Identifying Multiple Marginal Effects with a Single Binary Instrument or by Regression Discontinuity*

Carolina Caetano† *University of Rochester*

Juan Carlos Escanciano‡ *Indiana University*

October 5th, 2015.

Abstract

This paper proposes a new strategy for the identification of all the marginal effects of an endogenous multi-valued variable (which can be continuous, or a vector) in a regression model with one binary instrumental variable (IV). Identification is achieved by exploiting heterogeneity of the "first stage" in covariates that are separable from the endogenous variables of interest. The covariates themselves may be endogenous, and their endogeneity does not need to be modeled. With some modifications, the identification strategy is extended to the Regression Discontinuity Design (RDD). Ours is the first paper to obtain identification of nonparametric marginal effects in an RDD setting. This paper also provides parametric, semiparametric and nonparametric Two-Stage Least Squares (TSLS) estimators which are simple to implement, discusses their asymptotic properties, and shows that the estimators have satisfactory performance in moderate samples sizes. Finally, we apply our methods to the problem of estimating the effect of air quality on house prices, based on Chay and Greenstone (2005).

Keywords: Conditional Instrumental Variables; Endogeneity; Binary Instrument; Regression Discontinuity Design; Varying Coefficients; Nonparametric.

JEL classification: C13; C14; C21; D24

*We would like to thank Gregorio Caetano, Kenneth Chay and Matias Cattaneo for helpful discussions. We would also like to thank seminar participants at Boston College, Boston University, Harvard-MIT, Northwestern, University of Michigan, University of Colorado-Boulder and the 2015 World Congress of the Econometric Society for useful comments.

†Department of Economics, University of Rochester, 238 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627-0156, USA. E-mail: carol.caetano@rochester.edu. Web Page: http://www.carolinacaetano.net/.

‡Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/. Research funded by the Spanish Plan Nacional de I+D+I, reference number ECO2014-55858-P.

1 Introduction

Instrumental variables (IV) methods are well established as one of the most useful approaches to identify causal effects in econometric models. Consider the regression model

$$Y = g(X) + h(Z) + \varepsilon, \tag{1}$$

where Y is the dependent variable, g and h are unknown functions of X and Z respectively, and ε is an unobservable error term. X and Z are endogenous, so that $\mathbb{E}[\varepsilon|X, Z] \neq 0$ with positive probability. In addition to (Y, X, Z) we also observe an IV T which satisfies $\mathbb{E}[\varepsilon|T, Z] = \mathbb{E}[\varepsilon|Z]$.[1] We are interested in identifying the marginal effects of X on Y, given by g, while h is considered a nuisance parameter.

Depending on the nature of X and functional form assumptions on g and h, a traditional IV approach requires, among other things, that the IV T be sufficiently complex (see Newey and Powell (2003)). For instance, if X is continuous and we wish to identify g nonparametrically, then T must be continuous. If X is discrete with q points of support and we wish to identify all of its marginal effects, then T needs to have at least q points of support. If X is a vector of continuous variables, then T must have at least as many components as X, even if we impose substantial restrictions on the shape of g, such as linearity.

In this paper we propose a strategy for the identification of g in equation (1) (up to a constant) which applies to cases in which the support of X is larger than that of T (and thus it is impossible to achieve identification with a traditional IV approach.) For simplicity we focus on the case in which T is a binary variable, say $T \in \{0, 1\}$, while X takes 3 or more values, and may even be continuous, or a vector. In particular, our approach opens up the possibility of the identification of all the marginal effects of a complex variable X in cases where the instrument may be an experiment or a natural experiment.

Furthermore, we show that with some modifications our methodology can be extended to the Regression Discontinuity Design (RDD). To the best of our knowledge, we are the first to show how to identify all the marginal effects of a complex variable (e.g. continuous, discrete with more than two points of support, or a vector) using the RDD (see van der Klaauw (2008), Imbens and Lemieux (2008), Lee and Lemieux (2010), and Dinardo and Lee (2011) for surveys of the RDD literature, and Hahn, Todd and van der Klaauw (1990) for an analysis of identification of the classic RDD).

The identification strategy is based on the observation that if the population is categorized according to a covariate Z, the discrete variation that T induces on the distribution of X may vary with Z. This may allow us to recover a rich (e.g continuous) set of marginal effects. The following example illustrates some of the main ideas.

