Semiparametric IV Regression without Exclusion Restrictions *

Wayne Yuan Gao †
Department of Economics, University of Pennsylvania
and
Rui Wang ‡
Department of Economics, The Ohio State University

January 18, 2024

Abstract

We study identification and estimation of endogenous linear and nonlinear regression models without excluded instrumental variables, based on the standard mean independence condition and a nonlinear relevance condition. We propose two semiparametric estimators as well as a discretization-based estimator that does not require any nonparametric regressions. We establish their asymptotic normality and demonstrate via simulations their robust finite-sample performances with respect to exclusion restrictions violations and endogeneity. Our approach is applied to study the returns to education, and to test the direct effects of college proximity indicators as well as family background variables on the outcome.

Keywords: linear regression, quantile regression, endogeneity, instrumental variable, exclusion restriction, semiparametric two-stage estimation

*We thank Bertille Antoine, Jason Blevins, Xiaohong Chen, Xu Cheng, Stephen Cosslett, Francis Diebold, Keisuke Hirano, Robert de Jong, Lixiong Li, Xiao Lin, Ce Liu, Joris Pinkse, Neslihan Sakarya, Yuya Sasaki, Frank Schorfheide, Petra Todd, Bruce Weinberg, as well as participants at various conferences and seminars for helpful comments and suggestions.

†Department of Economics, University of Pennsylvania, 133 S 36th St, Philadelphia, PA 19104, USA. Email: waynegao@upenn.edu

‡Department of Economics, The Ohio State University, 1945 N High St, Columbus, OH 43210, USA. Email: wang.16498@osu.edu

1 Introduction

The method of instrumental variables (IV) has been a central approach to identify and estimate linear regression models with endogeneity. The conventional IV regression exploits excluded instrumental variables that have no direct effects on the outcome variable. However, finding valid instruments that satisfy the exclusion restriction can be challenging in many applications.

In this paper, we show that even in the absence of excluded instruments, the endogenous linear and (parametric) nonlinear regression models can still be identified by leveraging the nonlinear relevance between the included exogenous regressor and the endogenous variable. In contrast to the traditional IV regression that uses a linear first-stage projection, our approach applies conditional mean projections of the endogenous variables on the included exogenous variables in the first stage. We provide necessary and sufficient conditions for identification of the regression parameters, as well as semiparametric estimation and inference results that accommodate various types of first-stage nonparametric regression methods.

To illustrate, consider the following simple linear regression model:

$$Y_i = \alpha_0 + \beta_0 Z_i + \gamma_0 X_i + \epsilon_i \tag{1}$$

with a scalar endogenous variable X_i and a scalar exogenous variable Z_i satisfying the mean independence condition $\mathbb{E}[\epsilon_i | Z_i] = 0$. Taking conditional expectations of both sides of (1) given $Z_i = z$, we have

$$\mathbb{E}[Y_i | Z_i = z] = \alpha_0 + \beta_0 z + \gamma_0 \pi_0(z). \tag{2}$$

where $\pi_0(Z_i) := \mathbb{E}[X_i | Z_i]$ denotes the conditional mean of the endogenous regressor X_i given the exogenous regressor $Z_i = z$. Based on the simple observation that condition (2) can be viewed as a linear regression of Y_i on 1, Z_i, and $\pi_0(Z_i)$ with *no* endogeneity issue, we see that the model parameters can be identified and estimated provided that $\pi_0(z)$ is nonlinear in z.

More generally, we show that the above identification analysis extends to endogenous

nonlinear and quantile regression. By adopting a mean projection of the nonlinear function onto the included exogenous regressor, we show that local identification is achieved under a full-rank condition. In particular, for quantile regression, we show that the full-rank condition is equivalent to a different nonlinear relevance condition.

This strategy suggests two natural semiparametric least square (TSLS) estimators. Specifically, given $\hat{\pi}$ obtained via first-stage nonparametric regression of X_i on Z_i, we construct our first estimator $\hat{\theta}$ by

$$\text{regressing } Y_i \text{ on } 1,\ Z_i \text{ and } \hat{\pi}_0(Z_i) \quad \text{via OLS.}$$

We also propose an estimator that uses the mean projection $h_0(Z_i) := \mathbb{E}[Y_i|\, Z_i]$ of Y_i on Z_i as the dependent variable. Let $\hat{h}$ denote the estimator of nonparametric regression of Y_i on Z_i, the second estimator $\hat{\theta}^*$ is constructed by

$$\text{regressing } \hat{h}(Z_i) \text{ on } 1,\ Z_i \text{ and } \hat{\pi}_0(Z_i) \text{ via OLS.}$$

The only difference between $\hat{\theta}$ and $\hat{\theta}^*$ lies in the dependent variable used in the second step: $\hat{\theta}$ uses the raw observed variable Y_i, while $\hat{\theta}^*$ uses the fitted value $\hat{h}(Z_i)$ obtained through nonparametric regression of Y_i on Z_i. We propose the second estimator $\hat{\theta}^*$, as it can perform slightly better than $\hat{\theta}$ under some specifications in simulations.

We further propose a third estimator, $\hat{\theta}_{disc}$, which does not require any nonparametric estimation, based on a discretization of the support of Z_i into K (finite and fixed) partitions. Under this discretization, the first-stage estimation simplifies to sample averages in each partition. Furthermore, the estimator can be computed as a standard 2SLS estimator with partition dummies as IVs. While the discretization results in some loss of information and asymptotic efficiency, there is no "discretization bias" in our setting and the number of partition cells K is not required to grow large with the sample size.

We establish the $\sqrt{n}$-consistency of our three proposed estimators $\hat{\theta}$, $\hat{\theta}^*$, and $\hat{\theta}_{disc}$ for θ_0,

along with their asymptotic normality. We show that $\hat{\theta}$ and $\hat{\theta}^*$ share exactly the same asymptotic variance, while that of $\hat{\theta}_{disc}$ is in general different and, when error are homoskedastic, larger under the partial order of positive semi-definiteness.

Monte Carlo simulations support our theoretical results, and demonstrate the good finite-sample performance of the three estimators $\hat{\theta}$, $\hat{\theta}^*$, and $\hat{\theta}_{disc}$ with the presence of violation of the exclusion restriction and endogeneity. For comparison, we also implement the standard 2SLS estimator which treats the included regressor as excluded instrument, as well as the OLS estimator which does not account for endogeneity. The root mean squared error (RMSE) of the three estimators $\hat{\theta}$, $\hat{\theta}^*$, and $\hat{\theta}_{disc}$ are reasonably small, and the coverage probabilities of the 95% confidence intervals are very close to their nominal level, even with a relatively modest sample size of $n = 250$. In contrast, the standard 2SLS estimator has much larger bias and standard errors when the exclusion restriction is violated. As expected, the OLS estimator perform poorly in the presence of endogeneity.

Our approach is applied to study the returns to education and to test the direct effects of different instruments. Our first application, in line with Card (1993), studies two indicators of college proximity: the presence of a nearby 2-year college and a nearby 4-year college. Our findings show that after controlling for regional characteristics, the two college proximity indicators have no significant effects on the outcome. However, the 2SLS estimator varies substantially when using different instruments, while our estimators remain robust under various specifications. In the second application, we investigate two family background variables as potential instruments: parents' education and number of siblings. The results indicate that the number of siblings exerts no significant effect on wages, while parents' education significantly increases income. The estimated returns to education based on our three estimators appear to be smaller than those of 2SLS estimators, as our methods account for the direct effects of the two instruments.

Our paper is closely related to the line of econometric literature on the identification

of endogenous regression models without exclusion restrictions. See Lewbel (2019) for a comprehensive survey of related work on this topic. In the standard linear regression setting, Rigobon (2003), Klein and Vella (2010), Lewbel (2012), and Lewbel (2018) utilize heteroskedasticity of error terms, while Lewbel, Schennach, and Zhang (2023) works with a specific decomposition of error and imposes independence between them. Beyond the standard linear regression setting, Dong (2010) considers a binary response model with imposed independence assumption among error terms, Kolesár et al. (2015) studies a linear regression with "many IVs" under an orthogonality condition between the IVs' direct effects on the outcome variable and the effects on the endogenous covariates, and D'Haultfœuille, Hoderlein, and Sasaki (2021) considers a linear random coefficient model with exogeneity and independence assumptions on the random coefficients. Another relevant paper is Escanciano, Jacho-Chávez, and Lewbel (2016), who studies a more general framework of semiparametric conditional moment models and provides high-level conditions for identification without exclusions. They adopt the control function approach and impose the (full) conditional independence assumption of errors. Furthermore, they work with the moment equation conditional on both the endogenous regressor and the included exogenous regressor. In contrast, our approach is based on the moment equation given only the included exogenous regressor under the mean independence assumption of this regressor. A more recent paper by Tsyawo (2023) also exploits nonlinear relevance for the identification and estimation of endogenous linear regressions without exclusion restrictions, but utilizes the different framework of integrated conditional moment estimators. Our paper adopts a more standard semiparametric estimation and inference framework that accommodates a broad range of first-stage nonparametric regression methods. Moreover, we provide more detailed identification analyses that make explicit the necessary and sufficient conditions for identification, as well as generalizations of the identification analysis to endogenous nonlinear and quantile regression models.

Another relevant line of literature is on optimal IV and asymptotic efficiency in the

estimation of conditional moment restriction models: see, for example, Amemiya (1974, 1977), Chamberlain (1987) and Newey (1990, 1993), Ai and Chen (2003), and Newey (2004). The main focus of this line of literature is on asymptotic efficiency and typically assumes identification as a starting point. Consequently, this literature does not explicitly distinguish between included and excluded IVs or between linear and nonlinear revelance of IVs. In addition, many papers in this literature, such as Donald and Newey (2001), Hahn (2002) and Stock and Yogo (2005), are more concerned with the scenario where there are *many IVs* (which are often implicitly excluded IVs), while we focus on exactly the opposite scenario, where researchers *do not have any* excluded IVs. In addition, Escanciano (2018) considers endogenous linear regressions and proposes the "integrated IV estimator" as a simple and robust alternative to the optimal IV approach. However, the focus of Escanciano (2018) is on robustness (especially with weak instruments), and similarly it does not distinguish between included/excluded IVs or linear/nonlinear relevance of IVs.

In the special case where X_i is binary, there is also a connection between our paper and the literature on heterogeneous treatment effects. This literature, as exemplified by Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996), and Heckman and Vytlacil (2005), studies endogenous selection and instrumental variables within the potential outcome framework. See, e.g., Imbens (2014), Imbens and Rubin (2015), Mogstad and Torgovitsky (2018), and Abadie and Cattaneo (2018) for more comprehensive reviews. This framework allows for nonparametrically heterogeneous treatment effects, but usually imposes full conditional independence assumptions along with monotone relevance conditions on the IVs. Under this framework, the most closely related line of work is on the identification of treatment effects without exclusion restrictions: Manski and Pepper (2000), Flores and Flores-Lagunes (2013), and Mealli and Pacini (2013) establish partial identification without exclusion. Moreover, Hirano et al. (2000) relaxes the exclusion condition by applying the Bayesian approach, while Wang (2022) employs an additional instrument for identification.

The paper also relates to work on endogenous nonlinear and quantile regression models,

such as Newey and Powell (2003), Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007), and Chernozhukov and Hansen (2008). The existing studies explore nonparametric identification with excluded instruments. In contrast, our paper focuses on parametric models and investigates identification using only included exogenous regressors.

Our semiparametric two-stage estimation procedure with nonparametric regression of the endogenous/outcome variables on the included exogenous variables are also reminiscent of Robinson (1988), who considers a partially linear regression model without endogeneity. However, one of the key steps in Robinson (1988) is to transform the regression equation into a "differenced form" that is free of the unknown nonparametric function in the original equation. In contrast, the identification arguments in our linear regression setup does not involve the "differenced form" equation. Antoine and Lavergne (2014) and Antoine and Lavergne (2023) study estimation and inference under weaker identification through conditional moment restrictions. A recent paper by Antoine and Sun (2022) explores the partially linear model with endogenous covariates and also works with the differenced form in the style of Robinson (1988). However, these papers still rely on excluded IVs for identification.

The rest of the paper is organized as follows. Section 2 focuses on endogenous linear regression models, provides identification results and further discussions about the identification conditions, and establish the asymptotic normality of three proposed estimators. Section 3 extends the identification approach to nonlinear and quantile regressions. Section 4 presents simulation results about the finite-sample performances of our estimators. Section 5 studies the returns to education and examines the direct effects of various instruments. We conclude with Section 6.

2 Endogenous Linear Regression without Exclusion

2.1 Model and Identification

Consider the following linear regression model with endogeneity:

$$Y_i = \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0 + \epsilon_i, \tag{3}$$

where X_i is a d_x-dimensional endogenous regressor that can be dependent with ϵ_i, while Z_i is a d_z-dimensional included exogenous regressor satisfying the following mean independence, or strict exogeneity, assumption:

Assumption 1 (Mean Independence)**.** $\mathbb{E}[\epsilon_i | Z_i = z] = 0$ *for any* $z \in \mathcal{Z} := Supp(Z_i)$.

Writing $\theta_0 := \left(\alpha_0, \beta_0', \gamma_0'\right)' \in \mathbb{R}^{d:=1+d_x+d_z}$, we are interested in identifying and estimating θ_0. Section 3 explores the extension of endogenous nonlinear and quantile regressions.

Assumption 1 on Z_i leads to the following conditional moment restriction:

$$\mathbb{E}\left[Y_i - \alpha_0 - Z_i'\beta_0 - X_i'\gamma_0 \middle| Z_i = z\right] = 0, \tag{4}$$

which characterizes the identified set for θ_0. We show that the above restriction can point identify θ_0 under the no multicollinearity condition.

Define $\pi_0(z) := \mathbb{E}[X_i | Z_i = z]$. By employing the mean projection of X_i on Z_i, we can rewrite (4) by replacing the endogenous regressor X_i with $\pi_0(Z_i)$:

$$\mathbb{E}\left[Y_i - \alpha_0 - Z_i'\beta_0 - \pi_0(Z_i)'\gamma_0 \middle| Z_i = z\right] = 0.$$

When treating $\pi_0(Z_i)$ as a regressor, the above condition transforms into the moment restriction of a standard linear regression, which regresses Y_i on 1, Z_i, $\pi_0(Z_i)$. After applying the mean projection, there is no endogeneity since Z_i satisfies the strict exogeneity condition.

