$$

Model 4.3.4 *(Semiparametric Uniform, discrete case) Suppose that Model 4.2.7 holds. Let $\alpha_z = z'\pi + \kappa(z)$ and $\nu_z = \mu(z) - \kappa(z)$. This implies that $X^*|Z = z \sim U[\alpha_z, \alpha_z + \nu_z]$. In this case, the two parameters α_z and ν_z can be identified and estimated with the log-likelihood function*

$$\mathcal{L}(\alpha_z, \nu_z) = \sum_{i=1}^{n}\mathbf{1}(X_i = 0, Z_i = z)\log(-\alpha_z) - \mathbf{1}(Z_i = z)\log\nu_z.$$

Model 4.3.5 *(Semiparametric Exponential, discrete case) Suppose that Model 4.2.8 holds. Let* $\alpha_z = z'\pi + \kappa(z)$ *and* $\mu_z = \mu(z)$*. This implies that* $X^* = \alpha_Z + \zeta$ *almost surely, where* $\zeta|Z = z \sim Exp(\mu_z^{-1})$*. In this case, the two parameters* α_z *and* μ_z *can be identified and estimated with the log-likelihood function*

$$\mathcal{L}(\alpha_z, \mu_z) = \sum_{i=1}^{n} \mathbf{1}(X_i = 0, Z_i = z)\log(1-\exp(\alpha_z/\mu_z)) - \mathbf{1}(X_i > 0, Z_i = z)(\log \mu_z + (X_i - \alpha_z)/\mu_z).$$

Model 4.3.6 *(Conditional Symmetry) Suppose that* $F_{X|Z=z}(0) \leq 0.5$*. Assume that the distribution of* $\eta|Z = z$ *is symmetric around its mean, so that letting* $\mathbb{E}[\eta|Z = z] = \mu_z$, $F_{\eta|Z=z}(a) = 1 - F_{\eta|Z=z}(2\mu_z - a)$*. This implies that the distribution of* $X^*|Z = z$ *is also symmetric around its mean/median* $z'\pi + \mu_z$*. Thus, for all* $x < 0$, $F_{X^*|Z=z}(x) = 1 - F_{X^*|Z=z}(2(z'\pi + \mu_z) - x) = 1 - F_{X|Z=z}(2(z'\pi + \mu_z) - x)$*. Because* $F_{X|Z=z}(0) \leq 0.5$, $z'\pi + \mu_z = med(X^*|Z = z) = med(X|Z = z)$ *is identifiable, and thus so is* $F_{X^*|Z=z}(x) = 1 - F_{X|Z=z}(2med(X|Z = z) - x)$ *for all* x*. Therefore, if we calculate the expectation, we can identify the conditional expectation via change of variables as*

$$\mathbb{E}[X^*|X^* \leq 0, Z = z] = 2\,med(X|Z = z) - \mathbb{E}[X|X \geq 2\,med(X|Z = z), Z = z]. \qquad (1)$$

To estimate this quantity, simply substitute the sample equivalents. Note that full symmetry is testable. The distribution of $X^*|Z = z$ is observed in $(0, 2\,med(X|Z = z)]$. Therefore, the equality between the functions $F_{X|Z=z}(x)$ and $1 - F_{X|Z=z}(2\,med(X|Z = z) - x)$ for $x \in [0, med(X|Z = z)]$ can be tested, as discussed in Remark 4.1 in the paper.

2 Two Empirical Checks of the Linearity Assumption

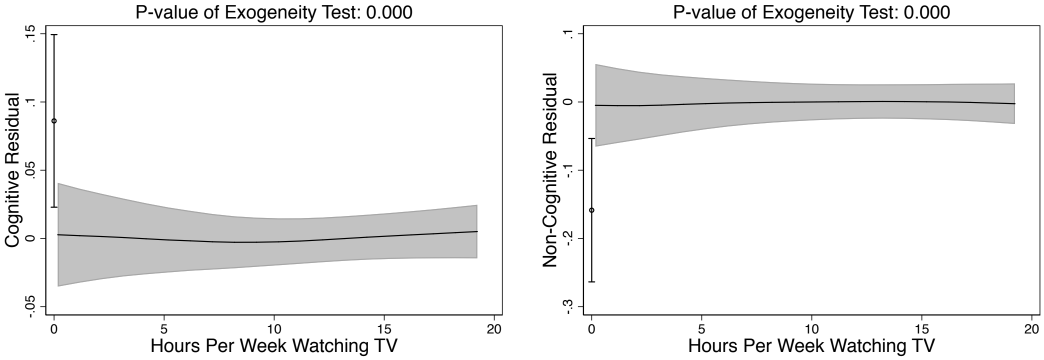
We suggest two checks of whether the main conclusions of our application would change under violations of the linearity assumption in equations (1) and (2). These checks proved to be of practical value both in this application and in Caetano et al. (2020).