Example 1.1 *(Binary linear case) Suppose that $X = (X_1, X_2)$ and g is linear, so the model in (1) is*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + h(Z) + \varepsilon$$

where $\mathbb{E}[\varepsilon|X_1, X_2, Z] \neq 0$. Even if h is assumed to be zero, standard IV methods are unable to identify β_1 and β_2 with a single binary instrument T, because the classical order condition fails.

[1] Such an IV is referred to as a Conditional IV (CIV) in the literature. See, e.g., Frolich (2007) and Kasy (2009).

For our approach, we require that T be excluded from the structural equation conditional on Z, i.e. $\mathbb{E}[\varepsilon|Z,T] = \mathbb{E}[\varepsilon|Z]$, but Z itself may be endogenous. Then

$$\begin{aligned}\mathbb{E}[Y|T=1,Z] - \mathbb{E}[Y|T=0,Z] = \beta_1 \left[\mathbb{E}[X_1|T=1,Z] - \mathbb{E}[X_1|T=0,Z]\right] \\ + \beta_2 \left[\mathbb{E}[X_2|T=1,Z] - \mathbb{E}[X_2|T=0,Z]\right].\end{aligned}$$

To identify β_1 and β_2 we need to invert this equation. The condition that guarantees that we can invert it is the following: there exists no pair $(\lambda_1, \lambda_2) \neq (0,0)$ such that

$$\lambda_1 \left(\mathbb{E}[X_1|T=1,Z] - \mathbb{E}[X_1|T=0,Z]\right) + \lambda_2 \left(\mathbb{E}[X_2|T=1,Z] - \mathbb{E}[X_2|T=0,Z]\right) = 0 \quad a.s.$$

This condition states that the "first stage" effects of T on X_1 and X_2 vary (in a linearly independent manner) with Z. In other words, our identification strategy exploits the heterogeneity in the "first stages" to separate the marginal effects of X_1 and X_2.

Note that Z itself may be endogenous, and thus the nuisance function h is not identified. In fact, our approach does not require at all the specification of $h(Z)$, or even of $h(Z) + \mathbb{E}[\varepsilon|Z]$. Note also that the constant β_0 is not identified.

Based on our identification results, we propose parametric, semiparametric and nonparametric estimators of g (up to a constant). In models that are linear in parameters we show that our identification strategy in the binary IV can be straightforwardly implemented with a simple Two-Stage Least Squares (TSLS) estimator that treats the possibly endogenous covariate Z as if it were exogenous and uses interaction terms between Z and T as instruments. For the RDD, we consider a semiparametric varying coefficients specification, which also leads to a simple TSLS implementation and provides a practically convenient way to incorporate information on covariates' heterogeneity in applications, while dealing with the "curse of dimensionality" problem present when the dimension of Z is moderate or large, which is often the case in applications. Additionally, we propose nonparametric estimators that relax the functional form assumptions, and discuss the rates of consistency based on results by Blundell, Chen and Kristensen (2007) both for the IV and the RDD cases. Our identification strategy in the nonparameric cases can also be implemented as TSLS regressions with off-the-shelf econometric software.

Our approach has some connections with the local average treatment effect literature (see Imbens and Angrist (1994) and Angrist, Graddy and Imbens (2000)) and the related RDD literature (see Hahn, Todd and van der Klaauw (1990)). However, both literatures have traditionally focused on the causal impact of a binary endogenous variable and discrete or continuous instruments, whereas here we are interested in complex endogenous variables and one binary instrument. Our results also complement alternative identification strategies for binary instruments and continuous endogenous variables in Chesher (2003), D'Haultfoeuille and Fevrier (2014), Torgovitsky (2014), D'Haultfoeuille, Hoderlein and Sasaki (2013) and Masten and Torgovitsky (2014). None of these papers exploit the heterogeneity of the "first stage" conditional on possibly endogenous covariates.

