Two-way Exclusion Restrictions in Models with Heterogeneous Treatment Effects*

Shenglong Liu[◇] Ismael Mourifié[†] Yuanyuan Wan[‡]

First version: May 2017
This version: May 2019

Abstract

In this paper we propose a novel method to identify the conditional average treatment effect partial derivative (CATE-PD) in an environment in which the treatment is endogenous, the treatment effect is heterogeneous, the candidate "instrumental variables" can be correlated with latent errors, and the treatment selection does not need to be (weakly) monotone. We show that CATE-PD is point identified under mild conditions if two-way exclusion restrictions exist: (i) an outcome-exclusive variable which affects the treatment but is excluded from the potential outcome equation, and (ii) a treatment-exclusive variable which affects the potential outcome but is excluded from the selection equation. We also propose an asymptotically normal two-step estimator and illustrate our method by investigating how the return to education varies across regions of different levels of development in China.

JEL Codes: C14, C31, C36
Keywords: Two-way exclusion, nonparametric identification, heterogeneous treatment effect, invalid instrumental variables

*We are grateful for comments from Victor Aguirregabiria, Loren Brandt, Dalia Ghanem, Marc Henry, Sung Jae Jun, Qi Li, Aureo de Paula, Joris Pinkse, Shu Shen, Thomas Russell. We benefited from discussions with participants at the Asia Meeting of Econometric Society (2017), Tsinghua International Conference on Econometrics, and department seminars at Ryerson University, UC Davis, TAMU, Academia Sinica, and Fudan University. Mourifié and Wan thank the support from SSHRC Insight Grant #435-2016-0045. All errors are ours.
◇ School of Public Policy and Management, Tsinghua University, liushenglong@mail.tsinghua.edu.cn.
† Corresponding author. Department of Economics, University of Toronto, 150 St. George Street, Toronto ON M5S 3G7, Canada. ismael.mourifie@utoronto.ca.
‡ Department of Economics, University of Toronto, yuanyuan.wan@utoronto.ca

1. Introduction

The validity of instrumental variables (IV) is one of most crucial questions in applying the IV approach to identify and estimate causal effects when the treatment is endogenous. When treatment effects are homogeneous, the concern is mainly about whether the instrumental variable is independent with the structural errors. When treatment effects are heterogeneous, the (weak) monotone response of treatment selection with respect to the IV is additionally required to identify meaningful causal parameters, (see discussions in Imbens and Angrist, 1994; Heckman and Vytlacil, 2005; Chaisemartin, 2016). Recently Kitagawa (2008, 2015), Huber and Mellace (2015), and Mourifié and Wan (2017) propose testing procedures to assess the joint assumption of instrument independence and (weak) monotonicity in the framework of local average treatment effect (LATE) models, and these papers find statistical evidence that the joint assumptions are violated in some empirical applications. However, to the best of our knowledge, few alternative approaches have been suggested to point identify meaningful causal parameters when the independence and/or (weak) monotonicity assumptions are violated.[1]

Motivated by this, our paper aims to provide an alternative method to identify meaningful parameters in an environment in which the treatment is endogenous, the treatment effect is heterogeneous (even after conditioning on all the observed covariates), the candidate "instrumental variables" can be correlated with latent errors, and the treatment selection does not need to be (weakly) monotone in these candidate "instrumental variables." We provide an alternative set of conditions under which the derivative of the conditional average treatment effect with respect to a particular regressor–a parameter which often has important policy implications–is point identified, while allowing both "sorting on the level" and "sorting on the gain," which are very crucial features of causal inference (see Heckman, Urzua, and Vytlacil, 2006). We also propose a simple nonparametric estimation procedure which is easy to implement. Our paper therefore complements to methods of identifying and estimating treatment effects with valid instrumental variables.

To fix the idea, we study the following potential outcome model

$$\begin{cases} Y = \sum_{d=0}^{T} \mathbf{1}[D=d] Y_d \\ Y_d = f_d(S,X) + U_d \end{cases}, \tag{1}$$

where for $d \in \mathscr{D} = \{0,1,2,\cdots,T\}$, Y_d is the potential outcome when the treatment variable D is externally set to d, $U_d \in \mathbb{R}$ is the treatment specific error term, Y is the observed outcome, $X \in \mathscr{X} \subset \mathbb{R}^{d_x}$ is a vector of covariates, $S \in \mathscr{S} \subset \mathbb{R}^{d_s}$ is a treatment-exclusive vector that only enters the potential outcome equation. We also assume that there exists an outcome-exclusive vector $Z \in \mathscr{Z} \subset \mathbb{R}^{d_z}$ which affects D but does not enter the f_d functions, as we will formally define in later sections. Following the convention in the literature, we use D_z to denote the potential treatment when Z be externally set to z. Without loss of generality, we focus on the univariate S case (i.e. $d_s = 1$) in the rest of the paper.

In this model environment with presence of hetereogenous treatment but where LATE assumptions fails to hold, our primarily concern will be the identification and estimation of $\beta_{d,d'}(s,x) = \partial \Delta_{d,d'}(s,x)/\partial s$, where $\Delta_{d,d'}(s,x) \equiv f_d(s,x) - f_{d'}(s,x)$. Note that if error terms

[1] One way to relax the monotonicity assumption is to consider the random coefficient model. Gautier and Hoderlein (2015) consider a random coefficient triangular model, where in the first stage equation, the coefficient of just one instrumental variable has a support on one side of zero but the others coefficients can take different signs.

have zero conditional mean, i.e., $\mathbb{E}[U_d|S,X]=0$ a.s, $\Delta_{d,d'}(s,x)\equiv f_d(s,x)-f_{d'}(s,x)$ is equal to the conditional average treatment effect (CATE) and $\beta_{d,d'}(s,x)$ the conditional average treatment effect partial derivative (CATE-PD).[2] We will show that under our conditions, $\beta_{d,d'}(s,x)$ is point identified for (s,x) in the interior of the joint support of S and X without valid instruments. $\beta_{d,d'}(s,x)$ is a useful parameter in many empirical environments. Often empirical researchers are interested in whether the CATE is different or monotone across subpopulations indexed by s.[3] $\beta_{d,d'}(s,x)$ can provide such information. For instance, if Y represents income, D represents education level, X is a set of demographic controls, and S is taken as a proxy of regional development level, then $\beta_{d,d'}(s,x)<0$ suggests the average return to education for individuals with observed characteristics x is higher in less developed regions.[4]

Furthermore, since $\beta_{d,d'}(\cdot,x)$ is point identified, the CATE difference $\Delta_{d,d'}(s,x)-\Delta_{d,d'}(s^*,x)$ is point identified for any pair (s,s^*). If information on $\Delta_{d,d'}(s^*,x)$ is available ex-ante — for example if a randomized experiment, or a quasi-experimental design such as a sharp Regression discontinuity design allows one to identify the CATE for the subpopulation characterized by s^* but only for this subpopulation, then we can use our parameter $\beta(\cdot,x)$ to recover $\Delta_{d,d'}(s,x)$ for all other values of s. Our method thus provides a robust way of extrapolating CATE without requiring the availability of valid instruments in the targeted subpopulations. Of course, if the researcher has in hand a continuous instrument Z that is indendenpent with all potential variables conditioning on (S,X) and treatment selection is monotone in Z, then he/she could estimate the marginal treatment effect (MTE) and then recover $\Delta_{d,d'}(s,x)$ and $\beta_{d,d'}(s,x)$, see Heckman and Vytlacil (2005). However, when no such instruments are available, it is an open question that if any interesting parameters can be point identified.[5] Our paper therefore contributes to the existing literature by providing point identification results for relevant interpretable parameter of interests, when the usual IV method assumptions fail to hold.

We do not make parametric assumptions or impose other shape restrictions on the potential outcome function f_d, other than assuming the error U_d is separable with $f_d(S,X)$. The additive separability assumption may appear to be a strong restriction here, however unlike Das (2005) and Florens and Malavolti (2002), but following Heckman and Vytlacil (2005), we allow the errors to be treatment specific, that is, $U_d\neq U_{d'}$. With such a specification, we do not rule out the feature of "essential heterogeneity" (see discussions in Heckman, Urzua, and Vytlacil, 2006) since $U_d-U_{d'}$ is allowed to be correlated with the treatment D (sorting on gains) after conditioning on all observed covariates.

Although we do not model the selection equation explicitly, we do allow for $D=$

[2]We focus on the case in which S is continuous and f_d is differentiable with respect to s. As we shall see later, our identification results hold straightforwardly when S is discrete.

[3]There have been tests developed for examining if certain treatment effect parameters are different across different observed subpopulations in other causal inference frameworks, see e.g. Hsu and Shen (2017, 2016); Hsu, Liu, and Shi (2016). These tests are all based on the validity of identifying assumptions. In regression discontinuity designs, Dong and Lewbel (2015) discuss identification and empirical relevance of treatment effect derivative (TED) – the derivative of the treatment effect with respect to the running variable at the cutoff.