2.1 Plotting Residuals from the Uncorrected Regression for $X > 0$

By equation (4) in the paper, when $X > 0$, $\mathbb{E}[Y|X, Z] = X'(\beta + \delta) + Z'(\pi - \gamma\delta)$. The expected value of the residuals of a regression of Y on X and Z for observations with $X > 0$ should thus be equal to zero for each positive value of X. Figure 1 shows the local linear fit of the estimated residuals for cognitive (left panel) and non-cognitive (right panel) skills in our application, which are always close to zero in the positive side.

The points at zero in Figure 1 represent the average of the residuals of the same regression among observations with $X = 0$. This corresponds exactly to an estimator of $\delta\mathbb{E}[X^*|X^* \leq 0]$, as discussed in Remark 2.1 in the paper. The positive and strongly significant result in the left

Figure 1: Evidence that Uncorrected Estimates Are Biased



Note: Each panel shows a plot of the local linear estimator (bandwidth equals to 10) of the residuals from a regression of Y onto X and Z, estimated using only observations with $X > 0$, where X represents the hours spent watching TV in a typical week. We also present the average of the residuals for $X = 0$ and show 90% confidence intervals everywhere. The caption shows the p-value of a test for whether the average residual at $X = 0$ is equal to zero. P-value is calculated using 1,000 bootstrapped samples.

panel shows that $\delta < 0$ for cognitive skills. Analogously, the right panel shows that $\delta > 0$ for non-cognitive skills. Incidentally, this implies a rejection of exogeneity in both cases (p-value shown in the caption of each panel).

2.2 Sequenced Sample Truncation

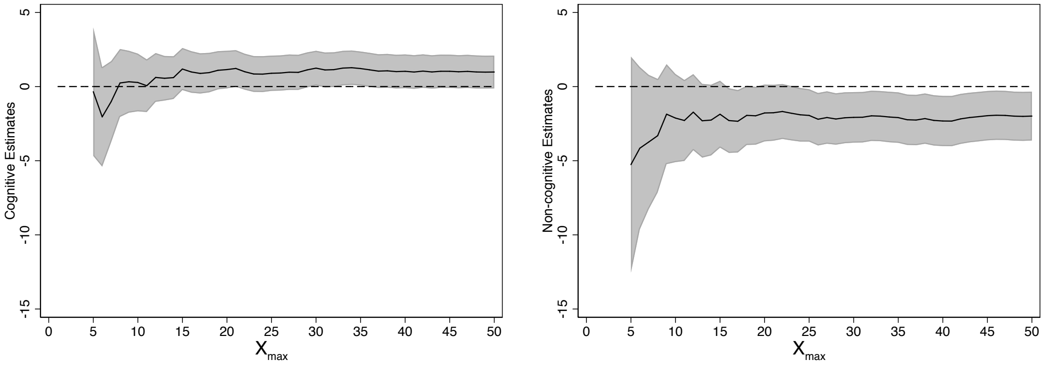
To fix ideas, consider two scenarios. In scenario 1, we restrict our sample to observations with $X \in [0, 1]$, and in scenario 2, we restrict our sample to observations with $X \in [0, 50]$. We exploit the idea that, under violations of the linearity assumption, the effect of the confounders at $X = 0$ is likely to be more similar to the effect of the confounders at $X = 1$ than at $X = 50$. Similarly, the effect of an additional hour of TV at $X = 0$ is more likely to be similar to the effect of an additional hour at $X = 1$ than at $X = 50$.