Let $W_i := \left(1, Z_i', \pi_0(Z_i)'\right)'$. We then apply the usual identification strategy by premultiplying both sides of the above equation by W_i and then taking unconditional expectations:

$$\mathbb{E}[W_i Y_i] = \mathbb{E}\left[W_i W_i'\right]\theta_0.$$

Since $\pi_0(z)$ is nonparametrically identified from data, the terms $W_i, \mathbb{E}\left[W_i W_i'\right]$, and $\mathbb{E}[W_i Y_i]$ are also identified. It is then clear that θ_0 is identified whenever $\mathbb{E}\left[W_i W_i'\right]$ is invertible, which boils down to the familiar requirement of no multicollinearity condition:

Assumption 2 (No Multicollinearity)**.** $\left(1, Z_i', \pi_0(Z_i)'\right)$ *are not (perfectly) multicollinear. Or equivalently,* $\mathbb{E}\left[W_i W_i'\right]$ *has full rank.*

The discussion regarding Assumption 2 is presented in Section 2.2. Under this assumption, θ_0 is identified as the standard OLS formula with W_i as the regressor:

$$\theta_0 = \left(\mathbb{E}\left[W_i W_i'\right]\right)^{-1}\mathbb{E}[W_i Y_i].$$

Since W_i is a deterministic function of Z_i, we can also project Y_i on Z_i and obtain an alternative expression for θ_0. Defining $h_0(Z_i) := \mathbb{E}[Y_i | Z_i]$, then θ_0 can be expressed as

$$\theta_0 = \left(\mathbb{E}\left[W_i W_i'\right]\right)^{-1}\mathbb{E}[W_i h_0(Z_i)],$$

which follows from the Law of Iterated Expectations. We conduct this additional projection because, through simulation, we find that the estimator based on this formula can exhibit slightly better performance under some specifications.

Theorem 1 (Identification with Included IV)**.** *Under Assumptions 1 and 2,*

$$\theta_0 = \left(\mathbb{E}\left[W_i W_i'\right]\right)^{-1}\mathbb{E}[W_i Y_i] = \left(\mathbb{E}\left[W_i W_i'\right]\right)^{-1}\mathbb{E}[W_i h_0(Z_i)]. \tag{5}$$

Theorem 1 suggests two natural semiparametric two-step estimators for θ_0. Specifically,

given first-stage nonparametric estimators $\hat{\pi}$ for π_0 and $\hat{h}$ for h_0, the second-stage plug-in estimators for θ_0 is given by, with $\hat{W}_i := \left(1, Z_i', \hat{\pi}\left(Z_i\right)'\right)'$,

$$\hat{\theta} := \left(\frac{1}{n}\sum_{i=1}^{n} \hat{W}_i \hat{W}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{W}_i Y_i,$$
$$\hat{\theta}^* := \left(\frac{1}{n}\sum_{i=1}^{n} \hat{W}_i \hat{W}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{W}_i \hat{h}\left(Z_i\right).$$

As shown in Section 2.3, both $\hat{\theta}$ and $\hat{\theta}^*$ are $\sqrt{n}$-consistent and asymptotically normal. Furthermore, they share the same asymptotic variance, and are thus asymptotically equally efficient. In the meanwhile, $\hat{\theta}$ does not need nonparametric estimation of h_0, and is thus simpler and faster to compute than $\hat{\theta}^*$. However, we do find that $\hat{\theta}^*$ can have better finite-sample performance under certain simulation setups. Hence, we keep the estimator $\hat{\theta}^*$ in our paper and provide results for it along with $\hat{\theta}$.

In Section 2.4, we propose a third estimator $\hat{\theta}_{disc}$ based on a discretization of the support of Z_i, which does not require any nonparametric regressions in the first stage. However, $\hat{\theta}_{disc}$ is not directly based on the sample analog of (5). Hence, we defer $\hat{\theta}_{disc}$ to Section 2.4.

2.2 Discussion about Assumption 2

In Section 2.1, we establish point identification of θ_0 from the perspective of the standard linear regression models, which naturally leads to the familiar "no multicollinearity" or "full rank" condition in Assumption 2. Since Assumption 2 is the foundation for the identification of θ_0, we now provide some equivalent, sufficient and necessary conditions for it, along with some more detailed discussions.

We start by stating an equivalent condition for Assumption 2, which also provides a slightly different perspective for identification, where we exploit the fact that the conditional moment equation (4) is a system of *deterministic* linear equations in θ across all $z \in \mathcal{Z}$. Therefore θ_0 is uniquely determined if the following condition holds:

Condition 1 (Full-Dimensional Support)**.** There exist $d = 1 + d_x + d_z$ distinct points $z_1, ..., z_d \in \mathcal{Z}$ such that

$$\text{rank} \begin{pmatrix} 1 & z_1' & \pi_0(z_1)' \\ 1 & z_2' & \pi_0(z_2)' \\ \vdots & \vdots & \vdots \\ 1 & z_d' & \pi_0(z_d)' \end{pmatrix} = d.$$

It turns out that Condition 1 is equivalent to Assumption 2, which is also intuitively so under linearity. Hence, the two perspectives for identification are equivalent.

Lemma 1. *Assumption 2* $\Leftrightarrow$ *Condition 1.*

Condition 1 provides an alternative perspective for identification from the support of the included instrument Z_i, under the feature that $\left(1, Z_i', \pi_0(Z_i)'\right)$ is a deterministic function of Z_i. We see that even though the dimension d_z of the included instrument Z_i is by construction smaller than the number of parameters d (e.g., a scalar Z_i), it is still possible for us to find d linearly independent *realizations* of $\left(1, Z_i', \pi_0(Z_i)'\right)$ on the support of Z_i, which will guarantee the required "no multicollinearity" assumption.

This perspective also motivates our third estimator $\hat{\theta}_{disc}$ by transforming the conditional moment equation into the following unconditional moment equation:

$$\mathbb{E}\left[\left(Y_i - \alpha_0 - Z_i'\beta_0 - X_i'\gamma_0\right)\mathbb{1}\{Z_i \in \mathcal{Z}_k\}\right] = 0,$$

where $(\mathcal{Z}_k)_{k=1}^K$ is a finite partition of the support of Z_i with $K \geq d$. The idea of transforming conditional moments into unconditional ones using instrumental functions such as indicator functions, has been well studied and applied in the literature: e.g., Khan and Tamer (2009b), Andrews and Shi (2013), and Shi, Shum, and Song (2018). See Section 2.4 for more details about the discretization-based estimator $\hat{\theta}_{disc}$.

We now provide a discussion of how Assumption 2 relates to nonlinearity, relevance, and

order conditions:

Condition 2 (No Multicollinearity in Z_i)**.** $\left(1, Z_i^{'}\right)$ are not multicollinear.

Condition 3 (Nonlinearity)**.** $\pi_{0,k}(z) := \mathbb{E}[X_{i,k}|\, Z = z]$ is nonlinear in z on $\mathcal{Z}$, for each component $k = 1, ..., d_x$.

Condition 4 (Relevance)**.** $\pi_{0,k}(z) := \mathbb{E}[X_{i,k}|\, Z = z]$ is not constant in z on $\mathcal{Z}$, for each component $k = 1, ..., d_x$.

Condition 5 (Order Condition on $\mathcal{Z}$)**.** The support of Z_i must contain d distinct points, i.e., $\#(\mathcal{Z}) \geq d = 1 + d_x + d_z$.

Clearly, all of the above are necessary conditions for Assumption 2:

Lemma 2. *(a) Assumption 2 implies Conditions 2 and 3; (b) Condition 3 implies Conditions 4 and 5.*

The no multicollinearity condition and the relevance condition are standard for linear regression models. Condition 3 requires Z_i to be relevant for X_i in a nonlinear manner. This requirement of nonlinearity marks the departure of our approach from the standard IV approach which utilizes a linear projection of X_i on Z_i.

The requirement of nonlinearity also imposes a restriction on the cardinality of the support of Z_i as in Condition 5. This is because it is always possible to fit a straight line between any two distinct points, and more generally, to fit a linear d-dimensional hyperplane across any d distinct points in $\mathbb{R}^d$. Hence, our order condition is on the cardinality of the support of Z_i, rather than the number of variables. Of course, if $\mathcal{Z}$ is a continuum, then the order condition is automatically satisfied.

When there is only one endogenous variable, then the converse of Lemma 2(a) is also true, effectively establishing the sufficiency of nonlinearity for point identification.

Lemma 3 (Sufficient Condition with Scalar X_i)**.** *Suppose that X_i is scalar-valued, i.e. $d_x = 1$. Then, Conditions 2 and 3 $\Rightarrow$ Assumption 2.*

Lemma 3 is particularly relevant when we are primarily worried about the endogeneity of a single treatment status variable X_i, which is often a discrete random variable. Then, if there exists some exogenous shifter Z_i that is relevant for X_i, $\pi_0(z)$ is naturally nonlinear given the discreteness of X_i.

Example 1 (Linear Treatment Effect Model with Selection)**.** Consider

$$Y_i = \alpha_0 + Z_i'\beta_0 + X_i\gamma_0 + \epsilon_i,$$
$$X_i = \mathbb{1}\{\varphi_0(Z_i) \geq u_i\},$$

with $\mathbb{E}[\epsilon_i | Z_i] = 0$, $u_i \perp Z_i$, and $u_i \sim F_u$. Then, the propensity score function $\pi_0(z) := \mathbb{E}[X_i | Z_i = z]$ is naturally nonlinear in z when $\#(Z_i) \geq 3$, i.e., the support of Z_i contains at least three points. As discussed in the introduction, the order condition $\#(Z_i) \geq 3$ can be satisfied even if Z_i just consists of two dummy variables. Hence, Condition 3 can be thought as a mild condition in this setting.

Lastly, we note that, when $d_x > 1$, we not only need each $\pi_{0,k}$ to be nonlinear in z, but also need each $\pi_{0,k}$ to be linearly independent (as a function) from 1, z, and all other $(\pi_{0,j})_{j\neq k}$ as well. We consider this condition relatively mild and easy to verify. Heuristically, whenever $\mathcal{Z}$ is a continuum, the space of functions on $\mathcal{Z}$ (under some regularity conditions) can be often viewed as an infinite-dimensional Hilbert space that admits a linear series representation under a certain orthonormal basis of functions $(b_k(\cdot))_{k=1}^{\infty}$ on $\mathcal{Z}$:

$$\mathcal{F} = \left\{ \sum_{k=1}^{\infty} c_k b_k(\cdot) : \sum_{k=1}^{\infty} c_k^2 < \infty \right\}.$$

Hence, linear independence among a finite number $(d = 1 + d_x + d_z)$ of "generic" functions from $\mathcal{F}$ seems heuristically as a "generic property".

2.3 Semiparametric Estimators $\hat{\theta}$ and $\hat{\theta}^*$

This section derives the asymptotic properties of the following two semiparametric estimators:

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\hat{W}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{W}_i Y_i,$$
$$\hat{\theta}^* = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\hat{W}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\hat{h}\left(Z_i\right).$$

We now lay out the regularity conditions for the $\sqrt{n}$-consistency and asymptotic normality of $\hat{\theta}$ and $\hat{\theta}^*$. The first one is a standard one on the existence of moments.[1]

Assumption 3 (Finite Fourth Moments)**.** $\mathbb{E}\left|\epsilon_i\right|^4, \mathbb{E}\left\|X_i\right\|^4$, *and* $\mathbb{E}\left\|Z_i\right\|^4$ *are finite.*

Below we give some high-level conditions about the first-stage nonparametric regressions, which can be satisfied with a wide variety of lower-level conditions and many types of nonparametric estimators. See, for example, Newey and McFadden (1994) and Chen (2007) for more information.

Assumption 4 (Smoothness and Nonparametric Convergence)**.** *Suppose that:*

(a) $h_0, \pi_0 \in \mathcal{H}$*, where* $\mathcal{H}$ *is a Sobolev function space of order* $s > \frac{d_z}{2}$ *on* $\mathcal{Z}$*.*

(b) The nonparametric estimators $\hat{h}$ *and* $\hat{\pi}$ *belong to* $\mathcal{H}$ *(with probability approaching 1) and are asymptotically linear.*

(c) The nonparametric estimators $\hat{h}$ *for* h_0 *and* $\hat{\pi}$ *for* π_0 *converge in* $L_2(Z)$*-norm faster than the* $n^{-1/4}$ *rate:* $\left\|\hat{h} - h_0\right\|_{L_2(Z)} = o_p\left(n^{-\frac{1}{4}}\right)$, $\left\|\hat{\pi} - \pi_0\right\|_{L_2(Z)} = o_p\left(n^{-\frac{1}{4}}\right)$.

Theorem 2 (Asymptotic Normality)**.** *Under Assumptions 1 - 4, we have:*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V_0\right), \quad \sqrt{n}\left(\hat{\theta}^* - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V_0\right),$$

[1] We impose this assumption on the fourth moment for subsequent variance estimation.

with

$$V_0 := \mathbb{E}\left[W_i W_i^{'}\right]^{-1} \mathbb{E}\left[\epsilon_i^2 W_i W_i^{'}\right] \mathbb{E}\left[W_i W_i^{'}\right]^{-1}.$$

The assumptions about h_0 and $\hat{h}$ can be dropped for $\hat{\theta}$ since it does not involve $\hat{h}$. Also, if $\mathbb{E}[\epsilon_i^2 | \, Z_i] \equiv \sigma_\epsilon^2$ (errors are homoskedastic), V_0 simplifies to $\sigma_\epsilon^2 \mathbb{E}\left[W_i W_i^{'}\right]^{-1}$.

The asymptotic variance can then be easily estimated via standard plug-in methods as in the following theorem. Based on the standard error estimates, confidence intervals and various test statistics can be computed in the standard manner.

Theorem 3 (Variance Estimation)**.** *Let $\hat{\epsilon}_i := Y_i - \hat{\alpha} - Z_i^{'}\hat{\beta} - X_i^{'}\hat{\gamma}$, and*

$$\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} \hat{W}_i \hat{W}_i^{'}, \quad \hat{\Omega} := \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 \hat{W}_i \hat{W}_i^{'}, \quad \hat{V} := \hat{\Sigma}^{-1}\hat{\Omega}\hat{\Sigma}^{-1}.$$

With homoskedasticity, $\hat{V} := \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2\right)\hat{\Sigma}^{-1}$. Under Assumptions 1 - 4, $\hat{V} \xrightarrow{p} V_0$.

2.4 Discretization-Based Estimator $\hat{\theta}_{disc}$

The two estimators $\hat{\theta}$ and $\hat{\theta}^*$ we proposed before both involve nonparametric regressions in the first stage. Alternatively, we propose a third estimator $\hat{\theta}_{disc}$ that does *not* require *any* nonparametric regression at all. Specifically, let $(\mathcal{Z}_k)_{k=1}^K$ be a partition of $\mathcal{Z}$ with K being a finite and fixed number such that $K \geq d = 1 + d_z + d_x$. To rule out redundant cells, we require each cell to have a positive probability.

Assumption 5 (Positive Probabilities)**.** $p_k := \mathbb{P}\left(Z_i \in \mathcal{Z}_k\right) > 0$ *for* $k = 1, ..., K$.