[4]The policy implications of such a result will be discussed later in the empirical application.

[5]For example, Heckman and Vytlacil (2005, pp.723) commented: *"Generalizing the MTE to the case of a nonseparable choice equation that violates the monotonicity condition, we can define but cannot identify the policy parameters of interest. If we make the model symmetrically heterogeneous in outcome and choice equations, the method of instrumental variables and our extensions of it break down in terms of estimating economically interpretable parameters. This case is beyond the outer limits of an entire literature.... More general structural methods are required*".

$\vartheta(Z, X, \varsigma)$, where ς is a vector of unobserved latent variables (possibly correlated with U_d, and Z) and the measurable function ϑ may not be monotone in either Z or in any elements of ς. We neither assume Z has a large support nor require it to be independent of (U, ς) conditional on other covariates. Instead, we focus on applications in which the regressor S affects the potential outcome but not the treatment (to be defined rigorously later). In our context, S and Z together form two-way exclusive restrictions, which will play a determinant role in our identification of $\beta_{d,d'}(s, x)$.

Similar to the conventional exclusion variable Z, the exclusion of S from the selection equation is more natural in some applications than others. Recently, Eisenhauer, Heckman, and Vytlacil (2015, EHV) consider a generalized Roy model with limited information by the agent and discuss scenarios in which such two-way exclusion variables exist. For example, the outcome-exclusive variable Z plays a similar role as the cost shifter of taking the treatment (in the terms of EHV), and the treatment-exclusive variable S likewise plays a similar role as the benefit shifter which is not perfectly foreseen at the time of treatment. The exclusion of S from selection equation is often justifiable in two-step decision problems. In such models, an economic agent first chooses to engage in an activity (or in a regime) and then subsequently decides the level of his/her activity. If the S is realized between the two decisions, then it is naturally excluded from selection equation. For instance, researchers may be interested in how the *contemporaneous local wage* or *experience* affects the return to college education, as studied in EHV. In this example, the first decision of a high school graduate is to go to college or not, which of course determines the level of his/her future outcomes. However, there is often a significant time gap between college enrollment and the realization of returns to college. The realization of local wage or experiences, however, were unknown to the student when he/she made the college enrollment decision. (although the student may form his/her expectation based on whatever information he/she had at the moment of the decision). Our paper is significantly different from EHV in that in EHV, the S variable, together with the treatment monotonicity and the IV independence assumption, provides identification for parameter (e.g. the cost of treatment) that is different from ours.

Another way of taking into account heterogeneous treatment effects is to use a model with a nonseparable structural function. Thereby, a part of the literature has been focusing on nonseparable structural functions with discrete endogenous variables, see for instance Chesher (2005), Jun, Pinkse, and Xu (2011), Shaikh and Vytlacil (2011), Mourifié (2015) among many others. However, most of the literature imposes additional restrictions on the selection equation, such as a threshold crossing model with a scalar-valued latent error term, and achieve, in general, set identification results.[6] Using the exclusion variable S and assuming the additively-separable structure of potential outcomes frees us from making shape restrictions on the selection equation.

We propose an asymptotically normal two-step estimator for $\beta_{d,d'}(s, x)$. In the first step, we estimate the conditional choice probability of treatment and the derivative of the conditional expectation of the outcome with respect to the treatment-exclusive variable S by a local linear estimation method. The second step estimator utilizes the fitted value of the first stage estimator and has an explicit form, which is analogous to the ordinary least

[6] A rare exception is Chernozhukov and Hansen (2005), who do not impose direct structure on the selection equation but require a type of monotone likelihood ratio condition and impose a rank similarity/invariance condition. Notice that recently, Dong and Shen (2017) propose testing procedures to assess the joint assumption of instrument independence, monotonicity, and rank similarity/invariance and also find statistical evidence that the joint assumptions are violated in some empirical applications. See discussions in Yu (2017) too.

squares estimator. We show our estimator is consistent and asymptotically normal, and provide variance estimators for conducting inference.

We apply our method to investigate how returns to education varies across prefectures with different levels of development using China's 2005 One Percent Population Survey. We consider the local *contemporaneous* average income as a proxy of local development and use it as the treatment exclusion variable. The exclusion of the local contemporaneous average income from selection equation is justified by the fact that this variable had not yet been realized when individuals made their schooling decision,[7] and as a market level variable it is likely independent of individual abilities. We are able to use mother's education as the outcome exclusion variable despite that it does not satisfy the conventional IV-validity assumptions in our data set. We find that returns to education significantly differs across areas with different levels of development in China, and it tends to be higher in poorer regions (western China). This empirical finding is consistent with the discussions in Heckman (2005).

The rest of the paper is organized as follows. We present our main identification result in Section 2. We propose the two-step estimator and derive its asymptotic properties in Section 3. As an empirical application, we use our method to investigate the regional differences in returns to education in China in Section 4. Section 5 concludes the paper.

2. Model and Identification

In this section, we will discuss the identifying assumptions and lay out the identification strategy. For the convinience of readers, we repeat the model below:

$$\begin{cases} Y = \sum_{d=0}^{T} \mathbf{1}[D = d] Y_d \\ Y_d = f_d(S, X) + U_d \end{cases},$$

where for $d \in \mathscr{D} = \{0, 1, 2, \cdots, T\}$. To simplify notation, we take S to be scalar-valued; our results can be extended straightforwardly to the cases with vector-valued S. We make the following assumptions:

Assumption 1 (Exclusion restrictions). *(i) The variable S is excluded from the observed treatment, i.e. $D = \vartheta(X, Z, \varsigma)$ for some unknown measurable functions ϑ and random vector ς. (ii) The variable Z does not enter $f_d(S, X)$ for each $d \in \mathscr{D} \equiv \{0, 1, \cdots, T\}$.*

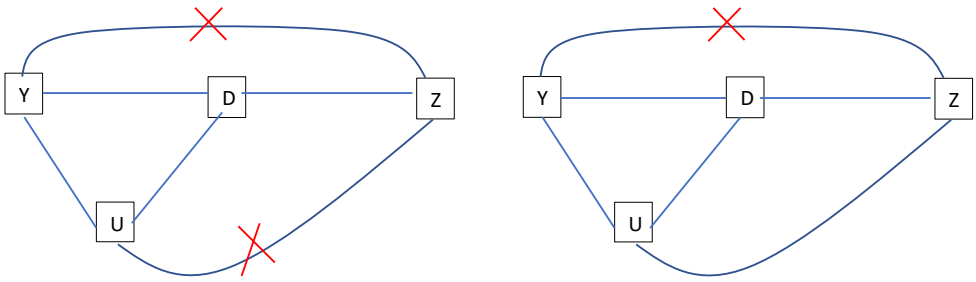
Assumption 2 (Independence). *$(U, \varsigma) \perp S | X, Z$, where $U = (U_0, U_1, \cdots, U_T)'$.*

Assumptions 1 and 2 are our key identifying assumptions. Assumption 1 is interpreted as follows: There is no direct causal effect of S on the treatment D, and there is no direct causal effect of Z on the potential outcome Y_d once we fix U_d. The first exclusion holds naturally in two-step decisions problems when D is chosen before the realization of S (the decision maker may know the distribution of S and has factored it into the function form of ϑ). The second exclusion is weak as it allows Z has indirect effect on Y_d through its correlation with U_d, as we will further elaborate in the following paragraph. Assumption 2 requires S to be statistically independent of (U, ς) conditional on (X, Z). This type of exogeneity assumption is often assumed in nonparametric identification of heterogeneous treatment effect models.

[7] Note that we allow individuals to take into account the distribution of future local income level, given all the information they have at the moment they make their educational choice.

For instance, Eisenhauer, Heckman, and Vytlacil (2015, Assumption 1) requires that the vector of unobservables to be statistically independent to all the covariates in the model, i.e $(U, \varsigma) \perp (S, X, Z)$, which implies our Assumption 2. Here, we just impose this type of exogeneity only for S.

It is worth emphasizing that Assumptions 1 and 2 impose weaker restriction on Z than the conventional instrumental variables framework. For an instrumental variable to be valid, it should be excluded from the outcome equation and should be independent with U. A valid instrumental variable can only affect the observed outcome (Y) indirectly though the endogenous treatment D. On the other hand, in our model, U and Z can be correlated and Z can affect the outcome not only through treatment D, but also through U. The following Figure 1 illustrates the difference between the usual IV assumption and the one that we consider here. The red crossings in each of the graphs show that such dependence links are not allowed in the corresponding framework. As a consequence, Z becomes easier to find or justify in empirical researches. For instance, in applications of estimating return to education, if one believes that parents' education affects children's talent U, then parents' education can not be used as an instrumental variable. However, as long as parents' education does not directly affect the outcome (Y) other than indirectly through children's education (D) and talent (U), it can serve as the Z variable in our model.