We build on this idea by first restricting the sample to reflect the first scenario and then progressively expanding the sample until it reaches the second scenario. In Figure 2, we show how our main estimate $\hat{\beta}$ for cognitive (left panel) and non-cognitive skills (right panel) changes for different truncations of our sample, ranging from $X_{\max} = 5$ to $X_{\max} = 50$).[1] Since over 99% of our sample spends less than 50 hours per week watching TV, the estimates in the far right of each panel are almost identical to the corresponding estimates reported in Table 1 of the paper. This approach also allows us to assess the robustness of our conclusions to

[1] We do not include the cases $X_{\max} = 1, 2, 3, 4$ in the plot because the confidence intervals are too large, which makes it hard to visualize what is happening in the rest of the plot. Nevertheless, the unreported results for these cases are consistent with the conclusions we draw from Figure 2. To keep everything else constant irrespective of $X_{\max}$, we keep $\hat{\mathbb{E}}[X^*|X^* \leq 0, Z]$ fixed across the different truncations using our preferred tail symmetry approach. Note that the identification of $\mathbb{E}[X^*|X^* \leq 0, Z]$ does not depend on the assumptions about equations (1) and (2) of the paper, which is what we are trying to test here.

the elimination of television time outliers from our sample.

Figure 2: Estimates for Different Sub-Samples of Data



Note: Each panel shows the estimates of β for cognitive (left panel) or non-cognitive (right panel) skills restricting the sample to only children who watch at most $X_{\max}$ hours of TV per week, for $X_{\max} = 5, \ldots, 50$, with the $X_{\max} = 50$ restriction including over 99% of the sample. The 90% bootstrapped confidence intervals are also shown.

For lower values of $X_{\max}$, the sample is smaller and the confidence intervals are larger, suggesting small-sample variability. Nevertheless, for both types of skills, the qualitative conclusions are stable for all values of $X_{\max}$. For cognitive skills, the point estimates are stable for $X_{max} \geq 15$, while for non-cognitive skills, the point estimates are stable for $X_{max} \geq 8$.

3 Monte Carlo

We assess the performance of our method with an empirical Monte Carlo based on the data and application introduced in Section 5 of the paper. Section 3.1 explains how we calibrate the parameters of the data generating process, Section 3.2 describes how we use these calibrated parameters to draw random samples with different distributions of $\eta|Z$, and Section 3.3 reports and discusses the Monte Carlo results.

3.1 Parameter Calibration

To estimate the parameters, we fit the semiparametric Tobit model (Model 4.3.1 in the paper.) All of the steps enumerated below are carried out on the sample used in Section 5 in the paper with 10 clusters defined as in the application.

1. For each cluster $k = 1, \ldots 10$ we run a Tobit regression of X on a constant. Denote by $\hat{\alpha}_k$ the estimated constant and by $\hat{\sigma}^2_{\eta,k}$ the estimated variance from this Tobit regression. Let $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)'$ be the $k-$th 10×1 canonical vector, then we estimate $\hat{\mathbb{E}}[X^*|X^* \leq 0, \hat{C}_{10} = e_k] = \hat{\alpha}_k - \hat{\sigma}_{\eta,k}\lambda\left(-\frac{\hat{\alpha}_k}{\hat{\sigma}_{\eta,k}}\right)$.

2. We run an OLS regression of Y on X, Z, and $X + \hat{\mathbb{E}}[X^*|X^* \leq 0, \hat{C}_{10} = e_k]\mathbf{1}(X = 0)$, and denote by $\hat{\beta}$, $\hat{a}$ and $\hat{\delta}$ the respective estimated coefficients. We denote by $\hat{\sigma}^2_{\epsilon,k}$ the variance (conditional on $\hat{C}_{10} = e_k$) of the residuals from this regression.

3. For each k, we estimate $\hat{F}_{X|\hat{C}_{10}=e_k}(0)$ by calculating the proportion of observations with $X_i = 0$ among all observations with $\hat{C}_{10i} = e_k$ (i.e., all observations in cluster k). We also estimate $\hat{p}_k = \hat{\mathbb{P}}(\hat{C}_{10} = e_k)$.

4. We estimate the vector $\hat{\gamma} = \hat{a} + \hat{\alpha}\hat{\delta}$, where $\alpha = (\alpha_1, \ldots, \alpha_{10})'$.

5. We estimate $\hat{\sigma}^2_{\eta,H} = \frac{1}{\sum_{k=1}^{10} \hat{p}_k^2} \sum_{k=1}^{10} \hat{p}_k^2 \hat{\sigma}^2_{\eta,k}$, and $\sigma^2_{\eta,k,\text{Mix}} = \hat{\sigma}^2_{\eta,k} - 49$.