The rest of the paper is organized as follows. Section 2 develops an intuitive example which shows explicitly what are the fundamental requirements of our method, as well as how our identification

approach works. Section 3 focuses on the binary IV case. It presents the identification results and also shows that the identification strategy can be implemented with a suitable TSLS estimator. Section 4 focuses on the RDD case. It extends the identification results of the binary IV case to the RDD, and it also proposes a TSLS estimator for a semiparametric varying coefficient specification. We present proofs in the Appendix discuss implementation in the fully non-parametric cases both in the binary IV case as well as in the RDD in the Appendix A.2. Section 5 reports the results of Monte Carlo experiments. Section 6 contains an empirical application of our method to the problem of estimating the effect of air quality on house prices, based on Chay and Greenstone (2005). This empirical example is particularly convenient because we are able to explore both a standard binary IV as well as an RDD design in the same problem, and using the same data. Finally, we conclude in Section 7, summarizing of our main findings and pointing out further applications of our results, which include an extension to non-separable models in unobservables (see also Remark 3.3) and tests of overidentification restrictions in models where classic IV provides just-identification.

2 An Example

The idea is best explained in an applied example. We will consider the problem of estimating the marginal effect of the amount a woman smokes during pregnancy (average daily number of cigarettes) on the baby's weight at birth (see Almond and Currie (2011) and Lumley et al. (2011) for discussions of the literature on this problem.) The variable of interest, "number of cigarettes," is naturally prone to endogeneity, given that there are many pre-existing selection factors associated both with smoking and with birth weight. A large literature has addressed this problem using experimental approaches (see Lumley et al. (2011)). Unfortunately, experiments provide only a binary instrument, and so it is impossible to uncover the effects of each marginal cigarette using the standard approach.

The following setup is entirely fictitious, but we believe that the association of our notation to a real problem can be helpful. Although the exposition below is rather informal, the rigorous arguments behind all claims can be seen in Example 3.1 in Section 3. In the maternal smoking context, the variable X represents the average number of cigarettes smoked per day during pregnancy, which may take several values. Suppose that an experiment implements an intervention in the treatment group that incentivizes women to reduce, and perhaps quit smoking, while the women in the control group are merely observed. The variable T is thus equal to one if the woman is in the treatment group, and zero otherwise. If we were able to observe the same woman under treatment and control, we could use the variation in the smoking dosages across the different women to recover the entire response function of birth weight to smoking. Precisely, we would learn the effect of the first cigarette from the women who smoked one cigarette, and then quit after the intervention. We would learn the effect of the second cigarette both from the women that smoked 2 cigarettes and then reduced to 1 cigarette under the intervention, or by subtracting the effect of the first cigarette from the effect found from the women that smoked 2 cigarettes and then quit after the intervention. We could continue with this strategy until we had the effects of each subsequent cigarette.

Unfortunately, we do not observe the same woman under both treatment and control. The main

insight of our method is that we can sometimes use a covariate to classify women into different groups, and thus generate artificially heterogeneous counterfactuals. For example, suppose that Z is the number of years of education of the mother. Although the experiment may have been performed without taking education into consideration at all, we can still separate women according to their education level, as long as this information is included in our data. The following table proposes a fictitious situation in which we do just this. Column (I) represents the years of education, and at first we are considering

Table 1: **Identification Idea**

(I) $\mathbf{Z}$	(II) $\bar{\mathbf{X}}_{\mathbf{0},\mathbf{Z}}$	(III) $\bar{\mathbf{X}}_{\mathbf{1},\mathbf{Z}}$	(IV) $\boldsymbol{\Delta}_{\mathbf{Y}}(\mathbf{Z})$	(V) $\mathbb{P}_{\mathbf{0},\mathbf{Z}}(\mathbf{0})$	(VI) $\mathbb{P}_{\mathbf{0},\mathbf{Z}}(\mathbf{1})$	(VII) $\mathbb{P}_{\mathbf{0},\mathbf{Z}}(\mathbf{3})$	(VIII) $\mathbb{P}_{\mathbf{1},\mathbf{Z}}(\mathbf{0})$	(IX) $\mathbb{P}_{\mathbf{1},\mathbf{Z}}(\mathbf{1})$	(X) $\mathbb{P}_{\mathbf{1},\mathbf{Z}}(\mathbf{3})$	(XI) row #
6	3	2	10	0	0	1	0	1/2	1/2	(1)
10	3	2	22	0	0	1	1/5	1/5	3/5	(2)
17	3	2	30	0	0	1	1/3	0	2/3	(3)

only 3 possibilities: 6, 10 and 17. Columns (II) and (III) show the average amount smoked by the women in the control and treatment groups, respectively, for the given number of years of education. Curiously, on average all groups reduced one cigarette because of the intervention. This is not a requirement of our method, we just want to show what can be achieved even when the "first stages" do not vary at all across the different values of Z. Column (IV) shows the average difference (in grams) in the birth weight between the treatment and the control groups for that level of education ($\Delta_Y(Z) = \mathbb{E}[Y|T=1,Z] - \mathbb{E}[Y|T=0,Z]$).