IV assumption Our assumption

Figure 1: Restrictions on Z

We impose almost no restrictions on the functional form of the selection equation $D = \vartheta(X, Z, \varsigma)$ other than S being excluded. We allow ς to be multi- or infinite dimensional and do not impose any monotonicity or threshold crossing conditions on ϑ.[8] This is also important because the plausibility of monotonicity has been questioned in some empirical contexts, for instance see the discussions of Barua and Lang (2016). See also Klein (2010) for the relaxation of the monotonicity.

Before discussing our main result, we make Assumption 3 just for simplification of notation. As will be clear later, our identification and estimation strategy can be easily extended to the case when S is discrete.

[8]The specification of function ϑ plays a similar role as the function $\delta(\cdot)$ in Chernozhukov and Hansen (2005, Assumption 3).

Assumption 3 (Differentiability). *S is continuous. Let $\mathscr{S}_x$ be the support of S conditional on $X = x$. Then for each $x \in \mathscr{X}$, $f_d(\cdot, x)$ for $d = 0, 1, \cdots, T$ is continuously differentiable in the interior of $\mathscr{S}_x$.*

We now describe our identification strategy below. Without loss of generality, we take the untreated group (those with $d = 0$) as the reference group. Model (2) can be equivalently written as follows:

$$\begin{aligned} Y &= \sum_{d=0}^{T} f_d(S, X) 1\{D = d\} + \sum_{d=0}^{T} U_d 1\{D = d\} \\ &= f_0(S, X) + \sum_{d=1}^{T} (f_d(S, X) - f_0(S, X)) 1\{D = d\} + \sum_{d=1}^{T} (U_d - U_0) 1\{D = d\} + U_0. \end{aligned}$$

We write $\tilde{f}_d(s, x) = f_d(s, x) - f_0(s, x)$. Let $W = (S, X,' Z')'$ and $\mathscr{W}$ be the support of W. Letting $\mathbb{E}[\cdot|w]$ represent the conditional expectation given $W = w$, we have:

$$\mathbb{E}[Y|w] = f_0(s, x) + \sum_{d=1}^{T} \tilde{f}(s, x) \mathbb{P}(D = d|w) + \mathbb{E}\left[\sum_{d=0}^{T} U_d 1\{D = d\}|w\right]. \tag{2}$$

Note that

$$\mathbb{E}\left[\sum_{d=0}^{T} U_d 1\{D = d\}|w\right] = \sum_{d=0}^{T} \mathbb{E}\left[U_d|D = d, w\right] \mathbb{P}(D = d|w).$$

Assumptions 1 and 2 imply that conditioning on (X, Z), S is independent with (U_d, D), hence $\mathbb{E}\left[U_d|D = d, w\right]$ does not depends on s. Similarly, under Assumptions 1 and 2, $\mathbb{P}[D = d|w] = \mathbb{P}(D = d|x, z)$ for all (x, z). By taking the derivative of the latter equation with respect to s, we have

$$\frac{\partial \mathbb{E}[Y|w]}{\partial s} = \sum_{d=1}^{T} \beta_d(s, x) \mathbb{P}(D = d|x, z) + \beta_0(s, x), \tag{3}$$

where $\beta_0(s, x) \equiv \frac{\partial f_0(s,x)}{\partial s}$ and $\beta_d(s, x) \equiv \beta_{d,0}(s, x) = \frac{\partial \tilde{f}_d(s,x)}{\partial s}$ for $d = 1, 2, \cdots, T$. Here we simplify the notation and write β_d for $\beta_{d,0}$. Note that β_d is a parameter of empirical interest. It measures how the average treatment effect changes with s. For example, it could measure how the returns to college education changes with variation in local labor market conditions.

To simplify the notation, let $m(w) = \partial \mathbb{E}[Y|\omega]/\partial s$, $\pi_d(x, z) = \mathbb{P}(D = d|x, z)$, $\boldsymbol{\pi}_0(x, z) = [\pi_1(x, z), ..., \pi_T(x, z)]'$, $\boldsymbol{\pi}(x, z) = [1, \boldsymbol{\pi}_0(x, z)']'$, and $\beta(s, x) = [\beta_0(s, x), \beta_1(s, x), ..., \beta_T(s, x)]'$. Therefore, Equation (3) can be rewritten as follows:

$$m(s, x, z) = \boldsymbol{\pi}(x, z)' \beta(s, x), \quad \forall (s, x, z) \in \mathscr{W}. \tag{4}$$

Equation (4) is the key identifying equation. m and $\boldsymbol{\pi}$ are identified directly from the data. β is the parameter of interest. As shall be clear soon, the conditional variation of Z given $(S, X) = (s, x)$ provides identification power for $\beta(s, x)$, which is summarized by the following Theorem.[9]

[9]Note that the separability plays an important role in the derivation, unless U_d and D are indepen-

Proposition 1. *Let (s, x) be a point from the joint support of S and X. Under Assumptions 1 to 3 and provided that corresponding expectations exist, $\beta(s, x)$ is identified if and only if the conditional variance $\mathbb{V}[\boldsymbol{\pi}_0(x, Z)|S = s, X = x]$ is positive definite. The identification equation is given by*

$$\beta(s,x) = \{\mathbb{E}[\boldsymbol{\pi}(x,Z)\boldsymbol{\pi}(x,Z)'|S=s,X=x]\}^{-1}\mathbb{E}[\boldsymbol{\pi}(x,Z)m(s,x,Z)|S=s,X=x]. \quad (5)$$

PROOF. *See Appendix A.1.* □

We have a few remarks. First, Proposition 1 generalizes Das (2005, Theorem 2.1). The key difference is that Proposition 1 allows treatment-specific error terms and invalid "instrumental variable" Z, by taking advantage of the treatment-exclusive regressor S. Second, when Z is discrete, the identification of $\beta(s, x)$ requires that conditional on $(S, X) = (s, x)$, the support of Z contains at least $T + 1$ distinct values: $\{z_1, z_2, \cdots, z_{T+1}\}$. It is also sufficient that the matrix $\Pi(x) = [\boldsymbol{\pi}(x, z_1)'; \boldsymbol{\pi}(x, z_2)'; \cdots, \boldsymbol{\pi}(x, z_{T+1})']$ has full rank (which is testable). For example, for the case of binary D and binary Z, the condition requires

$$\mathbb{P}(D=1|S=s,X=x,Z=1) \neq \mathbb{P}(D=1|S=s,X=x,Z=0), \quad \forall (s,x),$$

which is a rank condition that we are familiar with. Third, to identify β, we do not need to impose any location normalization on the distribution of U_d. Lastly, our identification result holds as long as the final term on the right hand side of Equation (2) does not depend on s; therefore the conditional independence between S and errors in the outcome equation (Assumption 2) is sufficient but not necessary.

When S is discrete, for a given vector of $(s, \tilde{s}, x, z)$, we can define

$$\begin{aligned} m(s,\tilde{s},x,z) &\equiv \mathbb{E}[Y|s,x,z] - \mathbb{E}[Y|\tilde{s},x,z] \\ &= f_0(s,x) - f_0(\tilde{s},x) + \sum_{d=1}^{T}(f_d(s,x) - f_d(\tilde{s},x) - f_0(s,x) + f_0(\tilde{s},x))\mathbb{P}(D=d|x,z), \quad (6) \end{aligned}$$

and analogously define $\beta_d(s, \tilde{s}, x) \equiv f_d(s, x) - f_d(\tilde{s}, x) - f_0(s, x) + f_0(\tilde{s}, x)$. In this scenario, the same identification strategy carries through for the identification of $\beta_d(s, \tilde{s}, x)$.

Under the current assumptions, the structural function $f_d(s, x)$ can only be identified under additional assumption. The following corollary shows that if the researcher is willing to impose an additional restriction, the well-known conditional average treatment effect (CATE) is identified.

Corollary 1. *If there is s^* such that $\Delta_{d,d'}(s^*, x)$ is known or identifyable, then $\Delta_{d,d'}(s, x) \equiv f_d(s, x) - f_{d'}(s, x)$ is identified*

$$\Delta_{d,d'}(s,x) = \Delta_{d,d'}(s^*,x) + \int_{s^*}^{s}(\beta_d(t,x) - \beta_{d'}(t,x))dt.$$

dent when conditional on observable W. To see this, suppose $Y_d = f_d(S, X, U_d)$ and define $ASF_d(w) = \mathbb{E}[f_d(s, x, U_d)|W = w]$, then in general $\mathbb{E}[Y|W] \neq \sum_{d=0}^{T} ASF_d(w)\mathbb{P}(D = d|W = w)$ because $f_d(s, x, U_d)$ and D can be correlated even conditioning on W. One could alternatively define $LASF_d(w) = \mathbb{E}[f_d(s, x, U_d)|W = w, D = d]$, then it can be shown that $\mathbb{E}[Y|W] = \sum_d LASF_d(w)\mathbb{P}(D = d|W = w)$, but this provide littles help for identification of $ASF(w)$.