Define the dummy variable for each of the K partition cells as $D_{i,k} := \mathbb{1}\left\{Z_i \in \mathcal{Z}_k\right\}$, and write $D_i := \left(D_{i1}, ..., D_{iK}\right)^{'}$. We can then use D_i as IVs to identify and estimate θ_0 based on the following transformation of equation (3):

$$\mathbb{E}\left[D_i Y_i\right] = \alpha_0 \mathbb{E}\left[D_i\right] + \mathbb{E}\left[D_i Z_i^{'}\right]\beta_0 + \mathbb{E}\left[D_i X_i^{'}\right]\gamma_0.$$

Since Z_i is averaged out within each partition cell, there is some information loss, and the no-multicollinearity condition for identification of θ_0 in Assumption 2 needs to be strengthened to a partitional version. To state the condition in a "lower-level" form, write $\overline{Z}_k := \mathbb{E}[Z_i | Z_i \in \mathcal{Z}_k]$, $\overline{X}_k := \mathbb{E}[X_i | Z_i \in \mathcal{Z}_k]$, and $\overline{W}_k := \left(1, \overline{Z}_k', \overline{X}_k'\right)'$.

Assumption 6 (No Partitional Multicollinearity)**.** *Suppose that* $\left(1, \overline{Z}_k', \overline{X}_k'\right)$ *are not multicollinear across* $k = 1, ..., K$*, or equivalently,*

$$rank \begin{pmatrix} 1 & \overline{Z}_1' & \overline{X}_1' \\ \vdots & \vdots & \vdots \\ 1 & \overline{Z}_K' & \overline{X}_K' \end{pmatrix} = d.$$

We note that Assumption 6 translates into the following standard full-rank condition written in terms of expectations (i.e., probability-weighted sums under discreteness), provided that each cell has a strictly positive probability.

Lemma 4. *Suppose that Assumption 5 holds. Then Assumption 6 holds if and only if* $\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k'$ *is invertible.*

Note that a necessary condition for Assumption 6 is the order condition $K \geq d$ already mentioned above. It is also easy to verify that Assumption 6 implies Assumption 2, but the converse is not generally true. However, Assumption 6 still remains as a condition nonparametrically identified from the observable distribution of data.

We can then construct $\hat{\theta}_{disc}$ as the standard two-stage least square (2SLS) estimator with the K-dimensional vector D_i as instruments. Formally, write $\tilde{W}_i := \left(1, Z_i', X_i'\right)'$, and let $Y, D, \tilde{W}$ denote the vector/matrix concatenation of the variables across all $i = 1, ..., n$, and each row of which contains $Y_i, D_i', \tilde{W}_i'$, respectively. Then

$$\hat{\theta}_{disc} := \left(\tilde{W}' P_D \tilde{W}\right)^{-1} \tilde{W}' P_D Y, \tag{6}$$

where $P_D := D\left(D'D\right)^{-1} D'$. Since D consists of partition cell dummies, the projection

matrix P_D is essentially computing cell-wise averages, and thus $\hat{\theta}_{disc}$ can be equivalently written as

$$\hat{\theta}_{disc} = \left(\sum_{k=1}^{K} \hat{p}_k \hat{\overline{W}}_k \hat{\overline{W}}_k' \right)^{-1} \sum_{k=1}^{K} \hat{p}_k \hat{\overline{W}}_k \hat{\mu}_{y,k}, \tag{7}$$

where $\hat{p}_k := \frac{n_k}{n}$, $n_k := \sum_{i=1}^{n} D_{ik}$, $\hat{\mu}_{y,k} := \frac{1}{n_k} \sum_{i=1}^{n} D_{ik} Y_i$, $\hat{\mu}_{z,k} := \frac{1}{n_k} \sum_{i=1}^{n} D_{ik} Z_i$, $\hat{\mu}_{x,k} := \frac{1}{n_k} \sum_{i=1}^{n} D_{ik} X_i$, and $\hat{\overline{W}}_k := \left(1, \hat{\mu}_{z,k}', \hat{\mu}_{x,k}'\right)'$.

Clearly, $\hat{\theta}_{disc}$ is very easy to compute. Researchers may use any standard 2SLS command with D_i as IVs (or with one of D_{ik}'s dropped if the constant is included), which yields equivalent results as (6) and (7). The asymptotic distribution of $\hat{\theta}_{disc}$ is derived as follows:

Theorem 4 (Asymptotic Normality of $\hat{\theta}_{disc}$)**.** *Under Assumptions 1, 3, 5, and 6,* $\sqrt{n}\left(\hat{\theta}_{disc} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V_{0,disc}\right)$ *with*

$$V_{0,disc} := \left(\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k' \right)^{-1} \sum_{k=1}^{K} p_k \overline{\sigma}^2_{\epsilon,k} \overline{W}_k \overline{W}_k' \left(\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k' \right)^{-1} \tag{8}$$

where $\overline{\sigma}^2_{\epsilon,k} := \mathbb{E}\left[\epsilon_i^2 \middle| Z_i \in \mathcal{Z}_k\right]$. *Furthermore, a consistent estimator* $\hat{V}_{disc}$ *for* $V_{0,disc}$ *can be constructed by plugging* $\hat{p}_k := \frac{n_k}{n}$ *in place of* p_k, $\hat{\overline{W}}_k$ *in place of* $\overline{W}_k$, *and* $\hat{\overline{\sigma}}^2_{\epsilon,k} := \frac{1}{n_k} \sum_{i: Z_i \in \mathcal{Z}_k} (Y_i - \tilde{W}_i' \hat{\theta}_{disc})^2$ *in place of* $\overline{\sigma}^2_{\epsilon,k}$ *in the formula (8) above.*

Under homoskedasticity $\mathbb{E}[\epsilon_i^2 | Z_i \in \mathcal{Z}_k] \equiv \sigma^2_\epsilon$, *the asymptotic variance simplifies to* $V_{0,disc} = \sigma^2_\epsilon \left(\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k' \right)^{-1}$ *with* $\hat{\sigma}^2_\epsilon := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \tilde{W}_i' \hat{\theta}_{disc})^2$ *consistent for* σ^2_ϵ.

While the established results for $\hat{\theta}_{disc}$ hold for any choice of partition $(\mathcal{Z}_k)_{k=1}^K$ that satisfy Assumptions 5 and 6, it is recommended in practice to choose $\mathcal{Z}_k$ in such a way that the cell probabilities p_k are comparable in magnitude. To illustrate, consider a simple example where we partition $\mathcal{Z}$ into $K = 10$ cells with $p_1 = 0.91$ but $p_2 = ... = p_{10} = 0.01$. Despite Assumption 6 holds (so that θ_0 is identified), the estimator $\hat{\theta}_{disc}$ is likely to perform badly, since there are only a few observations in cell $2, ..., K$ and thus the sample average estimation in those cells could be highly imprecise. Moreover, since $p_2, ..., p_{10}$ are close to zero, the smallest eigenvalue of $\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k'$ may be close to zero (unless the corresponding $\overline{W}_k$'s

are very large in magnitude, so that the product terms $p_k \overline{W}_k \overline{W}_k^{'}$ stay comparable across k). Since the estimator $\hat{\theta}_{disc}$ is based on the inverse of $\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k^{'}$, its variance can be large because of the imbalance between p_1 and $p_2, ..., p_{10}$.

If Z_i is a scalar, a natural strategy would be to choose the partition $(\mathcal{Z}_k)$ to be the K equally-sized quantile ranges, which would ensure that $p_k \equiv 1/K$ (or at least asymptotically so when sample quantiles are used in finite samples). If Z_i is vector, one could work with (empirical) vector quantiles as developed relatively recently in the literature based on the theory of optimal transport: see Galichon (2016) for an introduction, and, e.g., Chernozhukov et al. (2017), Hallin et al. (2021), and Ghosal and Sen (2022) for detailed discussions. Alternatively, one could start with a partition of the support of Z_i obtained as products of partitions in each dimension of Z_i, and adjust and/or merge certain cells (if necessary) to ensure that the sample proportions of observations in each cell are comparable across $k = 1, ..., K$.

We emphasize again that our results above apply for any choice of the partition $(\mathcal{Z}_k)$ as long as Assumptions 5 and 6 are satisfied. Hence, while we provide some suggestions for the choice of partitions above, there may be more appropriate partition choices depending on the specific applications and contexts.

3 Extensions: Nonlinear and Quantile Regressions

3.1 Endogenous Nonlinear Regressions

The identification strategy is not limited to linear models and can be also applied to analyze the following endogenous nonlinear regression models without exclusion restrictions:

$$Y_i = f(Z_i, X_i, \theta_0) + \epsilon_i,$$

where X_i is a d_x-dimensional endogenous regressor, Z_i is a d_z-dimensional exogenous regressor satisfying Assumption 1, and the function f is known up to the d-dimensional parameter

θ_0. The function f can be nonlinear and nonseparable in the covariates Z_i and X_i.

To identify θ_0, we adopt a similar strategy by projecting the entire functional term $f(Z_i, X_i, \theta_0)$ onto the included instrument Z_i. Let the function m_0 be defined as

$$m_0(Z_i, \theta_0) := \mathbb{E}[\, f(Z_i, X_i, \theta_0) |\, Z_i],$$

which is identified up to the parameter θ_0. Then under Assumption 1 (exogeneity) of the included instrument Z_i, we have the following conditional moment condition:

$$\mathbb{E}[Y_i - m_0(Z_i, \theta_0) |\, Z_i] = 0.$$

The above moment condition can be viewed as the moment restriction of the standard nonlinear regression model without endogeneity, while treating $m_0(Z_i, \theta_0)$ as the nonlinear regressor. The key distinction is that the function m_0 needs to be estimated. For standard nonlinear regression, local identification of θ_0 can be attained under the following condition.

Theorem 5. *Suppose that Assumption 1 holds, the function $m_0(z, \cdot)$ is continuously differentiable for any z, and $\mathbb{E}\left[\nabla_\theta m_0(Z_i, \theta_0) \nabla_{\theta'} m_0(Z_i, \theta_0)\right]$ has full rank, then θ_0 is locally identified.*

Local identification of θ_0 ensures that there exists a neighborhood Θ_0 of θ_0 on which θ_0 is identified. This result can be expanded to achieve global identification under additional assumptions, by invoking the global inversion theorem in Ambrosetti and Prodi (1995) (Chapter 3, Theorem 1.8). In the case of general nonlinear regressions, the interpretation of the full-rank condition depends on the specific functional form of $f(Z_i, X_i, \theta_0)$. This feature also applies to the standard IV regression with excluded instruments, where the identification conditions are contingent upon the specification of f as well.

Remark 1. *We focus on the nonlinear regression model, while the analysis also applies to*

a more general parametric model. Consider that we have the following moment condition:

$$\tilde{m}_0(Z_i, \theta_0) := \mathbb{E}[g(Y_i, Z_i, X_i, \theta_0) | Z_i] = 0,$$

where the function g is known up to the parameter θ_0 and g can be nonlinear and nonseparable in all variables (Y_i, Z_i, X_i). The moment function g may be derived from structural models, which is naturally nonlinear in all variables. In terms of the nonlinear regression model, the function g is given as $g(Y_i, Z_i, X_i, \theta_0) = Y_i - f(Z_i, X_i, \theta_0)$. Following Theorem 5, the parameter θ_0 is locally identified if $\mathbb{E}[\nabla_\theta \tilde{m}_0(Z_i, \theta_0) \nabla_{\theta'} \tilde{m}_0(Z_i, \theta_0)]$ has full rank. Section 3.2 explores the endogenous quantile regression model, where the moment condition is nonseparable in all observed variables (Y_i, Z_i, X_i).

Similar to linear regression models, we propose two semiparametric two-step nonlinear regression estimators. In the first step, nonparametrically regress $f(X_i, Z_i, \theta)$ on Z_i for each θ and get the predicted value $\hat{m}(Z_i, \theta)$; nonparametrically regress Y_i on Z_i and get the predicted value $\hat{h}(Z_i)$. In the second step, run the standard nonlinear regression using Y_i and $\hat{h}(Z_i)$ as the dependent variable, respectively:

$$\hat{\theta}_{nl} = \arg\min_{\theta \in \Theta_0} \frac{1}{n} \sum_i \left(Y_i - \hat{m}(Z_i, \theta)\right)^2,$$
$$\hat{\theta}^*_{nl} = \arg\min_{\theta \in \Theta_0} \frac{1}{n} \sum_i \left(\hat{h}(Z_i) - \hat{m}(Z_i, \theta)\right)^2.$$

Similar to Section 2.3, the two estimators $\hat{\theta}_{nl}, \hat{\theta}^*_{nl}$ are $\sqrt{n}$-consistent and have the same asymptotic variance:

$$\sqrt{n}\left(\hat{\theta}_{nl} - \theta_0\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, V_0), \quad \sqrt{n}\left(\hat{\theta}^*_{nl} - \theta_0\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, V_0),$$

with

$$V_0 := M_0^{-1} \mathbb{E}\left[\epsilon_i^2 \nabla_\theta m_0(Z_i, \theta_0) \nabla_{\theta'} m_0(Z_i, \theta_0)\right] M_0^{-1},$$

where $M_0 = \mathbb{E}\left[\nabla_\theta m_0(Z_i, \theta_0)\nabla_{\theta'} m_0(Z_i, \theta_0)\right]$. A consistent estimator $\hat{V}$ for the variance matrix V_0 can be developed by replacing ϵ_i with $Y_i - f(Z_i, X_i, \hat{\theta}_{nl})$, $\nabla_\theta m_0(Z_i, \theta_0)$ with its estimator $\nabla_\theta \hat{m}(Z_i, \hat{\theta}_{nl})$, and expectation with the sample mean.

Next, we investigates endogenous quantile regression as an illustration of Theorem 5.

3.2 Endogenous Quantile Regressions

In particular, our result from the previous subsection can be applied to study the following quantile regression model with endogenous regressors:

$$Y_i = \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0 + \epsilon_i, \quad \text{Quan}_\tau(\epsilon_i | \, Z_i) = 0,$$

where $\text{Quan}_\tau(\epsilon_i | \, Z_i)$ denotes the τ-th quantile of the conditional distribution of ϵ_i given Z_i. In this example, X_i is the potentially endogenous regressor and Z_i is the exogenous regressor that satisfies the conditional quantile restriction. We still study the identification and estimation of the coefficient θ_0 using only the included instrument Z_i.