6. We calculate

$$\hat{\sigma}^2_{\varepsilon,k} = \hat{\sigma}^2_{\epsilon,k} - \hat{\delta}^2\hat{\sigma}^2_{\eta,k}\left(1 + \frac{\hat{\alpha}_k}{\hat{\sigma}_{\eta,k}}\lambda\left(-\frac{\hat{\alpha}_k}{\hat{\sigma}_{\eta,k}}\right) - \lambda^2\left(-\frac{\hat{\alpha}_k}{\hat{\sigma}_{\eta,k}}\right)\right).$$

3.2 Drawing the Monte Carlo Samples

With the parameters estimated in the previous section in hand, we next produce the Monte Carlo samples using the general scheme: for observation $j \in \{1, 2, \ldots, N\}$:

1. Z_j is a random draw from the entire sample of Z_i's in the data set used in Section 5 of the paper. Suppose that Z_j belongs to the cluster k in the original analysis sample.

2. ε_j is a random draw from a $\mathcal{N}(0, \hat{\sigma}^2_{\varepsilon,k})$ distribution.

3. η_j is a random draw from one of the distributions below, depending on the simulation:[2]

   - Homoskedastic Normal: Draw η_j from a $\mathcal{N}(0, \hat{\sigma}^2_{\eta,H})$ distribution.
   - Heteroskedastic Normal: Draw η_j from a $\mathcal{N}(0, \hat{\sigma}^2_{\eta,k})$ distribution.
   - Heteroskedastic Logistic: Draw η_j from a Logistic$(0, \sqrt{3}\hat{\sigma}_{\eta,k}/\pi)$ where π is the mathematical constant pi.
   - Heteroskedastic Triangular: Let u_j be a draw from the Uniform[0,1] distribution. Then

$$\begin{aligned} \eta_j &= 3\hat{\sigma}_{\eta,k}\left(2\sqrt{u_j} - \sqrt{2}\right), \quad \text{if } u_j < 0.5 \\ &= 3\hat{\sigma}_{\eta,k}\left(\sqrt{2} - 2\sqrt{1 - u_j}\right), \quad \text{if } u_j \geq 0.5 \end{aligned}$$

   - Heteroskedastic Uniform: Draw η_j from a uniform distribution in the interval $\left[-\sqrt{3}\hat{\sigma}_{\eta,k}, \sqrt{3}\hat{\sigma}_{\eta,k}\right]$.

[2] Note that although we are drawing η_j from distributions with zero means, we are not operating under a zero mean assumption on the confounder. The mean of the confounder may be different from zero and is incorporated implicitly through the $\hat{\alpha}_k$ in step 4.

- Heteroskedastic Symmetric Mixture Normal: let u_j be a draw from a uniform[0,1] distribution. Then, $\tilde{\eta}_j^{(1)}$ is a draw from $\mathcal{N}(7, \sigma^2_{\eta,k,\text{Mix}})$ and $\tilde{\eta}_j^{(2)}$ be a draw from $\mathcal{N}(-7, \sigma^2_{\eta,k,\text{Mix}})$. Then, we obtain η_j as follows:

$$\begin{aligned} \eta_j &= \tilde{\eta}_j^{(1)} \text{ if } u_j < 0.5 \\ &= \tilde{\eta}_j^{(2)} \text{ if } u_j \geq 0.5 \end{aligned}$$

4. We calculate $X_j = \max\{0, \hat{\alpha}_k + \eta_j\}$.
5. We calculate $Y_j = \hat{\beta} X_j + Z_j' \hat{\gamma} + \hat{\delta} \eta_j + \varepsilon_j$
6. We keep $(Y_j, X_j, Z_j', \hat{C}_{10j}')$. This is one observation.
7. For each of the M Monte Carlo samples of size N, indexed by r, we estimate $\hat{\beta}_r(s)$ separately for each identification strategy $s \in$ {uncorrected with controls, Tobit, semi-parametric Tobit, and tail symmetry}. We then calculate the bias for each strategy as

$$\text{Bias}(s) = M^{-1} \sum_r (\hat{\beta}_r(s) - \hat{\beta}).$$

where $\hat{\beta}$ is the true value of the parameter in this Monte Carlo, obtained in step 2 of the previous section (this is equivalent to $\hat{\beta}$ in column (iv) in Table 1 of the paper). We calculate the standard deviation as

$$\text{SD}(s) = \sqrt{(M-1)^{-1} \sum_r (\hat{\beta}_r(s) - \bar{\hat{\beta}}(s))^2},$$

where $\bar{\hat{\beta}}(s) = M^{-1} \sum_r \hat{\beta}_r(s)$. We set M = 10,000 and consider three sample sizes: $N = 500$, $N = 1,000$, and $N = 5,000$.