Furthermore, if $\mathbb{E}[U_d|S,X] = 0$ *a.s. for all* d*, then* $\Delta_{d,d'}(s,x)$ *represents the well-known conditional average treatment effect* $\mathbb{E}[Y_d - Y_{d'}|S=s,X=x]$.

To illustrate the potential usefulness of Corollary 1, consider again an empirical study in which Y is income, D is the education level, X is a set of demographic controls, and S is taken as a proxy of regional development level. Suppose that in a specific region with $S = s^*$, the local government implements a compulsory school law such that the change to the minimums school-leaving age. As discussed in Oreopoulos (2006), the magnitude of its impact may provide an opportunity to consistently estimate the conditional average return to education in this locality, i.e. $\Delta_{d,d'}(s^*,x)$. Using Corollary 1, we can recover the conditional average return to education in another locality with $s \neq s^*$, even though such a compulsory school law is not available there. More broadly speaking, if a randomized controlled experiment or a natural/quasi experiment is conducted in subpopulation s^* so that researchers can credibly estimate $\Delta_{d,d'}(s^*,x)$, then $\Delta_{d,d'}(s,x)$ can be point identified under our assumptions.

For further illustration, consider another example in Eisenhauer, Heckman, and Vytlacil (2015), where S denotes the years of experience, D denotes years of education, and Y denotes wage income. We could expect that without any experience i.e., $s = 0$, two individuals with similar observable characteristics x but only one year difference in the level of education may have little difference in expected potential labor income: $\mathbb{E}[Y_d|S=s,X=x] = \mathbb{E}[Y_{d+1}|S=s,X=x]$, or equivalently $f_d(0,x) = f_{d+1}(0,x)$. For example, individuals with one or two years of college education but zero working experience may be evaluated in a similar way on the labor market. Such a situation could also be justified by the "sheepskin effects" hypothesis.[10] In this case, the requirement of Corollary 1 is satisfied at $s^* = 0$: $\Delta_{d,d'}(0,x) = f_d(0,x) - f_{d+1}(0,x) = 0$. Note that with the existence of treatment-specific unobserved heterogeneity, those two individuals may still have different outcome realizations since $Y_d - Y_{d+1} = U_d - U_{d+1} \neq 0$. Of course, if the gap in the years of education is wide, or an additional year of education upgrades the diploma to next level (e.g. from high school to college), such an assumption would be less likely to hold.

It is noted that the availability of the knowledge on $\Delta_{d,d'}(s^*,x)$ is case by case and researcher must have a clear empirical justification for it. If prior information on $\Delta_{d,d'}(s^*,x)$ is indeed available, then Linton and Härdle (1996)-type marginal integration estimation methods can be applied to estimate $\Delta_{d,d'}(s,x)$, which can in turn can serve as basis for estimating other parameters, e.g. $\partial\Delta_{d,d'}(s,x)/\partial x$. If prior information on $\Delta_{d,d'}(s^*,x)$ is not available, our identification strategy only leads to identification of $f_d(s,x)$ up to an unknown additive term $\psi_d(x)$. Identification up to location restriction is common in this literature, see, for example Das, Newey, and Vella (2003), Das (2005) and D'Haultfœuille and Maurel (2013), among many others. To answer some policy questions (as discussed in the introduction and in our empirical application), the point identification of β would be sufficient, but admittedly, there are also many cases in which the location term is desirable (see Heckman, 1990), then one can consider adapting Andrews and Schafgans (1998)'s methodology provided conditions therein are satisfied. We leave these questions for future research.

[10]The "sheepskin effect" is the hypothesis that obtaining an educational degree accompanied by a certificate made of parchment would yield a higher income than the same amount of studying without possession of a certificate. There often exists significant changes or differences in returns to schooling only at the year of schooling granted by a diploma. Please see Belman and Heywood (1991) and Jaeger and Page (1996).

3. Estimation

In this section, we propose an estimator for $\beta_d(s,x)$ for given d, s and x.[11] Since it is quite common in practice that the exclusion variable Z takes a finite set of values (as in our empirical application), we present the asymptotic results of our estimator by focusing on the case in which $Z \in \mathscr{Z} \equiv \{z_1, z_2, \cdots, z_J\}$ in this section. We leave the results for continuous Z in the online supplementary materials (Liu, Mourifie, and Wan, 2019, Section 2).

Throughout the rest of the paper, we use the product kernel; that is, for a generic d_t-dimensional variable, we let $\boldsymbol{K}_t(\cdot) = \Pi_{j=1}^{d_t} K(\cdot)$ where $K(\cdot)$ is a one-dimensional kernel. We use h to denote the bandwidth that converges to zero when $n \to \infty$. The following assumptions are maintained in this section.

Assumption 4. *$\{(Y_i, D_i, X_i, S_i, Z_i)\}_{i=1}^n$ are i.i.d. observations.*

Assumption 5. *The support of the conditional distribution of $Z|(S,X) = (s,x)$ does not depend on (s,x). Furthermore, $\mathbb{V}[\boldsymbol{\pi}_0(x,Z)]$ is positive definite.*

Assumption 5 is not necessary for either identification or estimation, but it does simplify the expression of our estimator. It holds when Z and (S,X) are independent. When Z is discrete, it just requires that $\mathbb{P}(Z = z|S = s, X = x) > 0$ for all $z \in \mathscr{Z}$ and (s,x) from the joint support of S and X, which is an assumption often satisfied in the literature. Under Assumption 5, Equation (4) implies that for the given (s,x), the following equation holds for each $z \in \mathcal{Z}$:[12]

$$m(s,x,z) = \boldsymbol{\pi}(x,z)'\beta(s,x),$$

which in turn implies

$$\beta(s,x) = \{\mathbb{E}[\boldsymbol{\pi}(x,Z)\boldsymbol{\pi}(x,Z)']\}^{-1}\mathbb{E}[\boldsymbol{\pi}(x,Z)m(s,x,Z)]. \tag{7}$$

Equation (7) suggests a simple two-step estimator for $\beta(s,x)$. In the first step, we estimate $\boldsymbol{\pi}$ and m nonparametrically; in the second step, we estimate $\beta(s,x)$ by plugging-in fitted values:

$$\begin{aligned}\hat{\beta}(s,x) &= \left(\frac{1}{n}\sum \hat{\boldsymbol{\pi}}(x,Z_i)\hat{\boldsymbol{\pi}}(x,Z_i)'\right)^{-1}\left(\frac{1}{n}\sum \hat{\boldsymbol{\pi}}(x,Z_i)\hat{m}(s,x,Z_i)\right)\\ &= \left(\sum_{j=1}^{J}\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)\hat{\boldsymbol{\pi}}(x,z_j)'\right)^{-1}\left(\sum_{j=1}^{J}\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)\hat{m}(s,x,z_j)\right),\end{aligned}$$

where $\hat{\omega}_j = (1/n)\sum_i 1\{Z_i = z_j\}$ is a consistent estimator for $\omega_j = \Pr(Z = z_j)$. $\hat{\pi}_d(x,z_j)$ can be obtained as the constant term of the local polynomial regression of $\mathbf{1}\{D_i = d\}$ on X_i using only the observations "i" such that $Z_i = z_j$. Likewise, $\hat{m}(s,x,z_j)$ can be obtained as the slope of the linear term $(S_i - s)$ of the local polynomial regression of Y_i on (S_i, X_i)

[11] We take X as continuous random variables. If X contains discrete regressors, we can also conduct the same analysis by conditioning on realizations of those regressors.

[12] Note that without Assumption 5, the support of the conditional distribution of Z given $S = s, X = x$ is not necessarily the same as the support of the marginal distribution of Z.

but again using only the corresponding subsample and with appropriate under-smoothing to eliminate the bias in the first stage.[13]

We define some notation before laying out the expression of the estimator. Let ϱ be a generic random vector and d_ϱ be its dimension. Let $\underline{r}^\varrho$ be a d_ϱ-vector of nonnegative integers and $|\underline{r}^\varrho| = \sum_{j=1}^{d_\varrho} r_j$. Let $\alpha_{\underline{r}^\varrho} \cdot \varrho^{\underline{r}^\varrho}$ be a polynomial of v of order $|\underline{r}^\varrho|$ with corresponding coefficients $\alpha_{\underline{r}^\varrho}$. We follow the convention that $v^{\underline{r}^\varrho} = 1$ when $|\underline{r}^\varrho| = 0$. For some positive integer p, let $\boldsymbol{\alpha}^\varrho$ be the stacked vector of $\alpha_{\underline{r}^\varrho}$ of which $0 \leq |\underline{r}^\varrho| \leq p$ and arranged in increasing order of $|\underline{r}^\varrho|$; for those with the same $|\underline{r}^\varrho|$, $\alpha_{\underline{r}^\varrho}$ are stacked in lexicographic order, that is, $\alpha_{(0,0,\cdots,|\underline{r}^\varrho|)}$ goes first and $\alpha_{(|\underline{r}^\varrho|,0,\cdots,0)}$ goes last. Let $\mathcal{P}(\varrho, \boldsymbol{\alpha}, p) = \sum_{0 \leq |\underline{r}^\varrho| \leq p} \alpha_{\underline{r}^\varrho} \cdot \varrho^{\underline{r}^\varrho}$ denote such a p-order polynomial.