The first fundamental assumption of our method is a validity condition, which requires that $\mathbb{E}[\varepsilon|T,Z] = \mathbb{E}[\varepsilon|Z]$. It implies that within each education level we can compare treatment and control groups to obtain the causal effect of the intervention for each of the education groups ($\Delta_Y(Z) = \mathbb{E}[g(X)|T=1,Z] - \mathbb{E}[g(X)|T=0,Z]$). In this example, the smoking variable X takes the values 0, 1 or 3 (generalization to a more complex X will become clear later). Of course, in this multi-valued setting we cannot translate the effect of the intervention directly into information about the actual causal effects of smoking on birth weight, which is described by the function g, because the treatment and control groups for a given education level are only comparable as a whole.

We do, however, observe exactly what each woman consumed. Columns (V) to (VII) show the smoking distribution in the control group. For example, the numbers in row (1) are the fractions of the women educated 6 years and in the control group that smoked 0, 1 or 3 cigarettes, respectively. In this example everyone in the control group smokes 3 cigarettes (one can think of an experiment that specifically recruited women who smoked 3 cigarettes to participate). This is also not a requirement of the method, but it simplifies the explanations to have one less moving part.

Columns (VIII) to (X) show the corresponding fractions in the treatment group. As we can see, each row is different. It means that the intervention did not affect all groups in the same way. Even

though each education group had the same reaction to the intervention on average, a reduction of one cigarette, the distribution of behaviors is very different. Among the women educated 6 years, half reduced their smoking by 2 cigarettes, while the other half did not modify their behavior. The women educated 10 years had more divided reactions, with 3/5 keeping their old habit, while the remaining were divided, half quitting, and half reducing to one cigarette per day. Among the women educated 17 years, an even higher fraction did not modify their behavior, 2/3, but the ones that did all quit. The resulting distributions of smoking levels across the different education levels are very diverse. It is this variation in the distributions that is at the heart of our approach. It does not matter that all the average effects are the same, it would not even matter if there were no first stage effects at all. Our ability to identify the marginal effects comes from the fact that the instrument affected the distribution of X differently across the different Z, as we will show below.

This variation in the distributions also explains the outcome differences seen on column (IV). Even though the women behaved the same on average, if the effects are nonlinear the resulting effects across the groups can vary. For example, if the birth weight is affected disproportionately more by the first cigarette than the rest, we may find that the group in row (3), where a higher proportion of women quit, may have a much more pronounced effect in the birth weight than the group in row (1), where nobody quit. It is important to notice that the different numbers in column (IV) could also be explained by a non-separability of X and Z in the structural equation, for example if smoking could interact with education in such a way that a reduction of one cigarette was more effective among women educated 10 years than among women educated 6 years. This explanation is, however, ruled out indirectly by the separability between g and the other terms in the model in (1) together with the validity condition. We can thus write $\mathbb{E}[Y|T=1,Z]-\mathbb{E}[Y|T=0,Z]=\mathbb{E}[g(X)|T=1,Z]-\mathbb{E}[g(X)|T=0,Z]$. Therefore any differences across the rows of column (IV) must be due to nonlinearities of g combined with differences in the distribution of X across the different levels of education.

We will use the variations in distributions in the following manner: the treatment and control groups are comparable for each level of education, but only as a whole. Hence, the differences in the outcomes between treatment and control are the result of the effects of the possible smoking behaviors combined with the differences in the probabilities of each smoking behavior. For example, row (1) gives us $10=\mathbb{E}[Y|T=1,Z=6]-\mathbb{E}[Y|T=0,Z=6]=0\cdot g(0)+0.5\cdot g(1)+0.5\cdot g(3)-[0\cdot g(0)+0\cdot g(1)+1\cdot g(3)]=0.5\cdot g(1)-0.5\cdot g(3)$. The resulting equation provides some information about g, but combining this with the differences for other Z's, we can get a system of equations:

$$0.5g(1)-0.5g(3)=10 \tag{2}$$

$$0.2g(0)+0.2g(1)-0.4g(3)=22 \tag{3}$$

$$0.33g(0)-0.33g(3)=30 \tag{4}$$

Notice that in order to combine the information learned across all the different levels of Z into a common system, it is imperative that Z's own effect be separable from that of X. The separability of the model is thus not a simplification, it is vital for the method.