3.3 Monte Carlo Results

Table 1 presents the Monte Carlo results for non-cognitive skills. The results for cognitive skills yield similar conclusions and are therefore omitted for brevity. We report both the average bias and the standard deviation of the Monte Carlo estimates in percentage points of the standard deviation of the outcome variable, the same units as in Table 1 of the paper (Section 5).

Irrespective of the sample size N, and irrespective of the distribution of $\eta|Z$, the bias of the uncorrected approach with controls (column (ii)) is very large. In all rows we strongly reject a t-test that the bias is equal to zero.

In the next columns of the table, we show the results using the correction strategies proposed in the paper under different distributional assumptions on $\eta|Z$. It is immediate that all of the correction strategies yield much smaller biases than the uncorrected strategy irrespective of the sample size and the true distribution of $\eta|Z$.

Table 1: Monte Carlo Results, Non-Cognitive Skills

	(ii) Uncorrected w/ Controls		(iii) Homoskedastic Tobit		(iv) Semiparametric Tobit		(v) Conditional Tail Symmetry	
N=500	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Hom. Normal	3.10**	0.45	0.08	3.50	0.24	3.37	0.42	3.43
Het. Normal	3.10**	0.46	0.03	3.56	0.21	3.39	0.38	3.47
Het. Logistic	3.12**	0.47	-0.85	3.86	-0.59	3.69	0.20	3.15
Het. Triangular	3.41**	0.30	0.16	1.45	0.27	1.42	0.33	1.45
Het. Uniform	3.00**	0.43	1.49	3.01	1.55	2.84	0.84	4.63
Het. Mixt. Norm.	3.07**	0.45	0.44	3.48	0.56	3.32	0.45	3.71
N=1,000	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Hom. Normal	3.09**	0.31	-0.00	2.42	0.08	2.37	0.17	2.38
Het. Normal	3.09**	0.31	-0.02	2.47	0.08	2.40	0.15	2.42
Het. Logistic	3.11**	0.32	-0.78	2.67	-0.64	2.61	0.12	2.18
Het. Triangular	3.40**	0.21	0.16	1.00	0.19	1.00	0.18	1.03
Het. Uniform	2.99**	0.30	1.55	2.08	1.57	2.01	0.60	3.57
Het. Mixt. Norm.	3.06**	0.31	0.36	2.35	0.42	2.30	0.19	2.59
N=5,000	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Hom. Normal	3.09**	0.14	0.02	1.07	0.03	1.06	0.05	1.07
Het. Normal	3.09**	0.14	-0.00	1.09	0.03	1.09	0.05	1.09
Het. Logistic	3.12**	0.14	-0.79	1.17	-0.74	1.16	0.01	0.95
Het. Triangular	3.40**	0.09	0.16	0.44	0.13	0.44	0.03	0.46
Het. Uniform	3.00**	0.13	1.54*	0.91	1.53*	0.89	0.12	1.73
Het. Mixt. Norm.	3.07**	0.13	0.38	1.05	0.38	1.04	0.05	1.18

Note: The values are reported in percentage points of the standard deviation of the outcome variable, as in Table 1 of the paper. Columns (iii), (iv) and (v) show results for corrections using Models 4.2.1, 4.3.1 and 4.3.2 in the paper, respectively. Simulations are based on 10,000 drawn samples of size N using 10 clusters. For each panel representing a different sample size N, each row represents the results from a Monte Carlo assuming a different distribution of $\eta|Z$. **: significant at the 5% level. *: significant at the 10% level.

Although all three correction strategies perform substantially better than the uncorrected strategy, a comparison among them reveals interesting patterns. First, the biases of the three correction methods are almost never significantly different from zero, including all instances in which the distribution assumption is wrong. The only exception is when the true distribution is the uniform and the correction strategies assume normality instead (columns (iii) and (iv)). Second, irrespective of the distribution of $\eta|Z$, the tail symmetry assumption yields biases that decline in magnitude as N grows. Third, both Tobit strategies perform better than the tail symmetry strategy under normality (two first rows in each panel), though the difference diminishes substantially when the sample increases.

Finally, note that the standard errors reported by the uncorrected approach are much smaller than the standard errors reported by any of the corrected approaches. This mirrors what we find in the paper (Table 1 in Section 5).

References

Caetano, C., Caetano, G., and Nielsen, E. (2020). Should children do more enrichment activities? Leveraging bunching to correct for endogeneity. *FEDS Working Paper No. 2020-036*.