With this notation, we can define our estimator for a given (s, x) and each $z_j \in \mathscr{Z}$ as:

$$\tilde{\boldsymbol{\alpha}}^m(s, x, z_j) = \text{argmin}_{\boldsymbol{\alpha}} \frac{1}{2n} \sum_{i=1}^{n} \mathbf{1}[Z_i = z_j] \boldsymbol{K}_s \left(\frac{S_i - s}{h} \right) \boldsymbol{K}_x \left(\frac{X_i - x}{h} \right) \times \left(Y_i - \mathcal{P}((S_i - s, X_i - x), \boldsymbol{\alpha}, p) \right)^2.$$

For $d = 1, 2, \cdots, T$, let

$$\tilde{\boldsymbol{\alpha}}^{\pi_d}(x, z_j) = \text{argmin}_{\boldsymbol{\alpha}} \frac{1}{2n} \sum_{i=1}^{n} \mathbf{1}[Z_i = z_j] \boldsymbol{K}_x \left(\frac{X_i - x}{h} \right) \times \left(1\{D_i = d\} - \mathcal{P}((X_i - x), \boldsymbol{\alpha}, p) \right)^2,$$

Finally, we can define our estimator $\hat{m}(s, x, z_j)$ to be the estimated coefficient corresponding to the linear term $(S_i - s)$ in the first regression and $\hat{\pi}_d(x, z_j)$ to be the coefficient associated with the constant term in the second regression. For the purpose of estimating the first-order derivative of the regression function $\mathbb{E}[Y|W = w]$ with respect to s, we follow the convention and use local quadratic regression; that is, we choose $p = 2$ so that the difference between the order of the polynomial and the order of the derivative to be estimated is odd (see Fan and Gijbels, 1996, section 3.3).

Assumption 6. *The bandwidth h is chosen such that $h \propto n^{-\frac{1}{6+(d_s+d_x)-\delta}}$ for some $0 < \delta < 1$.*

Assumption 7. *(i) The conditional density of (S, X) given $Z = z$ is bounded away from 0 and has bounded first-order derivative over its compact support for each $z \in \mathscr{Z}$. (ii) $\boldsymbol{\pi}(\cdot)$ and $\mathbb{E}[Y|W = \cdot]$ are $q+1$ times continuously differentiable for some $q \geq 2$. (iii) There exists some $\nu > 2$ such that $\mathbb{E}\|U\|^\nu$ is finite.*

Assumption 8. *The symmetric kernel $K(\cdot)$ has support $[-1, 1]$, integrates to one, and is continuously differentiable.*

Assumptions 6 to 8 are standard assumptions in local polynomial estimation of first-order derivatives of conditional expectation. Note that to estimate the first-order derivative, the

[13] We chose the local polynomial estimator for the first step for the following reasons. First, $m(s, x, \cdot)$ is the partial derivative of the conditional mean function of Y given $W = (s, x, z)$ with respect to s , which can be conveniently estimated by the local polynomial method. Second, for the continuous Z case, we will apply the uniform Bahadur representation results of Kong, Linton, and Xia (2010) to the first step estimators and derive the asymptotic distribution for the second step estimator. It should be noted $\boldsymbol{\pi}$ and m can also be estimated by other nonparametric methods, for example, sieve estimation.

variance is of order $\frac{1}{nh^{d+2}}$, where $d = d_s + d_x$. The bias is of order h^p since $p-1$ is odd. This implies that the optimal rate is $h = n^{-\frac{1}{2p+d+2}} = n^{-\frac{1}{6+d}}$. Assumption 6 implies that we undersmooth to eliminate the bias in estimating $m(s,x,z)$, that is $h^p \prec \frac{1}{\sqrt{nh^{d+2}}} \prec 1$, where "$a \prec b$" denotes "$a$ is of (strictly) smaller stochastic order than b" or equivalently $a/b \xrightarrow{p} 0$. We also need to ensure that the estimation error in $\hat{\pi}_d$ is negligible compared with that when estimating $m(s,x,z)$. The variance of $\hat{\pi}_d$ is of order $\frac{1}{nh^{d_x}}$ and the bias is of order h^{p+2} since p is even. Therefore, we know that both the variance and bias are asypmtotocially negligible since $\frac{1}{nh^{d_x}} \prec \frac{1}{nh^{d+2}} \prec 1$ and $h^{p+2} \prec h^p \prec \frac{1}{nh^{d+2}} \prec 1$. One could choose a different bandwidth to further improve the convergence rate of $\hat{\pi}_d$, but this just makes the estimation error in $\hat{\pi}_d$ "more negligible" and will not change the asymptotic properties of the second stage estimator (although finite sample performance may be improved).

Proposition 2. *Let(s,x) be an interior point of the joint support of (S,X). Suppose that Assumptions 1 to 8 are satisfied, then $\hat{\beta}(s,x) \xrightarrow{p} \beta(s,x)$. Furthermore, the first stage estimators satisfy $\sqrt{nh^{d_x}}\left(\hat{\boldsymbol{\pi}}(x,z_j) - \boldsymbol{\pi}(x,z_j)\right) = O_p(1)$ and $\sqrt{nh^{d_x+d_s+2}}(\hat{m}(s,x,z_j) - m(s,x,z_j)) \xrightarrow{d} N(0,\Omega_j(s,x))$, and*

$$\sqrt{nh^{d_x+d_s+2}}(\hat{\beta}(s,x) - \beta(s,x)) \xrightarrow{d} N(0, V^{-1}(x)\Omega(s,x)V^{-1}(x))$$

where $V(x) = \mathbb{E}[\boldsymbol{\pi}(x,Z_i)\boldsymbol{\pi}'(x,Z_i)]$ and $\Omega(s,x) = \sum_{j=1}^{J} \omega_j^2 \Omega_j(s,x)\boldsymbol{\pi}(x,z_j)'\boldsymbol{\pi}(x,z_j)$.

Proof. *See Appendix Appendix A.2.* □

Here V depends on x; Ω_j and Ω depend on (s,x). To simplify notation, we abbreviate them as V, Ω_j and Ω, respectively. Proposition 2 also provides a basis for inference provided that consistent estimators $\widehat{V}$ and $\widehat{\Omega}_j$ for V and Ω_j are available. For example, we can estimate V by

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\pi}}(x,Z_i)\hat{\boldsymbol{\pi}}'(x,Z_i), \tag{8}$$

which equals to $\sum_{j=1}^{J} \hat{\omega}_j \hat{\boldsymbol{\pi}}(x,z_j)\hat{\boldsymbol{\pi}}(x,z_j)'$ when Z is discrete; we estimate Ω by

$$\widehat{\Omega} = \sum_{j=1}^{J} \hat{\omega}_j^2 \widehat{\Omega}_j \hat{\boldsymbol{\pi}}(x,z_j)'\hat{\boldsymbol{\pi}}(x,z_j),$$

where $\widehat{\Omega}_j$ is a consistent estimator of Ω_j. We use sample analog for $\hat{\omega}_j$ and $\hat{\pi}$. For $\widehat{\Omega}_j$, we adopt the estimator proposed in Fan and Gijbels (1996, Section 4.3). To save space, we collect the formulas and verify the consistency of $\widehat{V}$ and $\widehat{\Omega}$ in Corollary 2 in Appendix B. To illustrate how to conduct inference, suppose that the treatment is binary ($D \in \{0,1\}$). Then $\beta(s,x) = (\beta_0(s,x), \beta_1(s,x))'$ is two-dimensional and $\beta_1(s,x)$ is the primary parameter of interest. Let s_2^2 be the $(2,2)$th element of $\widehat{V}^{-1}\widehat{\Omega}\widehat{V}^{-1}$. Then we can construct a $(1-\alpha)\times 100\%$ confidence interval for $\beta_1(s,x)$ as $\left[\hat{\beta}_1(s,x) \pm \frac{s_2 \Phi^{-1}(1-\alpha/2)}{\sqrt{nh^{d_x+d_s+2}}}\right]$, where Φ is the standard normal CDF. Based on Corollary 2, this confidence interval is asymptotically valid at a $1-\alpha$ level.