By the quantile exogeneity of Z_i, it yields the conditional moment restriction as follows:

$$\mathbb{E}\left[\mathbb{1}\{Y_i \leq \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0\} - \tau \middle| \, Z_i\right] = 0.$$

The moment condition above is naturally nonlinear and nonseparable in all variables (Y_i, Z_i, X_i). We project the whole indicator term on Z_i and define the function m_0 as

$$m_0(Z_i, \theta_0) := \mathbb{E}\left[\mathbb{1}\{Y_i \leq \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0\} \middle| \, Z_i\right].$$

According to Theorem 5, the coefficient θ_0 is locally identified if $\mathbb{E}\left[\nabla_\theta m_0(Z_i, \theta_0)\nabla_{\theta'} m_0(Z_i, \theta_0)\right]$ has full rank. Lemma 5 presents an alternative condition that is equivalent to this full rank condition, making it easier to interpret.

Assumption 7 (Continuous Errors)**.** *The error term ϵ_i conditional on (x, z) is continuously*

distributed with the density function $f_{\epsilon|X,Z}\left(\epsilon|\, x, z\right)$.

Assumption 7 is a standard assumption that simplifies the calculation of $\nabla_\theta m_0\left(Z_i, \theta_0\right)$.

Lemma 5. *Suppose that Assumption 7 holds and* $f_{\epsilon|Z}(0|\, z) > 0$ *for any* $z \in \mathcal{Z}$, *then* $\mathbb{E}\left[\nabla_\theta m_0(Z_i, \theta_0) \nabla_{\theta'} m_0(Z_i, \theta_0)\right]$ *has full rank if and only if* $1, Z_i$, *and* $\tilde{\pi}_0\left(Z_i\right) := \mathbb{E}\left[X_i|\, Z_i, \epsilon_i = 0\right]$ *are not multicollinear.*

In the case where the endogenous regressor X_i is a scalar, the full-rank condition is equivalent to the nonlinearity of $\tilde{\pi}_0\left(Z_i\right)$. This nonlinearity condition is analogous to the nonlinearity requirement of $\pi_0\left(Z_i\right)$ in the linear regression model, except it is also conditional on $\epsilon_i = 0$. Similarly, this nonlinearity relationship naturally arises when the endogenous regressor X_i is binary or discrete.

Based on the identification results, a natural two-step quantile regression estimator $\hat{\theta}_q$ can be obtained: in the first step, nonparametrically regress $\mathbb{1}\{Y_i \leq \alpha + Z_i'\beta + X_i'\gamma\}$ on Z_i for each θ and compute $\hat{m}(Z_i, \theta)$; in the second step, obtain the quantile estimator $\hat{\theta}_q$ as

$$\hat{\theta}_q := \arg\min_{\theta \in \Theta_0} \frac{1}{n} \sum_i \left(\hat{m}(Z_i, \theta) - \tau\right)^2.$$

Writing $S_i := f_{\epsilon|Z}\left(0|\, Z_i\right)\left(1, Z_i', \tilde{\pi}_0\left(Z_i\right)'\right)'$, the asymptotic distribution of the two-step quantile estimator $\hat{\theta}_q$ is given by

$$\sqrt{n}\left(\hat{\theta}_q - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V_{0,q}\right),$$

with $V_{0,q} := \left(\mathbb{E}\left[S_i S_i'\right]\right)^{-1} \mathbb{E}\left[\left(\mathbb{1}\{\epsilon_i \leq 0\} - \tau\right)^2 S_i S_i'\right] \left(\mathbb{E}\left[S_i S_i'\right]\right)^{-1} = \tau(1-\tau)\left(\mathbb{E}\left[S_i S_i'\right]\right)^{-1}$.

In contrast to the standard quantile regression without endogeneity, our approach allows for potential endogeneity in covariate X_i. While Chernozhukov and Hansen (2005) examines endogenous quantile models with excluded instruments, the key distinction is that our method establishes identification of θ_0 *using solely included regressors*. Our approach can be viewed as leveraging the derivative term $\nabla_\theta m_0(Z_i, \theta_0)$ as an instrumental function, which is

more informative than using Z_i as an instrument since it exploits the dependence between the indicator term $\mathbb{1}\{Y_i \leq \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0\}$ and the included regressor Z_i. Thus, our approach enables identification without exclusion restrictions. On the other hand, our method does require a parametric specification for Y_i, whereas Chernozhukov and Hansen (2005) allows for nonparametric identification with excluded instruments.

4 Simulation

This section examines the finite sample performances of $\hat{\theta}, \hat{\theta}^*$, and $\hat{\theta}_{disc}$, the three estimators proposed in Sections 2. We compare their performance with both the standard 2SLS estimator $\hat{\theta}_{2sls}$, which treats the included regressor Z_i as an excluded instrument, and the OLS estimator $\hat{\theta}_{ols}$ obtained by regressing Y_i on $(1, Z_i', X_i')$. We report four finite-sample performance measures for every estimator: "Bias", "SD" (standard deviation), "RMSE" (root mean squared error), and "CP" (coverage probability of 95% confidence interval). The confidence intervals are constructed using the standard $\pm 1.96 \times$ SE formula, where the standard error estimates SE are obtained based on the asymptotic variance estimators derived in previous sections, all of which allow for heteroskedasticity.[2] The four performance measures are computed based on $B = 2000$ simulations, and we table the performance measures under three sample sizes: $n = 250, 500, 1000$.

4.1 Binary X_i with Two Binary Z_{i1}, Z_{i2}

Our first simulation setup is as follows. In this setup, there are two binary included IVs Z_{i1}, Z_{i2}, randomly generated from $Bernoulli(0.5)$ independently. The endogenous regressor

[2]The standard errors of the OLS estimator are also calculated under heteroskedasticity.

X_i and the outcome variable Y_i are generated by

$$X_i = \mathbb{1}\{2Z_{i1}Z_{i2} + 2(1 - Z_{i1})(1 - Z_{i2}) - 1 \geq u_i\},$$
$$Y_i = \alpha_0 + \beta_{01}Z_{i1} + \beta_{02}Z_{i2} + \gamma_0 X_i + \epsilon_i,$$

where $\alpha_0 = \gamma_0 = 1$. To compare with the 2SLS estimator, which treats (Z_{i1}, Z_{i2}) as excluded instruments, we examine different values of the coefficients of the included regressors (Z_{i1}, Z_{i2}): $\beta_{01} = \beta_{02} = \{1, 0.5, 0\}$. The values of the coefficients (β_{01}, β_{02}) represent the degree of violation of the exclusion restriction, and the 2SLS estimator is only consistent when $\beta_{01} = \beta_{02} = 0$.

The two error terms (ϵ_i, u_i) are drawn, independently from (Z_{i1}, Z_{i2}), from the joint normal distribution with mean $(0, 0)$, variance $(1, 1)$, and correlation parameter ρ, which captures the extent of endogeneity between X_i and ϵ_i. We also consider different levels of endogeneity $\rho \in \{0.5, 0, -0.5\}$, with $\rho = 0$ corresponding to the case with no endogeneity issue (where OLS becomes unbiased and consistent).

Since the instrument $Z_i = (Z_{i1}, Z_{i2})$ is discrete, we use the sample averages to estimate the conditional expectations:

$$\hat{\pi}(z) = \frac{\sum_{i=1}^n X_i \mathbb{1}\{Z_i = z\}}{\sum_{i=1}^n \mathbb{1}\{Z_i = z\}}, \quad \hat{h}(z) = \frac{\sum_{i=1}^n Y_i \mathbb{1}\{Z_i = z\}}{\sum_{i=1}^n \mathbb{1}\{Z_i = z\}}.$$

In this case, the three estimators $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ are numerically equivalent.

Tables 1 and 2 report the performance of the five different estimators for γ_0 under various degrees of exclusion violations (with endogeneity $\rho = 0.5$) and different levels of endogeneity (with $\beta_{01} = \beta_{02} = 1$), respectively.[3] The results demonstrate the robust performance of our estimators in the presence of violations of the exclusion restriction and endogeneity. The root mean squared error (RMSE) of the three estimators are reasonably small, and the coverage probabilities of the 95% confidence intervals are close to the nominal level.

[3]Tables 11-16 in the Online Appendix report the performances of the estimators for the remaining coefficients α_0, β_{01}, and β_{02}.

In contrast, the 2SLS estimator has a very large standard deviation and bias in this simulation setup, even when the exclusion is satisfied $\beta_{01} = \beta_{02} = 0$, due to the small determinant of the matrix $X'P_Z X$. Additionally, the OLS estimator has a very small (close to zero) coverage probability with the presence of endogeneity. Our estimators' advantages become more significant as the sample size n increases due to the fast reduction in both bias and standard deviation, but the 2SLS and OLS estimators remain biased regardless of sample size. Also, the $\sqrt{n}$ convergence rate of our three estimators, are strongly demonstrated by the almost exact 50% reduction in SD and RMSE from $n = 250$ to $n = 1000$.

4.2 Binary X_i with Continuous Z_i

In this subsection, we consider a different DGP in which there is a continuous IV Z_i drawn from $\mathcal{N}(0, 2)$. The variables X_i and Y_i are generated by

$$X_i = \mathbb{1}\{2Z_i \geq u_i\}, \quad Y_i = \alpha_0 + \beta_0 Z_i + \gamma_0 X_i + \epsilon_i,$$

where $\alpha_0 = \gamma_0 = 1$, and we consider three values for $\beta_0 = \{1, 0.5, 0\}$. The error terms (u_i, ϵ_i) are again drawn from the joint normal distribution as in Section 4.1, independently from Z_i, with $\rho = \{0.5, 0, -0.5\}$.

For the two semiparametric estimators $\hat{\theta}$ and $\hat{\theta}^*$, we use the Nadaraya-Watson kernel estimator to nonparametrically estimate π_0 and h_0 in the first stage. We use the standard Gaussian kernel and set the bandwidth based on least square cross validation. We also find that the performances of the final estimators do not change much with other choices of kernels (e.g., an Epanechnikov kernel). For the discretization-based estimator $\hat{\theta}_{disc}$, we partition the support of Z_i into $K = 10$ cells defined by the (empirical) decile ranges. Our results stay similar if K is set to be larger, say, 30.

As shown in Tables 3 and 4 (and Tables 17-20 in the Online Appendix), all our three estimators perform uniformly well across different values of β_0 and ρ. Although the two

Table 1: Bin X with Bin Z_1, Z_2: Performance of $\hat{\gamma}$ (Coef. of X)
Different Degrees of Exclusion Violations

	$n = 250$				$n = 500$				$n = 1000$			
Est	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP
	$\beta_{01} = \beta_{02} = 1$											
$\hat{\theta}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}^*$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{disc}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{2sls}$	-0.826	34.302	34.312	0.889	-3.105	116.242	116.284	0.893	2.563	65.802	65.852	0.878
$\hat{\theta}_{ols}$	-0.485	0.121	0.500	0.024	-0.485	0.088	0.493	0.000	-0.485	0.062	0.489	0.000
	$\beta_{01} = \beta_{02} = 0.5$											
$\hat{\theta}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}^*$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{disc}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{2sls}$	-0.678	17.872	17.885	0.932	-1.759	59.329	59.355	0.917	0.984	32.927	32.942	0.896
$\hat{\theta}_{ols}$	-0.485	0.121	0.500	0.024	-0.485	0.088	0.493	0.000	-0.485	0.062	0.489	0.000
	$\beta_{01} = \beta_{02} = 0$											
$\hat{\theta}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}^*$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{disc}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{2sls}$	-0.530	3.734	3.771	0.991	-0.413	4.745	4.763	0.996	-0.594	3.571	3.620	0.993
$\hat{\theta}_{ols}$	-0.485	0.121	0.500	0.024	-0.485	0.088	0.493	0.000	-0.485	0.062	0.489	0.000

Table 2: Bin X with Bin Z_1, Z_2: Performance of $\hat{\gamma}$ (Coef. of X)
Different Degrees of Endogeneity

	$n = 250$				$n = 500$				$n = 1000$			
Est	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP
	$\rho = 0.5$											
$\hat{\theta}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}^*$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{disc}$	-0.003	0.182	0.182	0.956	0.002	0.137	0.137	0.939	0.005	0.094	0.094	0.952
$\hat{\theta}_{2sls}$	-0.826	34.302	34.312	0.889	-3.105	116.242	116.284	0.893	2.563	65.802	65.852	0.878
$\hat{\theta}_{ols}$	-0.485	0.121	0.500	0.024	-0.485	0.088	0.493	0.000	-0.485	0.062	0.489	0.000
	$\rho = 0$											
$\hat{\theta}$	-0.007	0.181	0.181	0.956	-0.000	0.137	0.137	0.938	0.004	0.093	0.093	0.952
$\hat{\theta}^*$	-0.007	0.181	0.181	0.956	-0.000	0.137	0.137	0.938	0.004	0.093	0.093	0.952
$\hat{\theta}_{disc}$	-0.007	0.181	0.181	0.956	-0.000	0.137	0.137	0.938	0.004	0.093	0.093	0.952
$\hat{\theta}_{2sls}$	-1.315	32.234	32.261	0.904	-0.301	38.531	38.532	0.894	0.362	74.036	74.036	0.877
$\hat{\theta}_{ols}$	-0.002	0.127	0.127	0.948	-0.001	0.091	0.091	0.946	0.002	0.064	0.064	0.942
	$\rho = -0.5$											
$\hat{\theta}$	-0.011	0.182	0.183	0.954	-0.003	0.138	0.138	0.937	0.003	0.093	0.093	0.950
$\hat{\theta}^*$	-0.011	0.182	0.183	0.954	-0.003	0.138	0.138	0.937	0.003	0.093	0.093	0.950
$\hat{\theta}_{disc}$	-0.011	0.182	0.183	0.954	-0.003	0.138	0.138	0.937	0.003	0.093	0.093	0.950
$\hat{\theta}_{2sls}$	0.913	59.738	59.745	0.900	1.477	58.951	58.969	0.882	0.576	74.167	74.169	0.872
$\hat{\theta}_{ols}$	0.483	0.124	0.499	0.023	0.485	0.087	0.492	0.002	0.485	0.062	0.489	0.000

semiparametric estimators $\hat{\theta}, \hat{\theta}^*$ involve nonparametric regressions in the first stage, they have reasonably good performances even with a small sample size ($n = 250$), with the corresponding CI coverage probabilities for γ_0 close to their nominal level 95%. In contrast, the 2SLS and OLS estimators are significantly biased under exclusion restriction violation and endogeneity, and their CI coverage probabilities are close to zero for all the sample sizes.

We also find that, the discretization-based estimator $\hat{\theta}_{disc}$ performs (surprisingly) well in finite samples. While the two semiparametric estimators $\hat{\theta}$ and $\hat{\theta}^*$ perform well overall, their small-sample biases induced by the first-stage nonparametric regressions are fairly noticeable when compared to that of $\hat{\theta}_{disc}$, especially in the estimation of γ_0 under $n = 250$. In contrast, the loss of asymptotic efficiency in $\hat{\theta}_{disc}$ seems to be fairly small and more than compensated by its smaller finite-sample bias.