This system has 3 equations and 3 unknowns. However, only two equations are linearly independent (since $0.4(2)+0.6(4)=(3)$). In fact, if we had used more values of the variable Z, we could have more

equations, but it would not change the fact that at most two equations would be independent. This is caused by the fact that the coefficients of each of these equations always add up to zero, as can be easily verified in the example above. The reason for this phenomenon is that the coefficients come from the subtraction of probabilities. Since probabilities always add up to one, the subtraction of two sets of probabilities always adds up to zero. This is true not only for this example, but for all cases. For example, if X assumes q values, then the maximum number of linearly independent equations we can hope to get is $q-1$.

Since we have 2 linearly independent equations, we cannot recover the values of $g(0)$, $g(1)$, and $g(3)$, but we can recover the value of any increment. It is straightforward to see in this example that, from equation (2), $g(3)-g(1)=-20$, from equation (4), $g(3)-g(0)=-90$, and combining both results, $g(1)-g(0)=-70$. In a situation where X assumes more values, say q, we can get all the increments provided we have $q-1$ linearly independent equations. Hence, the second fundamental requirement of this method is a relevance condition which requires that the change in the distribution of behaviors between treatment and control groups differs for the Z's. If X takes q values, there must be enough variation for at least $q-1$ values of Z (indirectly it requires that Z must assume at least $q-1$ values).

We never discussed why would such variations in the change of distributions occur. This depends on the example, the particular intervention, and the chosen Z. For example, suppose that the intervention in our example requires that the women in the treatment group read extensive material. This intervention could affect women of different levels of education in different manners. Less educated women might be less likely to get through the material, and therefore to quit as a result of the intervention. In the fictitious example we created the more educated women were indeed the most likely to quit, even if on average the behaviors were the same across all levels of education. In a real example it is unlikely that even the average behavior will be the same across all groups, thus possibly generating even more variation.

3 The Case with a Binary Instrumental Variable

3.1 Identification

Throughout this section we assume the random vector (Y,X,Z,T) satisfies the model

$$Y = g(X) + h(Z) + \varepsilon, \tag{5}$$

where the following exclusion restriction holds

Assumption 1 *(validity)* $\mathbb{E}[\varepsilon|Z,T] = \mathbb{E}[\varepsilon|Z]$ *a.s.*

We can thus write

$$\mathbb{E}[Y|Z,T=1] - \mathbb{E}[Y|Z,T=0] = \mathbb{E}[g(X)|Z,T=1] - \mathbb{E}[g(X)|Z,T=0] \text{ a.s.} \tag{6}$$

Identifying g (up to location) from this implicit equation depends on our ability to invert it. To better understand the conditions that guarantee the invertibility of equation (6) consider first the following

example for the case where X is discrete. This example formalizes the discussion in Section 2, and it extends naturally to the general case.

Example 3.1 *(X and Z discrete) Denote by $\mathcal{S}_X := \{x_1, ..., x_q\}$ and $\mathcal{S}_Z := \{z_1, ..., z_l\}$ the supports of the distributions of X and Z, respectively, with $q < \infty$ and $l < \infty$.*

Our identification strategy consists of inverting equation (6), which in this context can be written as $\mathbf{m} = A\mathbf{g}$, where, $\mathbf{m} := (m(z_1), \ldots, m(z_l))'$, ($a'$ denotes the transpose of a) with $m(z) := \mathbb{E}[Y|T = 1, Z = z] - \mathbb{E}[Y|T = 0, Z = z]$, $z \in \mathcal{S}_Z$, the matrix A is given by $P_1 - P_0$, where $P_t = (p_{tij})$ is the $l \times q$ matrix with entries $p_{tij} = \mathbb{P}[X = x_j|T = t, Z = z_i]$, $i = 1, ..., l$, $j = 1, ..., q$ and $t = 0, 1$, and $\mathbf{g} := (g(x_1), ..., g(x_q))'$. Notice that since P_0 and P_1 are matrices of probabilities, $A\iota = 0$, where ι denotes the $q \times 1$ vector of ones. Therefore, A is not full-rank, and thus g is not identified from (6). However, in this context we can identify linear functionals $c'\mathbf{g}$ with c in a space of dimension $rank(A)$. In particular, if $rank(A) = q - 1$, then all linear functionals $c'\mathbf{g}$ with $c'\iota = 0$ are identified. In this case, all increment effects $g(x_h) - g(x_j)$, $h \neq j$, are identified. Of course, this is only possible if the order condition $l \geq q - 1$ holds, so Z needs to assume at least $q - 1$ different values.