We conduct Monte Carlo simulations to demonstrate the finite sample performance of our estimator. For the compactness of the main text, we leave the details of simulation designs and all the results (e.g. mean squared errors and converge frequencies) in the online supplementary material (please see Liu, Mourifie, and Wan, 2019, Section 5).

4. Empirical Application

4.1. Empirical Question and Data

In this section we illustrate our method by studying how the average return to education in China varies as a function of local contemporaneous average income, which can be viewed as a proxy of local development level. Improving and equalizing return to education across regions has always been an important policy concern in China, as Heckman (2005, pp. 52) stated: "*A policy that equalizes returns across all investment types and across all regions increases economic growth. Current Chinese policy tends to ignore this fundamental rule...*". To justify such policies, it is therefore important to propose an adequate econometric method to examine whether the returns to education indeed differs significantly across regions with different levels of development. This is the main objective of this empirical illustration.

To be more specific, we will estimate how the returns to middle and high school education vary across prefectures at different levels of development in China, as measured by the log of local average monthly income. For this purpose, we constructed a sample from China's 2005 One Percent Population Survey, also known as the "mini census". We exclude individuals who are below age 18 or above 60 since they are likely students or are retired. In the end, we retain a sample of 176,458 observations from 31 provinces and 343 prefectures, so that our sample is quite representative regionally. Please see **??** for sizes of subsamples from each province. The key variables in our analysis are log of monthly income (Y), education level (D), prefecture level, average contemporaneous monthly log-income (S), and mother's education (Z).

We group the endogenous education level into three categories. $D = 0$ represents "elementary school and below", $D = 1$ represents "middle school", and $D = 2$ represents "high school and above." This categorization is motivated by the fact that there are middle school entrance exams and high school entrance exams after six-years of elementary school education and three years of middle school education, respectively. We use mother's education as the outcome exclusion variable Z and group it into the same three categories (see Table 1).[14] Following, Eisenhauer, Heckman, and Vytlacil (2015, EHV) who used "current local wages" as treatment exclusion or benefit shifter according their language, we consider as treatment exclusion variable S the "current log of average income at the prefecture level". In the context of our application, S can be viewed as a proxy of the local level of development. Note that since S is at the prefecture level, it is independent of individual-level latent abilities. The summary statistics of Y and S are reported in Table 2.[15]

In the literature of estimating return to education using parents' education as an instrumental variable, one concern is that parents' education can be correlated with talent. Indeed, we test the necessary implication of the joint assumptions of (i) independence between Z and (U_1, U_0) and (ii) the selection $D = \vartheta(X, Z, \varsigma)$ be monotone in Z at the provincial level using the method of Mourifié and Wan (2017), we find that the test rejects the joint assumptions in a significant portion of provinces, meaning that the dataset under analysis here shows strong evidences against the use of mother's education as a conventional IV to estimate

[14] We combine college education and high school education due to data limitations for the mother's education variable. The individuals in the constructed sample are 18 to 60 years old in 2005; their mothers are about 25 years older and the portion of college education among these mothers is very small.

[15] In this section, we mainly treat S as a continuous random variables. However, since S is at the prefecture level, we also conduct the estimation by explicitly taking S to be discrete, where we estimate the $m(s, \tilde{s}, z)$ (see Equation 6) using the method proposed in Racine and Li (2004); Li, Racine, and Wooldridge (2009). The results are qualitatively similar to what we reported here and are available upon request.

Table 1: Education (D) and Mother's Education (Z): Subsamples

	Z			
D	0	1	2	Total
0	36,153	1,309	187	37,649
1	75,943	19,141	2,253	97,337
2	20,552	12,721	8,199	41,472
Total	132,648	33,171	10,639	176,458

Table 2: Log-Income (Y) and Prefecture Average Log-Income (S)

	Mean	Standard Deviation	Min-Max
Y	6.05	0.87	1.61–11.51
S	6.30	0.48	5.31-7.48

the LATE. Please see the online supplementary materials Liu, Mourifie, and Wan (2019, Section 3) for the full details of the test. One advantage of our approach is that we allow the mother's education to be dependent with (U_1, U_0), and that the selection $D = \vartheta(X, Z, \varsigma)$ be non-monotone in Z.

4.2. Estimates Based on the Proposed Method

Next we proceed with our method. The estimation results are reported in Figure 2 for under-smoothing constant $c = -0.01$. In each figure, the upper panel plots $\hat{\beta}_1(s)$, the CATE-PD and its 95% level confidence interval (piontwise), that is, the derivative of the causal effect of increasing the education level from elementary school to middle school with respect to s. The middle panel is the CATE-PD from middle school to high school, and the lower is the sum of the previous two: from elementary school to high school.

In addition to the main results, we also conduct robustness checks, including (i) consider controlling some demographic variables such as age, gender, and ethnic groups, (ii) consider the issue of internal migration, (iii) re-categorize the education levels, and (iv) consider other under-smoothing constant e.g. $c = 0$ and $c = -0.03$. The additional results carry the same quality messages and therefore reported in the supplementary materials Liu, Mourifie, and Wan (2019, Section 4).

The first main result is that CATE-PD significantly differ from 0 over most of s values in all three panels, meaning that the return to education is indeed different across regions/prefectures with different levels of development. This result suggests that schooling investment must be done at a different rate across those areas in order to equalize returns to education.

Another common feature across the three panels is that the CATE partial derivative tend to be negative for low value of s (our proxy of local development) and positive for the very high value of s, as graphically shown in Figure 3. This suggest that the return to education, as a function of s, has a U-shape. It suggests that the return to education is relatively high in very poor regions and rich regions. One possible explanation is that in the poor regions, the supply of highly educated people on the labor market is more scarce,

Figure 2: Estimates of Treatment Derivatives, $c = -0.01$

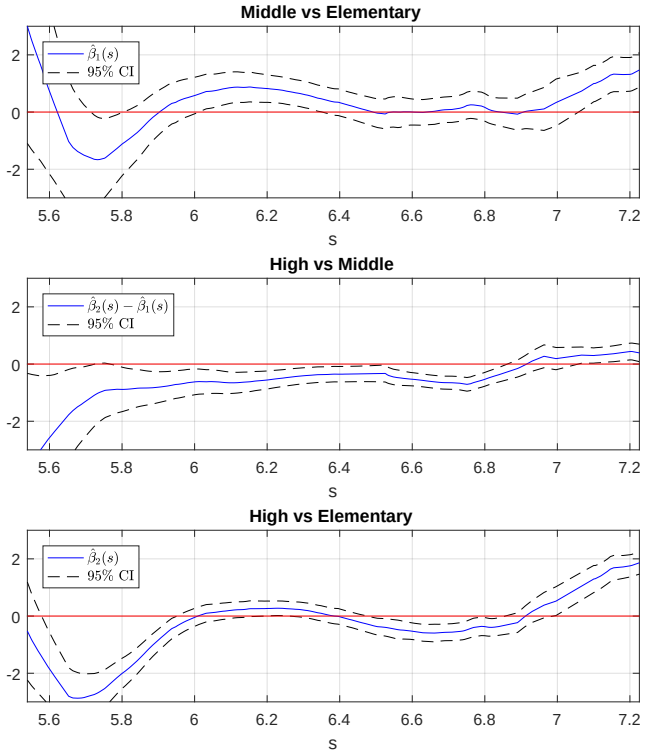
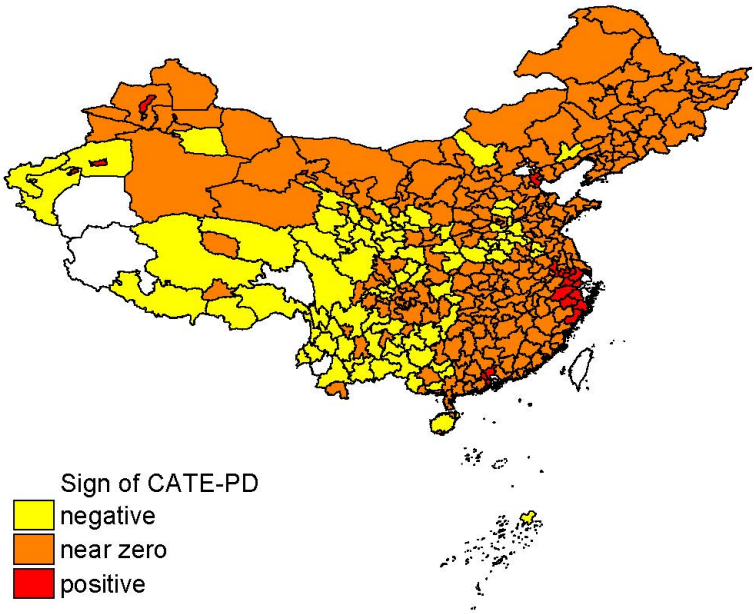


Figure 3: Estimates of CATE-PD High vs Elementary



which in turn causes a rise in their relative price (wages) in those local labor markets. This is consistent with the observation in Heckman (2005, pp. 63): "*.....Western China and rural areas currently have low incomes and hence poor support for education but a very high return to it....*" On the other hand, a possible explanation for higher return in richer regions is that higher wages is needed to compensate higher living costs in richer regions. It is also because high-paid job opportunities (such as those in financial sector) are more available in richer regions than other regions.