5 Empirical Applications

We apply our methodology to examine the returns to education, a topic of substantial attention in the literature (see, e.g., Card (2001) for a review of various studies on this topic). A key concern in investigating the causal impact of education is its potential endogeneity, and it is challenging to find valid instruments that are excluded from the model. Our approach allows us to include all potential instruments in the regression model and test their validity.

We conduct two applications and test the direct effects of different instruments for education. The first application explores the college proximity indicators, which are proposed in Card (1993). We find that after controlling for regional characteristics, the presence of a nearby college does not significantly affect income. However, the 2SLS estimator varies substantially with different instruments and can become insignificant, while our estimators remain more robust regardless of the choice of the instruments. In the second application, we examine the validity of family background variables as instruments. Our findings show that the number of siblings has no significant effect on wages rates, while parents' education

Table 3: Bin X with Cts Z: Performance of $\hat{\gamma}$ (Coef. of X)
Different Degrees of Exclusion Violations

	$n = 250$				$n = 500$				$n = 1000$			
Est	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP
	$\beta_0 = 1$											
$\hat{\theta}$	0.044	0.326	0.329	0.942	0.036	0.220	0.223	0.954	0.024	0.155	0.156	0.948
$\hat{\theta}^*$	-0.099	0.306	0.321	0.942	-0.072	0.209	0.221	0.948	-0.059	0.148	0.159	0.942
$\hat{\theta}_{disc}$	-0.031	0.321	0.323	0.950	-0.016	0.223	0.223	0.955	-0.014	0.161	0.161	0.951
$\hat{\theta}_{2sls}$	5.165	0.298	5.174	0.000	5.159	0.200	5.163	0.000	5.165	0.147	5.167	0.000
$\hat{\theta}_{ols}$	-0.477	0.198	0.516	0.322	-0.485	0.140	0.505	0.060	-0.486	0.097	0.495	0.000
	$\beta_0 = 0.5$											
$\hat{\theta}$	0.044	0.326	0.329	0.942	0.036	0.220	0.223	0.954	0.024	0.155	0.156	0.948
$\hat{\theta}^*$	-0.150	0.293	0.329	0.930	-0.114	0.204	0.234	0.936	-0.092	0.146	0.173	0.906
$\hat{\theta}_{disc}$	-0.031	0.321	0.323	0.950	-0.016	0.223	0.223	0.955	-0.014	0.161	0.161	0.951
$\hat{\theta}_{2sls}$	2.584	0.206	2.592	0.000	2.578	0.140	2.582	0.000	2.582	0.102	2.584	0.000
$\hat{\theta}_{ols}$	-0.477	0.198	0.516	0.322	-0.485	0.140	0.505	0.060	-0.486	0.097	0.495	0.000
	$\beta_0 = 0$											
$\hat{\theta}$	0.044	0.326	0.329	0.942	0.036	0.220	0.223	0.954	0.024	0.155	0.156	0.948
$\hat{\theta}^*$	-0.287	0.310	0.422	0.828	-0.220	0.223	0.313	0.802	-0.168	0.160	0.232	0.782
$\hat{\theta}_{disc}$	-0.031	0.321	0.323	0.950	-0.016	0.223	0.223	0.955	-0.014	0.161	0.161	0.951
$\hat{\theta}_{2sls}$	0.003	0.164	0.164	0.948	-0.002	0.113	0.113	0.954	-0.001	0.081	0.081	0.950
$\hat{\theta}_{ols}$	-0.477	0.198	0.516	0.322	-0.485	0.140	0.505	0.060	-0.486	0.097	0.495	0.000

Table 4: Bin X with Cts Z: Performance of $\hat{\gamma}$ (Coef. of X)
Different Degrees of Endogeneity

	$n = 250$				$n = 500$				$n = 1000$			
Est	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP	Bias	SD	RMSE	CP
	$\rho = 0.5$											
$\hat{\theta}$	0.044	0.326	0.329	0.942	0.036	0.220	0.223	0.954	0.024	0.155	0.156	0.948
$\hat{\theta}^*$	-0.099	0.306	0.321	0.942	-0.072	0.209	0.221	0.948	-0.059	0.148	0.159	0.942
$\hat{\theta}_{disc}$	-0.031	0.321	0.323	0.950	-0.016	0.223	0.223	0.955	-0.014	0.161	0.161	0.951
$\hat{\theta}_{2sls}$	5.165	0.298	5.174	0.000	5.159	0.200	5.163	0.000	5.165	0.147	5.167	0.000
$\hat{\theta}_{ols}$	-0.477	0.198	0.516	0.322	-0.485	0.140	0.505	0.060	-0.486	0.097	0.495	0.000
	$\rho = 0$											
$\hat{\theta}$	0.074	0.325	0.334	0.933	0.056	0.218	0.225	0.948	0.035	0.153	0.157	0.943
$\hat{\theta}^*$	-0.093	0.305	0.319	0.946	-0.065	0.207	0.218	0.953	-0.056	0.147	0.158	0.943
$\hat{\theta}_{disc}$	0.001	0.325	0.325	0.953	0.002	0.222	0.222	0.958	-0.005	0.161	0.161	0.954
$\hat{\theta}_{2sls}$	5.165	0.298	5.173	0.000	5.158	0.198	5.161	0.000	5.165	0.147	5.167	0.000
$\hat{\theta}_{ols}$	-0.004	0.204	0.204	0.938	-0.003	0.146	0.146	0.940	-0.003	0.101	0.101	0.948
	$\rho = -0.5$											
$\hat{\theta}$	0.108	0.323	0.340	0.917	0.074	0.217	0.230	0.937	0.046	0.153	0.160	0.936
$\hat{\theta}^*$	-0.081	0.302	0.312	0.955	-0.060	0.206	0.214	0.959	-0.054	0.147	0.156	0.947
$\hat{\theta}_{disc}$	0.038	0.321	0.324	0.952	0.019	0.222	0.223	0.960	0.004	0.161	0.161	0.952
$\hat{\theta}_{2sls}$	5.163	0.294	5.172	0.000	5.157	0.196	5.161	0.000	5.165	0.147	5.167	0.000
$\hat{\theta}_{ols}$	0.481	0.197	0.519	0.298	0.477	0.139	0.497	0.068	0.479	0.098	0.489	0.002

significantly increases wages.

5.1 Application I: College Proximity Indicators

We use the same dataset as in Card (1993), drawn from the National Longitudinal Survey of Young Men (NLSYM). This data contains information of $n = 3010$ male observations in 1976, documenting their educational attainment, wage, race, age, and assorted demographic characteristics. Card (1993) proposes to use the presence of a 4-year college as an instrument for education, which is likely to affect an individual's educational attainment but may not have direct effects on their earnings. However, Card (1993) also raises a potential concern with this instrument, as the presence of a college might be correlated with superior school quality and, consequently, could lead to higher earnings. We study two specifications that investigate one of the two indicators of college proximity respectively: the presence of a nearby 2-year college (nearc2) and the presence of a nearby 4-year college (nearc4).

Following Card (1993), the dependent variable is the log of hourly wage in 1976, the endogenous variable is education, and the control variables include experience, experience squared, a black indicator, indicators for southern residence and residence in an SMSA in 1976, and indicators for region in 1966 and living in an SMSA in 1966. Distinct from Card (1993), our approach also includes the college proximity instrument in the model and allows for testing its direct effect on the outcome by examining the significance of the coefficient. Table 5 presents the summary statistics of the primary variables.

We present the results of six different estimators. The first two estimators $\hat{\theta}, \hat{\theta}^*$ are introduced in Section 2.3. For the estimation of $\hat{\pi}(Z_i), \hat{h}(Z_i)$, we employ the Support Vector Machine (SVM) method, a broadly applied machine learning technique for high-dimensional regressors. The neural network approach is also implemented, yielding similar results and the same significance of all coefficients.[4] For the discretization estimator $\hat{\theta}_{disc}$, we divide the experience variable into three partitions using empirical quantiles and generate dummy

[4]We adopt the function 'svm' from the e1071 package and the function 'neuralnet' from the neuralnet package in R to implement the SVM approach and the neural network method.

Table 5: Summary Statistics with College Proximity Indicators

	mean	s.d.	minimum	maximum
log(wage) in 1976	6.262	0.444	4.605	7.785
education	13.263	2.677	1.000	18.000
experience	8.856	4.142	0.000	23.000
experience squared	95.579	84.618	0.000	529.000
black	0.234	0.423	0.000	1.000
nearc2	0.441	0.497	0.000	1.000
nearc4	0.682	0.466	0.000	1.000
live in SMSA in 1966	0.650	0.477	0.000	1.000
live in SMSA in 1976	0.713	0.452	0.000	1.000
live in South in 1976	0.404	0.491	0.000	1.000

Notes: the experience variable is constructed using the conventional meansure: experience=age-education-6.

variables for each partition. Then we construct the instrument for education using the product of any two indicator variables. The estimator $\hat{\theta}_{ols}$ is the OLS estimator that includes the instrument in the regression, while $\tilde{\theta}_{ols}$ represents the OLS estimator that does not include the instrument. The last one $\hat{\theta}_{2sls}$ is the 2SLS estimator using the college proximity indicator as the excluded instrument for education.

Table 6 and Table 7 display the outcomes of the coefficients for education and the college proximity instruments using SVM and neural network methods. The results from our three estimators $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ demonstrate that, after controlling for all the regional factors in 1966 and 1976, having a nearby 2-year college or 4-year college has no significant effects on wages. This finding supports the validity of using college proximity indicators as excluded instruments. The standard deviations of the instrument coefficients with $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ are very close to that of $\hat{\theta}_{ols}$, reinforcing the good performance of these estimators.

The estimated returns to education from $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ are uniformly positive and significant under various specifications and nonparametric estimation methods. Moreover, the standard deviations of the education coefficients, derived from the three estimators, are also reasonably

small across different specifications and are smaller than the one obtained from the 2SLS estimator. The coefficients on education from $\hat{\theta}, \hat{\theta}_{disc}$ are all higher than those from the two OLS estimators, suggesting that the OLS estimators may underestimate education's impact. The coefficient from $\hat{\theta}^*$ can be lower than $\hat{\theta}_{ols}$, as it uses a different dependent variable. Overall, the three estimators $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ yield very similar results, which corroborates our theoretical results on their asymptotic variances.

For the 2SLS estimator $\hat{\theta}_{2sls}$, the coefficients on education vary substantially when using the two different instruments. It becomes insignificant when using the presence of a nearby 2-year college as an instrument, due to the large standard deviation. In addition, the estimated coefficients on education from $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ are uniformly smaller than the one obtained from the 2SLS estimator, since our three estimators allow for the direct effect of the instrument.

Table 6: Returns to Education: College Proximity Indicators (SVM)

	education	nearc2	education	nearc4
	nearc2		nearc4	
$\hat{\theta}$	0.083**	0.028	0.078**	0.019
	(0.012)	(0.015)	(0.012)	(0.017)
$\hat{\theta}^*$	0.067**	0.024	0.065**	0.015
	(0.012)	(0.015)	(0.012)	(0.017)
$\hat{\theta}_{disc}$	0.092**	0.029	0.082**	0.019
	(0.014)	(0.015)	(0.013)	(0.017)
$\hat{\theta}_{ols}$	0.075**	0.027	0.074**	0.018
	(0.004)	(0.015)	(0.004)	(0.017)
$\tilde{\theta}_{ols}$	0.075**	-	0.075**	-
	(0.004)		(0.004)	
$\hat{\theta}_{2sls}$	0.293	-	0.132**	-
	(0.186)		(0.054)	

Notes: the symbol ** denotes significant coefficients from zero at 95% level.

Table 7: Returns to Education: College Proximity Indicators (Neural Network)

	education	nearc2	education	nearc4
	nearc2		nearc4	
$\hat{\theta}$	0.087**	0.024	0.099**	0.012
	(0.018)	(0.015)	(0.017)	(0.017)
$\hat{\theta}^*$	0.081**	0.030	0.099**	0.012
	(0.018)	(0.015)	(0.017)	(0.018)
$\hat{\theta}_{disc}$	0.092**	0.029	0.082**	0.019
	(0.014)	(0.015)	(0.013)	(0.017)
$\hat{\theta}_{ols}$	0.075**	0.027	0.074**	0.018
	(0.004)	(0.015)	(0.004)	(0.017)
$\tilde{\theta}_{ols}$	0.075**	-	0.075**	-
	(0.004)		(0.004)	
$\hat{\theta}_{2sls}$	0.293	-	0.132**	-
	(0.186)		(0.054)	

Notes: the symbol ** denotes significant coefficients from zero at 95% level.

5.2 Application II: Family Background Variables

As data about nearby colleges might not always be available, family background variables are often used as instruments for education. In this application, we explore two family background variables as instruments: parents' average education and number of siblings. To conduct this study, we utilize the dataset 'NLSY79', which conducts interviews of $n = 10800$ young individuals, both male and female, ranging in age from 14 to 21 in 1979. This survey records various characteristics of the individuals, such as gender, marriage status, work-related factors, region indicators, as well as family background variables. Table 8 displays the summary statistics of the key variables.

We compare our three estimators $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ with two OLS estimators and three 2SLS estimators. We still apply both the SVM and the neural network methods to estimate $\hat{\pi}(Z_i)$ and $\hat{h}(Z_i)$. For the discretization estimator, we construct dummy variables for experience, hour, parents' average education, and number of siblings, based on whether each variable is above the median. The included instruments for education are then constructed using the

Table 8: Summary Statistics with Family Background Variables

	mean	s.d.	minimum	maximum
log(wage)	2.780	0.604	0.756	5.284
education	13.678	2.475	0.000	20.000
female	0.500	0.500	0.000	1.000
black	0.100	0.300	0.000	1.000
marriage	0.652	0.476	0.000	1.000
experience	16.977	4.373	0.827	23.808
hour	40.831	8.925	10.000	60.000
live in North-Central	0.325	0.468	0.000	1.000
live in North-Eastern	0.162	0.368	0.000	1.000
live in Southern	0.360	0.480	0.000	1.000
parents' average education	11.703	2.738	0.000	20.000
number of siblings	3.165	2.139	0.000	17.000

Notes: parents' average education is computed by (mother's education+father's education)/2.

product of any two variables. For the two OLS estimators, $\hat{\theta}_{ols}$ includes the two family background variables, while $\tilde{\theta}_{ols}$ does not. Additionally, we evaluate three 2SLS estimators. The first one $\hat{\theta}^{both}_{2sls}$ uses both parents' education and number of siblings as excluded instruments for education. The second estimator $\hat{\theta}^{edu}_{2sls}$ employs only parents' education as an instrument, while the third one $\hat{\theta}^{sib}_{2sls}$ utilizes solely the number of siblings.

Table 9 and Table 10 display the results of the eight estimators. The findings from the three estimators $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ show that the number of siblings does not have significant effects on wages, whereas parents' education significantly increases income. This result is consistent with our intuitive reasoning, as parents' education could influence an individual's wage by creating a more favorable educational environment. The three 2SLS estimators appear to overestimate the returns to education, especially the two estimators $\hat{\theta}^{both}_{2sls}, \hat{\theta}^{edu}_{2sls}$ which involve using parents' education as instruments. The three estimators $\hat{\theta}, \hat{\theta}^*, \hat{\theta}_{disc}$ all have significantly positive coefficients on education, and their results are smaller than those of the three 2SLS estimators, given that they control for direct effects of parents' education.

Table 9: Returns to Education: Family Background Variables (SVM)

	education	parents' education	number of siblings
$\hat{\theta}$	0.116**	0.014**	0.001
	(0.004)	(0.002)	(0.002)
$\hat{\theta}^*$	0.101**	0.019**	-0.001
	(0.004)	(0.002)	(0.002)
$\hat{\theta}_{disc}$	0.109**	0.019**	-0.004
	(0.012)	(0.008)	(0.005)
$\hat{\theta}_{ols}$	0.103**	0.019**	0.002
	(0.002)	(0.002)	(0.002)
$\tilde{\theta}_{ols}$	0.113**	-	-
	(0.002)		
$\hat{\theta}^{both}_{2sls}$	0.147**	-	-
	(0.005)		
$\hat{\theta}^{edu}_{2sls}$	0.149**	-	-
	(0.005)		
$\hat{\theta}^{sib}_{2sls}$	0.123**	-	-
	(0.009)		

Notes: the symbol ** denotes significant coefficients from zero at 95% level.

Table 10: Returns to Education: Family Background Variables (Neural Network)

	education	parents' education	number of siblings
$\hat{\theta}$	0.109**	0.017**	0.004
	(0.006)	(0.003)	(0.002)
$\hat{\theta}^{*}$	0.081**	0.029**	-0.001
	(0.006)	(0.003)	(0.002)
$\hat{\theta}_{disc}$	0.109**	0.019**	-0.004
	(0.012)	(0.008)	(0.005)
$\hat{\theta}_{ols}$	0.103**	0.019**	0.002
	(0.002)	(0.002)	(0.002)
$\tilde{\theta}_{ols}$	0.113**	-	-
	(0.002)		
$\hat{\theta}^{both}_{2sls}$	0.147**	-	-
	(0.005)		
$\hat{\theta}^{edu}_{2sls}$	0.149**	-	-
	(0.005)		
$\hat{\theta}^{sib}_{2sls}$	0.123**	-	-
	(0.009)		

Notes: the symbol ** denotes significant coefficients from zero at 95% level.

6 Conclusion

This paper offers an alternative approach to identify and estimate endogenous regression models in the absence of excluded instruments. The key idea of this approach is to leverage the nonlinear dependence between the included exogenous regressor and the endogenous variable. For estimation, we introduce two semiparametric estimators and a easy-to-compute discretization-based estimator. The asymptotic properties of all three estimators are derived and their robust finite sample performances are demonstrated through Monte Carlo simulations. As an application, we apply the approach to study returns to education, and to test the direct effects of college proximity indicators as well as family background variables.

References

Abadie, A. and M. D. Cattaneo (2018): "Econometric methods for program evaluation," *Annual Review of Economics*, 10, 465–503.

Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

Ambrosetti, A. and G. Prodi (1995): *A primer of nonlinear analysis*, 34, Cambridge University Press.

Amemiya, T. (1974): "The nonlinear two-stage least-squares estimator," *Journal of econometrics*, 2, 105–110.

——— (1977): "The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model," *Econometrica: Journal of the Econometric Society*, 955–968.

Andrews, D. W. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 91, 444–455.

Antoine, B. and P. Lavergne (2014): "Conditional moment models under semi-strong identification," *Journal of Econometrics*, 182, 59–69.

——— (2023): "Identification-robust nonparametric inference in a linear IV model," *Journal of Econometrics*, 235, 1–24.

Antoine, B. and X. Sun (2022): "Partially linear models with endogeneity: a conditional moment-based approach," *The Econometrics Journal*, 25, 256–275.

Card, D. (1993): "Using geographic variation in college proximity to estimate the return to schooling," .

——— (2001): "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica*, 69, 1127–1160.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restriction," *Journal of Econometrics*, 34, 3.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry (2017): "Monge–Kantorovich depth, quantiles, ranks and signs," *The Annals of Statistics*, 223–256.

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261.

——— (2008): "Instrumental variable quantile regression: A robust inference approach," *Journal of Econometrics*, 142, 379–398.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139, 4–14.

D'Haultfœuille, X., S. Hoderlein, and Y. Sasaki (2021): "Testing and relaxing the exclusion restriction in the control function approach," *Journal of Econometrics*.

Donald, S. G. and W. K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161–1191.

Dong, Y. (2010): "Endogenous regressor binary choice models without instruments, with an application to migration," *Economics Letters*, 107, 33–35.

Escanciano, J. C. (2018): "A simple and robust estimator for linear regression models with strictly exogenous instruments," *The Econometrics Journal*, 21, 36–54.

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016): "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7, 561–589.

Flores, C. A. and A. Flores-Lagunes (2013): "Partial identification of local average treatment effects with an invalid instrument," *Journal of Business & Economic Statistics*, 31, 534–545.

Galichon, A. (2016): *Optimal transport methods in economics*, Princeton University Press.

Ghosal, P. and B. Sen (2022): "Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing," *The Annals of Statistics*, 50, 1012–1037.

Hahn, J. (2002): "Optimal inference with many instruments," *Econometric Theory*, 18, 140–168.

Hallin, M., E. Del Barrio, J. Cuesta-Albertos, and C. Matrán (2021): "Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach," *The Annals of Statistics*, 49, 1139–1165.

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 73, 669–738.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000): "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics*, 1, 69–88.

Honoré, B. E. and L. Hu (2020): "Selection without exclusion," *Econometrica*, 88, 1007–1029.

——— (2022): "Sample selection models without exclusion restrictions: Parameter heterogeneity and partial identification," *Journal of Econometrics*.

Imbens, G. (2014): "Instrumental variables: an econometrician's perspective," Tech. rep., National Bureau of Economic Research.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica: journal of the Econometric Society*, 467–475.

Imbens, G. W. and D. B. Rubin (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Khan, S. and E. Tamer (2009a): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152, 104–119.

——— (2009b): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152, 104–119.

Klein, R. and F. Vella (2010): "Estimating a class of triangular simultaneous equations models without exclusion restrictions," *Journal of Econometrics*, 154, 154–164.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015): "Identification and inference with many invalid instruments," *Journal of Business & Economic Statistics*, 33, 474–484.

Lewbel, A. (2012): "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models," *Journal of Business & Economic Statistics*, 30, 67–80.

——— (2018): "Identification and estimation using heteroscedasticity without instruments: The binary endogenous regressor case," *Economics Letters*, 165, 10–12.

——— (2019): "The identification zoo: Meanings of identification in econometrics," *Journal of Economic Literature*, 57, 835–903.

Lewbel, A., S. M. Schennach, and L. Zhang (2023): "Identification of a triangular two equation system without instruments," *Journal of Business & Economic Statistics*, 1–35.

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

Mealli, F. and B. Pacini (2013): "Using secondary outcomes and covariates to sharpen inference in instrumental variable settings," *Journal of the American Statistical Association*, 108, 1120–1131.

Mogstad, M. and A. Torgovitsky (2018): "Identification and extrapolation of causal effects with instrumental variables," *Annual Review of Economics*, 10, 577–613.

Newey, K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Newey, W. K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica: Journal of the Econometric Society*, 809–837.

——— (1993): "16 Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics*, Elsevier, vol. 11, 419–454.

——— (2004): "Efficient semiparametric estimation via moment restrictions," *Econometrica*, 72, 1877–1897.

Newey, W. K. and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578.

Rigobon, R. (2003): "Identification through heteroskedasticity," *Review of Economics and Statistics*, 85, 777–792.

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," *Econometrica: Journal of the Econometric Society*, 931–954.

Shi, X., M. Shum, and W. Song (2018): "Estimating semi-parametric panel multinomial choice models using cyclic monotonicity," *Econometrica*, 86, 737–761.

Stock, J. and M. Yogo (2005): "Asymptotic distributions of instrumental variables statistics with many instruments," *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 6, 109–120.

Tsyawo, E. S. (2023): "Feasible IV regression without excluded instruments," *The Econometrics Journal*, 26, 235–256.

Wang, R. (2022): "Point Identification of LATE with Two Imperfect Instruments," *Working Paper*.

A Discussion

In this section, we provide some additional discussion about how our identification relates to the following two approaches for identification without exclusion restrictions: (i) the Heckman correction approach and (ii) the approach of using a known nonlinear instrumental function.

A.1 Comparison: Heckman Correction Approach

The conventional Heckman correction approach can also achieve identification without exclusion restrictions, under distributional assumptions or parametric functions. This approach typically focuses on a binary endogenous regressor and examines the following specification:

$$\begin{aligned} Y_i &= \alpha_0 + Z_i^{'}\beta_0 + X_i\gamma_0 + \epsilon_i, \\ X_i &= \mathbb{1}\left\{Z_i'\eta_0 \geq u_i\right\}, \\ (\epsilon_i, u_i)' &\sim \mathcal{N}\left((0,0)', (1, \rho_0; \rho_0, 1)\right). \end{aligned}$$

Under the joint distribution of the two error terms (ϵ_i, u_i), it yields the following conditional moment restriction:

$$\mathbb{E}[Y_i|\, X_i, Z_i] = \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0 + \rho_0 \frac{\phi(Z_i'\eta_0)}{\Phi(Z_i'\eta_0)},$$

which can identify (θ_0, ρ_0) without exclusion restrictions. The Heckman correction approach can be extended to nonbinary and multi-dimensional endogenous regressor X_i. We can still look at the conditional expectation of Y_i given all regressors (X_i, Z_i):

$$\mathbb{E}[Y_i|\, X_i, Z_i] = \alpha_0 + Z_i'\beta_0 + X_i'\gamma_0 + \mathbb{E}[\epsilon_i|\, X_i, Z_i].$$

If a parametric form on the selection bias term $\mathbb{E}[\epsilon_i|\, X_i, Z_i]$ is imposed as follows:

$$\mathbb{E}[\epsilon_i|\, X_i, Z_i] = s(X_i, Z_i, \eta_0),$$

and this function s is nonlinear in (X_i, Z_i), then the coefficient θ_0 is identified.

The Heckman correction approach exploits the mean projection of the error ϵ_i on all regressors (X_i, Z_i). To achieve identification, this approach requires a parametric form (or parametric distributions of errors) for s as well as the nonlinearity of s. However, since the

function s involves the unobserved error term ϵ_i, its nonlinearity cannot be directly tested.

In contrast, our approach applies the mean projection of the endogenous X_i on the exogenous Z_i. The identification relies on the nonlinearity of the function $\pi_0(Z_i) = \mathbb{E}[X_i|\, Z_i]$, but does not require further functional form assumption on π_0. Moreover, the function π_0 only depends on observed variables (X_i, Z_i), making its nonlinearity a testable condition.

A.2 Relationship to the Instrumental Function Approach

We establish identification of θ_0 from the viewpoint of a standard linear regression model, while treating $\pi_0(Z_i)$ as an exogenous regressor. Point identification is then obtained under the no-multicollinearity condition, which translates into a nonlinearity requirement on π_0. Another approach to address endogeneity is to use a (known) nonlinear function $g(Z_i)$ as an instrument for the endogenous regressor X_i. In this section, we will discuss the connections between our identification approach and this alternative approach.

To illustrate, consider the case where both Z_i and X_i are scalar variables. Using $(1, Z_i, g(Z_i))$ as IVs, we can obtain the following moment restrictions for θ_0:

$$\mathbb{E}\left[(Y_i - \alpha_0 - \beta_0 Z_i - \gamma_0 X_i)\begin{pmatrix} 1 \\ Z_i \\ g(Z_i) \end{pmatrix}\right] = 0.$$

The parameter θ_0 is identified from the above equation if

$$H_g := \begin{bmatrix} 1 & \mathbb{E}[Z_i] & \mathbb{E}[X_i] \\ \mathbb{E}[Z_i] & \mathbb{E}[Z_i^2] & \mathbb{E}[X_i Z_i] \\ \mathbb{E}[g(Z_i)] & \mathbb{E}[Z_i g(Z_i)] & \mathbb{E}[X_i g(Z_i)] \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}[Z_i] & \mathbb{E}[\pi_0(Z_i)] \\ \mathbb{E}[Z_i] & \mathbb{E}[Z_i^2] & \mathbb{E}[\pi_0(Z_i) Z_i] \\ \mathbb{E}[g(Z_i)] & \mathbb{E}[Z_i g(Z_i)] & \mathbb{E}[\pi_0(Z_i) g(Z_i)] \end{bmatrix}$$

has full rank, which depends on the functional form of π_0 and the choice of g.

Clearly, a necessary condition for the full rank requirement is the nonlinearity of π_0.

Otherwise, the third column of H_g will be a linear combination of the first two columns,[5] and hence H_g cannot have full rank, irrespective of the choice of g. Our identification results thus make explicit the dependence of the identifiability on the nonlinearity of the π_0 function.

Moreover, it is worth noting that not all nonlinear functions can serve as valid IVs for X_i in the sense of satisfying the full rank condition on H_g. For example, if $Z_i \sim \mathcal{N}(0,1)$, $\pi_0(z) = z^3$, and $g(z) = z^2$, then H_g has deficient rank, and thus Z_i^2 is not a valid IV.

Our identification results in Theorem 1 can be interpreted as using $\pi_0(Z_i)$ as an instrument for the endogenous regressor X_i, which is an unknown function that can be identified from data. As shown in Chamberlain (1987) and Newey (1990), π_0 is in fact the optimal instrument under homoskedasticity. The identification results in our paper, Lemma 3 in particular, further imply that, with $\pi_0(Z_i)$ used as the IV, the nonlinearity of π_0 becomes *sufficient* for the full rank condition. Therefore, $\pi_0(Z_i)$ is not only the instrumental function that minimizes the asymptotic variance under homoskedasticity, but also the instrumental function that requires the minimum assumption for identification.