Finally, we can also see there is heterogeneity across the three panels of Figure 2. Let's consider the CATE-PD of middle school v.s. elementary school first. For most of middle-income prefectures, the 95% confidence interval contains zero, suggesting that the causal effect of increasing education from elementary school to middle school is about the same across those middle-income prefectures. On the other hand, the causal effect of increasing education from middle school to high school tends to decrease in the local income level within the middle-income regions.

5. Conclusion

In this paper we discussed identification and estimation of nonparametric structural functions in heterogenous treatment effect models with a discrete endogenous treatment. We focused on applications in which the potential outcome function is additively-separable with a treatment-specific structural error and in which there exists two-way exclusion variables: this frees us from assuming the candidate "instrumental variable" is independent of latent errors and from assuming the treatment response is monotone. We proposed a two-stage nonparametric estimators to consistently estimate the treatment effect derivatives. We also provided Monte Carlo simulations and an empirical illustration for our method.

References

Andrews, D. W., and M. M. Schafgans (1998): "Semiparametric estimation of the intercept of a sample selection model," *Review of Economic Studies*, pp. 497–517.

Barua, R., and K. Lang (2016): "School Entry, Educational Attainment, and Quarter of Birth: A Cautionary Tale of a Local Average Treatment Effect," *Journal of Human Capital*, 10(3), 347–376.

Chaisemartin, C. d. (2016): "Defying the LATE? Identification of Local Treatment Effects When the Instrument Violates Monotonicity," *Quantitative Economics, Forthcoming*.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

Chesher, A. (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73(5), 1525–1550.

Das, M. (2005): "Instrumental variables estimators of nonparametric models with discrete endogenous regressors," *Journal of Econometrics*, 124(2), 335–361.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric estimation of sample selection models," *The Review of Economic Studies*, 70(1), 33–58.

D'Haultfœuille, X., and A. Maurel (2013): "Inference on an extended Roy model, with an application to schooling decisions in France," *Journal of Econometrics*, 174(2), 95–106.

Dong, Y., and A. Lewbel (2015): "Identifying the Effect of Changing Policy Threshold in Regression Discontinuity Models," *The Review of Economics and Statistics*, 97(5), 1081–1092.

Dong, Y., and S. Shen (2017): "Testing for rank invariance or similarity in program evaluation," *Review of Economics and Statistics, forthcoming*.

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The generalized Roy model and the cost-benefit analysis of social programs," *Journal of Political Economy*, 123(2), 413–443.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Florens, J.-P., and L. Malavolti (2002): "Instrumental regression with discrete variables," Discussion paper, Mimeo.

Gautier, E., and S. Hoderlein (2015): "A triangular treatment effect model with random coefficients in the selection equation," *arXiv preprint arXiv:1109.0362v4*.

Heckman, J. (1990): "Varieties of selection bias," *The American Economic Review*, pp. 313–318.

Heckman, J. J. (2005): "China's human capital investment," *China Economic Review*, 16(1), 50–70.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding instrumental variables in models with essential heterogeneity," *The Review of Economics and Statistics*, 88(3), 389–432.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 73(3), 669–738.

Hsu, Y.-C., C.-A. Liu, and X. Shi (2016): "Testing Generalized Regression Monotonicity," Discussion paper, Institute of Economics, Academia Sinica, Taipei, Taiwan.

Hsu, Y.-C., and S. Shen (2016): "Testing for treatment effect heterogeneity in regression discontinuity design," Discussion paper, Institute of Economics, Academia Sinica, Taipei, Taiwan.

——— (2017): "Monotonicity Test for Local Average Treatment Effects Under Regression Discontinuity," Discussion paper, Institute of Economics, Academia Sinica, Taipei, Taiwan.

Huber, M., and G. Mellace (2015): "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 97(2), 398–411.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Jun, S. J., J. Pinkse, and H. Xu (2011): "Tighter bounds in triangular systems," *Journal of Econometrics*, 161(2), 122–128.

Kitagawa, T. (2008): "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model," Working paper.

——— (2015): "A Test for Instrument Validity," *Econometrica*, 83(5), 2043–2063.

Klein, T. J. (2010): "Heterogeneous treatment effects: Instrumental variables without monotonicity?," *Journal of Econometrics*, 155(2), 99–116.

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, 26(05), 1529–1564.

Li, Q., and J. S. Racine (2007): *Nonparametric econometrics: theory and practice*. Princeton University Press.

Li, Q., J. S. Racine, and J. M. Wooldridge (2009): "Efficient estimation of average treatment effects with mixed categorical and continuous data," *Journal of Business & Economic Statistics*, 27(2), 206–223.

Linton, O., and W. Härdle (1996): "Estimation of additive regression models with known links," *Biometrika*, 83(3), 529–540.

Liu, S., I. Mourifie, and Y. Wan (2019): "Two-way Exclusion Restrictions in Models with Heterogeneous Treatment Effects: Supplementary Materials," Discussion paper.

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187(1), 74–81.

Mourifié, I., and Y. Wan (2017): "Testing Local Average Treatment Effect Assumptions," *Review of Economics and Statistics*, 99(2), 305–313.

Oreopoulos, P. (2006): "Estimating average and local average treatment effects of education when compulsory schooling laws really matter," *The American Economic Review*, 96(1), 152–175.

Racine, J., and Q. Li (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119(1), 99–130.

Shaikh, A. M., and E. J. Vytlacil (2011): "Partial identification in triangular systems of equations with binary dependent variables," *Econometrica*, 79(3), 949–955.

Yu, P. (2017): "Testing Conditional Rank Similarity With and Without Covariates," *Working Paper*.

Appendix A. Proofs of Main Results

Appendix A.1. Proof to Proposition 1

The idea of proof is similar to Das (2005). To show the sufficiency, the identification Equation (4) implies that almost surely with respect to the conditional distribution of Z given $(S, X) = (s, x)$,

$$m(s, x, Z) = \boldsymbol{\pi}(x, Z)'\beta(s, x),$$

which in turn implies

$$\boldsymbol{\pi}(x,Z)m(s,x,Z) = \boldsymbol{\pi}(x,Z)\boldsymbol{\pi}(x,Z)'\beta(s,x).$$

Taking expectation with respect to the conditional distribution of Z gives

$$\mathbb{E}[\boldsymbol{\pi}(x,Z)m(s,x,Z)|S=s,X=x] = \mathbb{E}[\boldsymbol{\pi}(x,Z)\boldsymbol{\pi}(x,Z)'|S=s,X=x]\beta(s,x).$$

Since $\mathbb{V}[\boldsymbol{\pi}_0(x,Z)|S=s,X=x]$ is positive definite and

$$|\mathbb{E}[\boldsymbol{\pi}(x,Z)\boldsymbol{\pi}(x,Z)'|S=s,X=x]| = |\mathbb{V}[\boldsymbol{\pi}_0(x,Z)|S=s,X=x]| > 0,$$

where $|\cdot|$ denotes determinant of a matrix, we can obtain Equation (5) by pre-multiply $\{\mathbb{E}[\boldsymbol{\pi}(x,Z)\boldsymbol{\pi}(x,Z)'|S=s,X=x]\}^{-1}$ to both sides. The necessity holds because when $\mathbb{V}[\boldsymbol{\pi}_0(x,Z)|S=s,X=x]$ does not have full rank, there are multiple β satisfying Equation (4) a.s. in $Z|(S,X)=(s,x)$.