B Proofs

B.1 Proof of Lemma 1

Proof. We first prove that Condition 1 $\Longrightarrow$ Assumption 2 by contradiction. Suppose that Assumption 2 fails. Then there exists $c \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ s.t. $\left(1, \pi_0'(z), z'\right) c = 0,\ \forall z \in \mathcal{Z}$. For any distinct $z_1, ..., z_d \in \mathcal{Z}$, define

$$A(z_1, ..., z_d) := \begin{pmatrix} 1 & z_1' & \pi_0(z_1)' \\ 1 & z_2' & \pi_0(z_2)' \\ \vdots & \vdots & \vdots \\ 1 & z_d' & \pi_0(z_d)' \end{pmatrix}, \quad r(z_1, ..., z_d) := \operatorname{rank}\left(A(z_1, ..., z_d)\right).$$

[5] Writing $\pi_0(Z_i) = a + bZ_i$, we have $\mathbb{E}[\pi_0(Z_i)] = a + b\mathbb{E}[Z_i]$, $\mathbb{E}[\pi_0(Z_i)Z_i] = a\mathbb{E}[Z_i] + b\mathbb{E}[Z_i^2]$, and $\mathbb{E}[\pi_0(Z_i)g(Z_i)] = a\mathbb{E}[g(Z_i)] + b\mathbb{E}[Z_i g(Z_i)]$.

We have $A(z_1, ..., z_d)c = \mathbf{0} \Rightarrow r(z_1, ..., z_d) < d$.

We now prove that Assumption 2 $\Longrightarrow$ Condition 1. Suppose that Assumption 2 holds, i.e., $\left(1, Z_i^{'}, \pi_0\left(Z_i\right)^{'}\right)$ are not multicollinear. This means that for any $c \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, there must exist some $z \in \mathcal{Z}$ s.t.

$$\left(1, z^{'}, , \pi_0\left(z\right)^{'}\right) c \neq 0. \tag{9}$$

If $\#\left(\mathcal{Z}\right) = K < d$, then (9) cannot true. Hence $\#\left(\mathcal{Z}\right) \geq d$. For any d distinct points $z_1, ..., z_d$, if $r(z_1, ..., z_d) = d$, then we are done. If $r(z_1, ..., z_d) < d$, then there exists some $c \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ s.t. $A(z_1, ..., z_d)c = \mathbf{0}$. Now, by (9) there must exists $z_{d+1} \in \mathcal{Z}$ s.t. $\left(1, z_{d+1}^{'}, \pi_0\left(z_{d+1}\right)^{'}\right) c \neq 0$, which implies that $\left(1, z_{d+1}^{'}, \pi_0\left(z_{d+1}\right)^{'}\right)$ is linearly independent from $\left\{\left(1, z_k^{'}, \pi_0\left(z_k\right)^{'}\right) : k = 1, .., d\right\}$ and thus $r(z_1, ..., z_d, z_{d+1}) = r(z_1, ..., z_d) + 1$. If $r(z_1, ..., z_d, z_{d+1}) = d$, we stop; otherwise we can repeat the argument above and find some $z_{d+2} \in \mathcal{Z}$ such that $r\left(z_1, ..., z_{d+2}\right) = r\left(z_1, ..., z_{d+1}\right) + 1$. This recursion must stop at most $k^* \leq d - r\left(z_1, ..., z_d\right)$ steps, with $r\left(z_1, ..., z_{d+k^*}\right) = d$. Then we pick d distinct points from $\{z_1, ..., z_{d+k^*}\}$ such that its rank is d, which is precisely Condition 1. □

B.2 Notation for Asymptotic Theory

We first formally set up our notation. For any $\theta = \left(\alpha, \beta^{'}, \gamma^{'}\right)^{'} \in \mathbb{R}^d$ and any functions $h : \mathbb{R}^{d_z} \to \mathbb{R}$ and $\pi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$, define

$$g^*\left(z; \theta, h, \pi\right) := h\left(z\right) - w\left(z, \pi\right)^{'} \theta, \quad g\left(y, z; \theta, \pi\right) := y - w\left(z, \pi\right)^{'} \theta$$

with $w\left(z, \pi\right) := \left(1, z^{'}, \pi\left(z\right)^{'}\right)^{'}$ so that

$$g^*\left(Z_i; \theta_0, h_0, \pi_0\right) = h_0\left(Z_i\right) - w\left(Z_i, \pi_0\right)^{'} \theta_0 \equiv 0,$$
$$g\left(Y_i, Z_i; \theta_0, \pi_0\right) = Y_i - w\left(Z_i, \pi_0\right)^{'} \theta_0 = \epsilon_i + u_i^{'}\gamma_0,$$

where $u_i := X_i - \mathbb{E}[X_i|Z_i]$ and

$$\mathbb{E}[g(Y_i, Z_i; \theta_0, h_0, \pi_0)|Z_i] = \mathbb{E}\left[\epsilon_i + u_i'\gamma_0 \middle| Z_i\right] = 0.$$

We construct the following quadratic population criterion function:

$$Q^*(\theta, h, \pi) := \frac{1}{2}\mathbb{E}\left[g^{*2}(Z_i; \theta, h, \pi)\right], \quad Q(\theta, \pi) := \frac{1}{2}\mathbb{E}\left[g^2(Y_i, Z_i; \theta, \pi)\right],$$

so that $Q^*(\theta_0; h_0, \pi_0) = 0$, and $Q(\theta_0; \pi_0) = \mathbb{E}\left[\left(\epsilon_i + u_i'\gamma_0\right)^2\right]$.

Corollary 1. *Under Assumptions 1-2, θ_0 is the unique minimizer of $Q^*(\cdot, h_0, \pi_0)$ and $Q(\cdot, \pi_0)$, i.e., $\theta_0 = \arg\min_{\theta\in\mathbb{R}^d} Q^*(\theta, h_0, \pi_0) = \arg\min_{\theta\in\mathbb{R}^d} Q(\theta, \pi_0)$.*

Proof. Note that $Q^*(\theta, h_0, \pi_0) = \frac{1}{2}\mathbb{E}[g^{*2}(Z_i; \theta, h_0, \pi_0)] = 0$ implies $g^*(Z_i; \theta, h_0, \pi_0) = 0$ almost surely, and thus $\mathbb{E}[w(Z_i, \pi_0) g^*(Z_i; \theta, h_0, \pi_0)] = \mathbb{E}\left[W_i W_i'\right]\theta - \mathbb{E}[W_i h_0(Z_i)] = \mathbf{0}$, which implies that $\theta = \theta_0$. In the meanwhile, the first-order condition for the minimization of $Q(\theta, \pi_0) = \frac{1}{2}\mathbb{E}[g^2(Y_i, Z_i; \theta, \pi_0)]$ is given by $\mathbb{E}[w(Z_i, \pi_0) g(Y_i, Z_i; \theta_0, \pi_0)] = \mathbb{E}\left[W_i\left(Y_i - W_i'\theta_0\right)\right] = \mathbf{0}$, which is equivalent to $\mathbb{E}\left[W_i W_i'\right]\theta_0 - \mathbb{E}[W_i Y_i] = \mathbf{0}$. □

B.3 Proof of Theorem 2

Proof. It is well known that a Sobolev space of order $s > \frac{d_z}{2}$ is a Donsker class of functions. Since the residual functions in our setup are given by

$$w(Z_i, \pi) g^*(Z_i, \theta, h, \pi) = \left(h(Z_i) - \alpha - Z_i'\beta - \pi(Z_i)'\gamma\right)\begin{pmatrix} 1 \\ Z_i \\ \pi(Z_i) \end{pmatrix},$$

$$w(Z_i, \pi) g(Y_i, Z_i, \theta, \pi) = \left(Y_i - \alpha - Z_i'\beta - \pi(Z_i)'\gamma\right)\begin{pmatrix} 1 \\ Z_i \\ \pi(Z_i) \end{pmatrix},$$

which are smooth functions of (θ, h, π), the function classes

$$\mathcal{F}^* := \left\{ w(\cdot, \pi)\, g^*(\cdot, \theta, h, \pi) - w(\cdot, \pi_0)\, g^*(\cdot, \theta, h_0, \pi_0) : \theta \in \mathbb{R}^d, h \in \mathcal{H}, \pi \in \mathcal{H} \right\},$$
$$\mathcal{F}^* := \left\{ w(\cdot, \pi)\, g(\cdot, \cdot, \theta, \pi) - w(\cdot, \pi_0)\, g(\cdot, \cdot, \theta, \pi_0) : \theta \in \mathbb{R}^d, \pi \in \mathcal{H} \right\},$$

are also Donsker, and thus satisfy the stochastic equicontinuity condition.

We then proceed to derive the influence functions for $\hat{\theta}^*$ and $\hat{\theta}$ separately.

(a) For $\hat{\theta}^*$, recall that $g^*(z; \theta, h, \pi) = h(z) - \alpha - z'\beta - \pi(z)'\gamma$ with $g^*(Z_i; \theta_0, h_0, \pi_0) \equiv 0$. Hence, $\nabla_\theta Q^*(\theta, h_0, \pi_0) = -\mathbb{E}\left[w(Z_i, \pi_0)\, g^*(Z_i; \theta, h_0, \pi_0)\right]$, with $\nabla_\theta Q^*(\theta_0, h_0, \pi_0) = -\mathbb{E}[W_i 0] = \mathbf{0}$ and $\nabla_{\theta\theta} Q^*(\theta, h_0, \pi_0) = \mathbb{E}\left[w(Z_i, \pi_0)\, w(Z_i, \pi_0)'\right] = \mathbb{E}\left[W_i W_i'\right] = \Sigma_0$. Furthermore,

$$\begin{aligned}
& D_{(h,\pi)}\left[\nabla_\theta Q^*(\theta_0, h_0, \pi_0), h - h_0, \pi - \pi_0\right] \\
:= & \lim_{t \searrow 0} \frac{1}{t} \left(\nabla_\theta Q^*(\theta_0, h_0 + t(h - h_0), \pi_0 + t(\pi - \pi_0)) - \nabla_\theta Q^*(\theta, h_0, \pi_0) \right) \\
= & -\lim_{t \searrow 0} \frac{1}{t} \mathbb{E}\left[t \begin{pmatrix} h - h_0 - (\pi - \pi_0)'\gamma_0 \\ Z_i \left(h - h_0 - (\pi - \pi_0)'\gamma_0 \right) \\ \pi_0 \left(h - h_0 - (\pi - \pi_0)'\gamma_0 \right) + (\pi - \pi_0)\, g^*(\cdot, \theta_0, h_0, \pi_0) \end{pmatrix}_{Z_i} \right] \\
= & -\mathbb{E}\left[\begin{pmatrix} h - h_0 - (\pi - \pi_0)'\gamma_0 \\ Z_i \left(h - h_0 - (\pi - \pi_0)'\gamma_0 \right) \\ \pi_0 \left(h - h_0 - (\pi - \pi_0)'\gamma_0 \right) \end{pmatrix}_{Z_i} \right],
\end{aligned}$$

where the subscript Z_i means that all the functions $h, h_0, \pi, \pi_0, g(\cdot, \theta_0, h_0, \pi_0)$ are evaluated at Z_i. and the last equality uses the observation that $g^*(z, \theta_0, h_0, \pi_0) \equiv 0$.

Define

$$\psi^*(Y_i, X_i, Z_i) := -\begin{pmatrix} Y_i - h_0(Z_i) - (X_i - \pi_0(Z_i))' \gamma_0 \\ Z_i \left(Y_i - h_0(Z_i) - (X_i - \pi_0(Z_i))' \gamma_0\right) \\ \pi_0(Z_i) \left(Y_i - h_0(Z_i) - (X_i - \pi_0(Z_i))' \gamma_0\right) \end{pmatrix}$$
$$= -\begin{pmatrix} \epsilon_i \\ Z_i \epsilon_i \\ \pi_0(Z_i) \epsilon_i \end{pmatrix} = -\epsilon_i W_i$$

since, by (3) and (4), $Y_i - h_0(Z_i) - (X_i - \pi_0(Z_i))' \gamma_0 = \epsilon_i$. Hence,

$$\mathbb{E}\left[\psi^*(Y_i, X_i, Z_i)\right] = -\mathbb{E}\left[\mathbb{E}\left[\epsilon_i | Z_i\right] W_i\right] = \mathbf{0},$$

and $\mathbb{E}\left[\left\|\psi^*(Y_i, X_i, Z_i)\right\|^2\right] < \infty$. Noting that $g^*(Z_i, \theta_0, h_0, \pi_0) \equiv 0$, then by the standard theory for semiparametric two-stage estimation, e.g. Theorems 8.1 & 8.2 of Newey and McFadden (1994), we have

$$\sqrt{n}\left(\hat{\theta}^* - \theta_0\right) = -\Sigma_0^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(-g^*(Z_i, \theta_0, h_0, \pi_0) W_i + \psi^*(Y_i, X_i, Z_i)\right) + o_p(1)$$
$$= -\Sigma_0^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi^*(Y_i, X_i, Z_i) + o_p(1) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, V_0 = \Sigma_0^{-1} \Omega_0 \Sigma_0^{-1}\right)$$

with $\Omega_0 = \mathbb{E}\left[\psi^*(Y_i, X_i, Z_i) \psi^*(Y_i, X_i, Z_i)'\right] = \mathbb{E}\left[\epsilon_i^2 W_i W_i'\right]$.

(b) For $\hat{\theta}$, recall $g(y, z; \theta, \pi) := y - w(z, \pi)' \theta$ and $g(Y_i, Z_i; \theta_0, \pi_0) = \epsilon_i + u_i' \gamma_0$. Hence,

$$\nabla_\theta Q(\theta, \pi_0) = -\mathbb{E}\left[w(Z_i, \pi_0) g(Y_i, Z_i; \theta, \pi_0)\right],$$

with $\nabla_\theta Q(\theta_0, \pi_0) = -\mathbb{E}\left[w(Z_i, \pi_0)\left(\epsilon_i + u_i' \gamma_0\right)\right] = \mathbf{0}$ and

$$\nabla_{\theta\theta} Q(\theta, \pi_0) = \mathbb{E}\left[w(Z_i, \pi_0) w(Z_i, \pi_0)'\right] = \mathbb{E}\left[W_i W_i'\right] = \Sigma_0.$$

Furthermore,

$$
\begin{aligned}
& D_{\pi}\left[\nabla_{\theta}Q\left(\theta_{0},\pi_{0}\right),\pi-\pi_{0}\right] \\
:= & \lim_{t\searrow 0}\frac{1}{t}\left(\nabla_{\theta}Q\left(\theta_{0},\pi_{0}+t\left(\pi-\pi_{0}\right)\right)-\nabla_{\theta}Q\left(\theta,\pi_{0}\right)\right) \\
= & -\lim_{t\searrow 0}\frac{1}{t}\mathbb{E}\left[t\begin{pmatrix} -\left(\pi-\pi_{0}\right)'\gamma_{0} \\ -Z_{i}\left(\pi-\pi_{0}\right)'\gamma_{0} \\ -\pi_{0}\left(\pi-\pi_{0}\right)'\gamma_{0}+\left(\pi-\pi_{0}\right)g\left(Y_{i},Z_{i},\theta_{0},\pi_{0}\right)\end{pmatrix}_{Z_{i}}\right] \\
= & \mathbb{E}\left[\begin{pmatrix} \left(\pi-\pi_{0}\right)'\gamma_{0} \\ Z_{i}\left(\pi-\pi_{0}\right)'\gamma_{0} \\ \pi_{0}\left(\pi-\pi_{0}\right)'\gamma_{0}\end{pmatrix}_{Z_{i}}\right]
\end{aligned}
$$

where the last equality follows from the Law of Iterated Expectations and

$$\mathbb{E}\left[g\left(Y_{i},Z_{i},\theta_{0},\pi_{0}\right)\middle|\,Z_{i}\right]=\mathbb{E}\left[\epsilon_{i}+u_{i}'\gamma_{0}\middle|\,Z_{i}\right]=0.$$

Defining

$$\psi\left(Y_{i},X_{i},Z_{i}\right):=\begin{pmatrix} \left(X_{i}-\pi_{0}\left(Z_{i}\right)\right)'\gamma_{0} \\ Z_{i}\left(X_{i}-\pi_{0}\left(Z_{i}\right)\right)'\gamma_{0} \\ \pi_{0}\left(Z_{i}\right)\left(X_{i}-\pi_{0}\left(Z_{i}\right)\right)'\gamma_{0}\end{pmatrix}=u_{i}'\gamma_{0}W_{i},$$

we have $\mathbb{E}\left[\psi\left(Y_{i},X_{i},Z_{i}\right)\right]=\mathbb{E}\left[W_{i}\mathbb{E}\left[u_{i}'\middle|\,Z_{i}\right]\gamma_{0}\right]=\mathbf{0}$. Again, based on the standard results for the asymptotic theory of semiparametric two-stage estimators, such as The-

orems 8.1 & 8.2 of Newey and McFadden (1994), we have

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}-\theta_0\right) &= -\Sigma_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(-g\left(Z_i,\theta_0,\pi_0\right)W_i+\psi\left(Y_i,X_i,Z_i\right)\right)+o_p\left(1\right)\\ &= -\Sigma_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(-\left(\epsilon_i+u_i^{'}\gamma_0\right)W_i+u_i^{'}\gamma_0 W_i\right)+o_p\left(1\right)\\ &= \Sigma_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i W_i+o_p\left(1\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},V_0=\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}\right).\end{aligned}$$

□

B.4 Proof of Theorem 3

Proof. Given finite fourth moment in Assumption 3, we have

$$\frac{1}{n}\sum_{i=1}^{n}W_iW_i^{'}\xrightarrow{p}\Sigma_0,\quad \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2W_iW_i^{'}\xrightarrow{p}\Omega_0.$$

Moreover, given the consistency of the first-stage nonparametric estimator $\hat{\pi}$ in Assumption 4 and the consistency of estimators $\hat{\theta}$ in Theorem 2, we have

$$\begin{aligned}\hat{W}_i-W_i &= \left(0,\mathbf{0}',\hat{\pi}\left(Z_i\right)'-\pi_0\left(Z_i\right)'\right)'\xrightarrow{p}\mathbf{0},\\ \hat{\epsilon}_i-\epsilon_i &= \alpha_0-\hat{\alpha}-Z_i^{'}\left(\hat{\beta}-\beta_0\right)-X_i^{'}\left(\hat{\gamma}-\gamma_0\right)\xrightarrow{p}0,\end{aligned}$$

and thus

$$\begin{aligned}\hat{\Sigma}-\Sigma_0 &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{W}_i\hat{W}_i^{'}-W_iW_i^{'}\right)+\frac{1}{n}\sum_{i=1}^{n}W_iW_i^{'}-\Sigma_0\xrightarrow{p}\mathbf{0},\\ \hat{\Omega}-\Omega_0 &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\epsilon}_i^2\hat{W}_i\hat{W}_i^{'}-\epsilon_iW_iW_i^{'}\right)+\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2W_iW_i^{'}-\Omega_0\xrightarrow{p}\mathbf{0}.\end{aligned}$$

Hence, $\hat{V}:=\hat{\Sigma}^{-1}\hat{\Omega}\hat{\Sigma}^{-1}\xrightarrow{p}V_0=\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}$. □

B.5 Discussion about the Asymptotic Variance of $\hat{\theta}_{disc}$

Since $\hat{\theta}_{disc}$ is constructed based on averages over each partition cell $\mathcal{Z}_k$, there is in general some information loss, and thus $\hat{\theta}_{disc}$ tends to be less efficient than $\hat{\theta}$ and $\hat{\theta}^*$. Our next result formalizes this efficiency loss in the setting where ϵ_i is homoskedastic.

Theorem 6. *Suppose that $\mathbb{E}[\epsilon_i^2|\, Z_i] \equiv \sigma_\epsilon^2$. Then $V_{0,disc} - V_0$ is positive semi-definite.*

Proof. Under homoskedasticity, the formulas for V_0 and $V_{0,disc}$ simplify to

$$V_0 = \sigma_\epsilon^2 \left(\mathbb{E}\left[W_i W_i^{'}\right]\right)^{-1}, \quad V_{0,disc} = \sigma_\epsilon^2 \left(\sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k^{'}\right)^{-1}.$$

Recalling that $\overline{W}_k = \mathbb{E}[W_i|\, Z_i \in \mathcal{Z}_k]$, we have

$$\begin{aligned} V_0^{-1} - V_{0,disc}^{-1} &= \frac{1}{\sigma_\epsilon^2}\left(\mathbb{E}\left[W_i W_i^{'}\right] - \sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k^{'}\right) \\ &= \frac{1}{\sigma_\epsilon^2}\sum_{k=1}^{K} p_k \left(\mathbb{E}\left[W_i W_i^{'}\middle|\, Z_i \in \mathcal{Z}_k\right] - \overline{W}_k \overline{W}_k^{'}\right) = \frac{1}{\sigma_\epsilon^2}\sum_{k=1}^{K} p_k \operatorname{Var}(W_i|\, Z_i \in \mathcal{Z}_k), \end{aligned}$$

which is positive semi-definite. Hence, $V_{0,disc} - V_0$ is positive semi-definite. □

Despite the efficiency loss, the discretization-based estimator $\hat{\theta}_{disc}$ is very simple and user-friendly. Applied researchers just need to create dummy variables for a chosen partition of $\mathcal{Z}$ and run a standard 2SLS command. Furthermore, in our simulations, we find that $\hat{\theta}_{disc}$ performs surprisingly well in finite sample: the efficiency loss of $\hat{\theta}_{disc}$ tends to be quite small and more than compensated by its smaller finite-sample bias as a 2SLS estimator that does not require nonparametric regressions.

Furthermore, in the special case where Z_i are discrete variables with finite support, there is clearly no information loss from discretization. In this case, expectations simplify to weighted sums over the K realizations of Z_i (weighted by the probability mass p_k), and it can be easily shown that the asymptotic variance of $\hat{\theta}_{disc}$ coincides with the one for $\hat{\theta}$ and $\hat{\theta}^*$ in Theorem 2 (without the homoskedasticity assumption).

Corollary 2. *Suppose that $\mathcal{Z} = \{z_1, ..., z_K\}$ for some finite K. Then, under the element-by-element partition, i.e., $\mathcal{Z}_k := \{z_k\}$, we have $V_{0,disc} = V_0$.*

Proof. Since $\hat{\theta}_{disc}$ is a 2SLS estimator, it is $\sqrt{n}$-consistent and asymptotic normal, with the asymptotic variance given by the formula $V_{0,disc} := \Sigma_{0,disc}^{-1}\Omega_{0,disc}\Sigma_{0,disc}^{-1}$ with

$$\begin{aligned}\Sigma_{0,disc} &:= \mathbb{E}\left[W_i' D_i\right]\left(\mathbb{E}\left[D_i D_i'\right]\right)^{-1}\mathbb{E}\left[D_i' W_i\right] \\ &= \sum_{k=1}^{K}\left(p_k \overline{W}_k \cdot \frac{1}{p_k}\cdot p_k \overline{W}_k'\right) = \sum_{k=1}^{K} p_k \overline{W}_k \overline{W}_k',\end{aligned}$$

and

$$\begin{aligned}\Omega_{0,disc} &:= \mathbb{E}\left[W_i' D_i\right]\left(\mathbb{E}\left[D_i D_i'\right]\right)^{-1}\mathbb{E}\left[\epsilon_i^2 D_i D_i'\right]\left(\mathbb{E}\left[D_i D_i'\right]\right)^{-1}\mathbb{E}\left[D_i W_i'\right] \\ &= \sum_{k=1}^{K}\left(p_k \overline{W}_k \cdot \frac{1}{p_k}\cdot p_k \overline{\sigma}_k \cdot \frac{1}{p_k}\cdot p_k \overline{W}_k'\right) = \sum_{k=1}^{K} p_k \overline{\sigma}_k^2 \overline{W}_k \overline{W}_k'.\end{aligned}$$

Note that, in the proofs above, we exploited the fact that each D_{ik} is a partition cell dummy along with the associated properties such as, for all k and $j \neq k$, $\mathbb{E}[D_{ik}] = p_k$, $D_{ik}^2 = D_{ik}$, $D_{ik}D_{ij} = 0$. □

B.6 Proof of Theorem 5

Proof. Under Assumption 1, the parameter θ_0 should satisfy $g(z, \theta_0) := \mathbb{E}[Y_i | Z_i = z] - m_0(z, \theta_0) = 0$, for any $z \in \mathcal{Z}$. The function $g(z, \cdot)$ is continuously differentiable since the function m_0 is continuously differentiable by assumption. Then by the local inverse theorem in Ambrosetti and Prodi (1995) (Chapter 2, Theorem 2), θ_0 is locally identified if there exists

$d = dims(\theta_0)$ distinct points $z_1, ..., z_d \in \mathcal{Z}$ such that the following condition holds:

$$G_0 := \begin{bmatrix} \partial_{\theta_1} g(z_1, \theta_0), & \partial_{\theta_2} g(z_1, \theta_0), & ..., & \partial_{\theta_d} g(z_1, \theta_0) \\ \partial_{\theta_1} g(z_2, \theta_0), & \partial_{\theta_2} g(z_2, \theta_0), & ..., & \partial_{\theta_d} g(z_2, \theta_0) \\ \vdots & \vdots & \vdots & \vdots \\ \partial_{\theta_1} g(z_d, \theta_0), & \partial_{\theta_2} g(z_d, \theta_0), & ..., & \partial_{\theta_d} g(z_d, \theta_0) \end{bmatrix} \text{ has full rank.}$$

As shown in Lemma 2, the above full rank condition is equivalent to the requirement that $\partial_{\theta_1} g(Z_i, \theta_0), \partial_{\theta_2} g(Z_i, \theta_0), ..., \partial_{\theta_d} g(Z_i, \theta_0)$ are not multicollinear. The no multicollinearity condition is also equivalent to the requirement that $\mathbb{E}\left[\nabla_\theta g(Z_i, \theta_0) \nabla_{\theta'} g(Z_i, \theta_0)\right] = \mathbb{E}\left[\nabla_\theta m_0(Z_i, \theta_0) \nabla_{\theta'} m_0(Z_i, \theta_0)\right]$ has full rank.

□

B.7 Proof of Lemma 5

Proof. Under Assumption 7, the moment function can be expressed as follows:

$$\begin{aligned} g(z, \theta) &= \mathbb{E}\left[\mathbb{1}\{Y_i \leq \alpha + Z_i'\beta + X_i'\gamma\} \middle| Z_i = z \right] - \tau \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{ Y_i \leq \alpha + Z_i'\beta + X_i'\gamma \right\} \middle| X_i \right] \middle| Z_i = z \right] - \tau \\ &= \mathbb{E}\left[\int \mathbb{1}\left\{ \epsilon \leq \alpha - \alpha_0 + z'(\beta - \beta_0) + X_i'(\gamma - \gamma_0) \right\} f_{\epsilon|X,Z}(\epsilon | X_i, z)\, d\epsilon \middle| Z_i = z \right] - \tau. \end{aligned}$$

The derivative of $\nabla_\theta g(z, \theta) = \nabla_\theta m_0(z, \theta)$ with respect to θ is given as

$$\begin{aligned} &\nabla_\theta g(z, \theta) = \nabla_\theta m_0(z, \theta) \\ =&\mathbb{E}\left[\nabla_\theta \int \mathbb{1}\left\{ \epsilon \leq \alpha - \alpha_0 + z'(\beta - \beta_0) + X_i'(\gamma - \gamma_0) \right\} f_{\epsilon|X,Z}(\epsilon | X_i, z)\, d\epsilon \middle| Z_i = z \right] \\ =&\mathbb{E}\left[f_{\epsilon|X,Z}\left(\alpha - \alpha_0 + z'(\beta - \beta_0) + X_i'(\gamma - \gamma_0) \middle| X_i, z \right) \left(1, z', X_i'\right)' \middle| Z_i = z \right]. \end{aligned}$$

Evaluating at θ_0, the derivative $m_0(z,\theta)$ is simplified as

$$\nabla_\theta m_0(z,\theta_0) = \mathbb{E}\left[f_{\epsilon|X,Z}(0|X_i,z) \begin{pmatrix} 1 \\ z \\ X_i \end{pmatrix} \middle| Z_i = z \right] = f_{\epsilon|Z}(0|z) \begin{pmatrix} 1 \\ z \\ \mathbb{E}\left[\frac{f_{\epsilon|X,Z}(0|X_i,z)}{f_{\epsilon|Z}(0|z)} X_i \middle| Z_i = z \right] \end{pmatrix}.$$

Applying Bayes' rule, we have

$$\begin{aligned} \mathbb{E}\left[\frac{f_{\epsilon|X,Z}(0|X_i,z)}{f_{\epsilon|Z}(0|z)} X_i \middle| Z_i = z \right] &= \int \frac{f_{\epsilon|X,Z}(0|x,z)}{f_{\epsilon|Z}(0|z)} x f_{X|Z}(x|z)\, dx \\ &= \int x f_{X|\epsilon,Z}(x|z,\epsilon=0)\, dx = \mathbb{E}[X_i|Z_i=z,\epsilon_i=0]. \end{aligned}$$

Therefore,

$$\nabla_\theta m_0(Z_i,\theta_0) = f_{\epsilon|Z}(0|Z_i) \begin{pmatrix} 1 \\ Z_i \\ \mathbb{E}[X_i|Z_i,\epsilon_i=0] \end{pmatrix}.$$

Since $f_{\epsilon|Z}(0|Z_i=z) > 0$ for any $z \in \mathcal{Z}$, the full rank condition of $\mathbb{E}[\nabla_\theta m_0(Z_i,\theta_0)\nabla_{\theta'} m_0(Z_i,\theta_0)]$ is equivalent to the no multicollinearity of 1, Z_i, $\tilde{\pi}_0(Z_i) := \mathbb{E}[X_i|Z_i,\epsilon_i=0]$.

□