Appendix A.2. Proof to Proposition 2

Since at (s,x) and all $z_j \in \mathscr{Z}$, $\hat{\omega}_j$, $\hat{\boldsymbol{\pi}}(x,z_j)$, $\hat{m}(s,x,z_j)$ are consistent estimator for ω, $\boldsymbol{\pi}(x,z_j)$ and $m(s,x,z_j)$, respectively, it follows from continuous mapping theorem, Slutsky theorem and law of large number that

$$\hat{\beta}(s,x) = \left(\sum_{j=1}^{J}\omega_j\boldsymbol{\pi}(x,z_j)\boldsymbol{\pi}(x,z_j)'\right)^{-1}\left(\sum_{j=1}^{J}\omega_j\boldsymbol{\pi}(x,z_j)m(s,x,z_j)\right)+o_p(1) = \beta(s,x)+o_p(1).$$

For asymptotic normality, note first that for a given $z_j \in \mathscr{Z}$,

$$\begin{aligned}\sqrt{nh^{d_x+d_s+2}}&\left(\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)\hat{m}(s,x,z_j) - \omega_j\boldsymbol{\pi}(x,z_j)m(s,x,z_j)\right)\\ &= \sqrt{nh^{d_x+d_s+2}}\left(\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j) - \omega_j\boldsymbol{\pi}(x,z_j)\right)\hat{m}(s,x,z_j)\\ &\quad + \sqrt{nh^{d_x+d_s+2}}\left(\omega_j\boldsymbol{\pi}(x,z_j)\hat{m}(s,x,z_j) - \omega_j\boldsymbol{\pi}(x,z_j)m(s,x,z_j)\right).\end{aligned}$$

The first RHS term converges in probability to zero since $\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)-\omega_j\boldsymbol{\pi}(x,z_j) = O_p(1/\sqrt{nh^{d_x}})$. The second term converges in distribution to $N(0,\omega_j^2\Omega_j(s,x,z_j)\boldsymbol{\pi}(x,z_j)\boldsymbol{\pi}'(x,z_j))$ by standard result of local polynomial regression, for instance, see Li and Racine (2007, Theorem 2.10) and the fact that we are using under-smoothing.

Now consider another $z_{j'} \in \mathscr{Z}$, it is straightforward to see that $\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)\hat{m}(s,x,z_j)$ and $\hat{\omega}_{j'}\hat{\boldsymbol{\pi}}(x,z_{j'})\hat{m}(s,x,z_{j'})$ are independent given Assumption 4 and the definition of the estimator. Hence

$$\begin{aligned}\sqrt{nh^{d_x+d_s+2}}&\left(\sum_j\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)\hat{m}(s,x,z_j) - \sum_j\omega_j\boldsymbol{\pi}(x,z_j)m(s,x,z_j)\right)\\ &\xrightarrow{d} N(0,\sum_j\omega_j^2\Omega_j(s,x,z_j)\boldsymbol{\pi}(x,z_j)\boldsymbol{\pi}'(x,z_j)) \equiv N(0,\Omega).\end{aligned}$$

Finally, by observing that $\sum_{j=1}^{J} \omega_j \boldsymbol{\pi}(x, z_j)\boldsymbol{\pi}(x, z_j)' \xrightarrow{p} V$, it then follows that

$$\sqrt{nh^{d_x+d_s+2}}(\hat{\beta}(s,x) - \beta(s,x)) \xrightarrow{d} N(0, V^{-1}\Omega V^{-1}).$$

Appendix B. Variance Estimation in Section 3

The formula for $\Omega_j(s,x)$ is standard and can be obtained, for example, from Li and Racine (2007, Theorem 2.10). To be more specific, under our rate condition,

$$\sqrt{nh^{d_x+d_s+2}}(\hat{m}(s,x,z_j) - m(s,x,z_j)) \xrightarrow{d} N\left(0, \frac{\sigma_m^2(s,x,z_j)\kappa}{g_{(S,X)|Z}(s,x|z_j)}\right),$$

where $\sigma_m^2(s,x,z_j) = \mathbb{E}[\epsilon_m^2|S=s, X=x, Z=z_j]$ with $\epsilon_m \equiv Y - \mathbb{E}[Y|S,X,Z]$, and $g_{(S,X)|Z}(s,x|z_j)$ is the conditional density of (S,X) given $Z=z_j$ evaluated at (s,x), and κ is a constant which only depends on the kernel function and can be computed based on the formula in Li and Racine (2007, Section 2.7.3). One can estimate the variance matrix by estimating $g_{(S,X)|Z}(s,x|z_j)$ and $\sigma_m^2(s,x,z_j)$, respectively and consider plug in. We here adopt the approach suggested in Fan and Gijbels (1996, Section 4.2). We define some additional notation. Let $\mathbb{X}_j$ be a $n_j \times |\tilde{\alpha}^m|$ dimensional matrix, where $n_j = \sum_i \mathbf{1}\{Z_i = z_j\}$, $|\tilde{\alpha}^m|$ is the length of vector $\tilde{\alpha}^m$, or equivalently, number of terms in the polynomial $\mathcal{P}((S_i - s, X_i - x), \boldsymbol{\alpha}, p)$. Therefore, a typical row in $\mathbb{X}_j$ is

$$\left(1, X_i - x, S_i - s, (X_i - x)^2, \cdots, (S_i - s)^p\right),$$

where i is such that $Z_i = z_j$. Let $\mathbb{W}_j$ be a $n_j \times n_j$ diagonal matrix with a typical element $\boldsymbol{K}_{x,h}(X_i - x)\boldsymbol{K}_{s,h}(S_i - s)$ for corresponding i, and finally let $\mathbb{Y}_j$ be the n_j-vector of corresponding Y_is. Let e_s be a vector of zeros with the same dimension as the $\left(1, X_i - x, S_i - s, (X_i - x)^2, \cdots, (S_i - s)^p\right)$ and with its third element being replaced by 1. With this notation, it is easy to see that

$$\hat{m}(s,x,z_j) = e_s \left(\mathbb{X}_j'\mathbb{W}_j\mathbb{X}_j\right)^{-1} \mathbb{X}_j'\mathbb{W}_j\mathbb{Y}_j,$$

where e_s is a zeros vector in which the third element being replaced by 1. Under local homogeneity, the conditional variance of $\hat{m}(s,x,z_j)$ given $(S,X,Z=z_j)$ can be approximated as (see Fan and Gijbels, 1996, section 4.3)

$$e_s \left(\mathbb{X}_j'\mathbb{W}_j\mathbb{X}_j\right)^{-1} \left(\mathbb{X}_j'\mathbb{W}_j^2\mathbb{X}_j\right) \left(\mathbb{X}_j'\mathbb{W}_j\mathbb{X}_j\right)^{-1} e_s'\sigma_m^2(s,x,z_j).$$

Let $\hat{\sigma}_m^2(s,x,z_j)$ be a consistent estimator as describable in Fan and Gijbels (1996, Equation 4.49), that is,

$$\hat{\sigma}_m^2(s,x,z_j) = \frac{\sum_i (Y_i - \hat{Y}_i)\boldsymbol{K}_{s,h}(S_i - s)\boldsymbol{K}_{x,h}(X_i - x)\mathbf{1}[Z_i = z_j]}{\mathrm{Tr}\left(\mathbb{W}_j - \mathbb{W}_j\mathbb{X}_j\left(\mathbb{X}_j'\mathbb{W}_j\mathbb{X}_j\right)^{-1}\mathbb{X}_j'\mathbb{W}_j\right)},$$

then we can consistently estimate $\Omega_j(s,x)$ by

$$\widehat{\Omega}_j = nh^{d_x+d_s+2} e_s \left(\mathbb{X}_j'\mathbb{W}_j\mathbb{X}_j\right)^{-1} \left(\mathbb{X}_j'\mathbb{W}_j^2\mathbb{X}_j\right) \left(\mathbb{X}_j'\mathbb{W}_j\mathbb{X}_j\right)^{-1} e_s'\hat{\sigma}_m^2(s,x,z_j) \tag{B.1}$$

Corollary 2. *Let $\widehat{\Omega}_j$ be defined as in Equation* (B.1) *and $\widehat{V}$ be defined as in Equation* (8) *and suppose $\sigma_m(s,x,z_j)$ is constant in a small neighborhood of (s,x) for each z_j, then*

$$\left(\widehat{V}^{-1}\widehat{\Omega}\widehat{V}^{-1}\right)^{-\frac{1}{2}}\sqrt{nh^{d_x+d_s+2}}(\hat{\beta}(s,x)-\beta(s,x)) \xrightarrow{d} N(0,I).$$

PROOF. *Note first $\hat{\omega}_j \xrightarrow{p} \omega_j$. By Lemma 2 of the supplementary materials,* $\sup_{x,z}|\hat{\pi}_d(x,z)-\pi_d(x,z)| = o_p(1)$. *Therefore $\widehat{V} \xrightarrow{p} V$ since $\widehat{V} = \sum_{j=1}^{J}\hat{\omega}_j\hat{\boldsymbol{\pi}}(x,z_j)\hat{\boldsymbol{\pi}}(x,z_j)'$. In addition, since $\widehat{\Omega} = \sum_{j=1}^{J}\hat{\omega}_j^2\widehat{\Omega}_j\hat{\boldsymbol{\pi}}(x,z_j)'\hat{\boldsymbol{\pi}}(x,z_j)$, the consistency of $\widehat{\Omega}$ is implied by the consistency of $\widehat{\Omega}_j$, which holds by Fan and Gijbels (1996, Section 4.11).* □