In contrast, the classic nonparametric IV strategy is based on the equation

$$\mathbb{E}[Y|T] = \mathbb{E}[g(X) + h(Z)|T].$$

If $\mathbb{E}[h(Z)|T] = 0$ (for example if $h \equiv 0$ or if T is independent of Z), then the equation above translates into the system of equations $\mathbf{r} = P\mathbf{g}$, where $\mathbf{r} := (\mathbb{E}[Y|T = 0], \mathbb{E}[Y|T = 1])'$, and $P = (p_{tj})$ is the $2 \times q$ matrix with entries $p_{tj} := \mathbb{P}(X = x_j|T = t)$, $t = 0, 1$, $j = 1, \ldots, q$. In this classic setting the matrix P has a rank of at most 2, and so g is not identified if $q > 2$ (see Newey and Powell (2003)). In fact, we can only identify linear functionals $c'\mathbf{g}$ where c is spanned by the two rows of P (see Severini and Tripathi (2006, 2012)), which are not necessarily of interest. Having $\mathbb{E}[h(Z)|T] \neq 0$ does not reduce the identified set for g, as the number of estimating equations remains the same but the number of unknowns increases.

This discussion of Example 3.1 extends to the general case as follows. With some abuse of notation, we write equation (6) also as

$$m = Ag, \tag{7}$$

where now $Ag := \mathbb{E}[g(X)|Z, T = 1] - \mathbb{E}[g(X)|Z, T = 0]$ is a continuous (i.e. bounded) linear operator, $A : \mathcal{G} \subseteq L_2(X) \to L_2(Z)$, where henceforth, for a generic random vector ζ, $L_2(\zeta)$ denotes the Hilbert space of square-integrable functions with respect to the distribution of ζ. Here $\mathcal{G}$ is a subspace of $L_2(X)$ that may incorporate prior restrictions on g such as functional form restrictions or shape restrictions. We introduce our identification assumption as follows. Define $\mathcal{N}(A) = \{g \in \mathcal{G} : Ag = 0\}$, the null space of A. Our fundamental relevance condition requires that the null space of A is composed exclusively of the constant functions:

Assumption 2 *(relevance) $\mathcal{N}(A) = \{f \in \mathcal{G} : f \equiv c \in \mathbb{R}\}$.*

Notice that the identification condition in Example 3.1 that $rank(A) = q-1$ is equivalent to Assumption 2 in the discrete support case, since $\dim(\mathcal{N}(A)) + rank(A) = q$. In the general case, Assumption 2 is a nonparametric rank condition which is the analogue of the L_2-completeness condition required in nonparametric IV (see Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Andrews (2011) and D'Haultfoeuille (2011) for discussions on completeness). We compare formally these two identification assumptions in Remark 3.4 below.

Theorem 3.1 *Under Assumptions 1 and 2, g is identified up to location.*

Proof. Note that $A\widetilde{g} = m$, with $\widetilde{g}(X) = g(X) - \mathbb{E}[g(X)]$ and A is invertible on the orthocomplement of $\mathcal{N}(A)$, which by Assumption 2 is given by $\mathcal{N}^{\perp} = \{\lambda \in \mathcal{G} : \mathbb{E}[\lambda(X)] = 0\}$. Thus, since $\widetilde{g} \in \mathcal{N}^{\perp}$ it holds that $\widetilde{g} = A^{-1}m$. ■

In general, necessary conditions for Assumption 2 in the nonparametric case are that X and Z have the same level of complexity (e.g. both are continuous) and that they are not functionally dependent (nonparametric separability). Intuitively, what is needed for Assumption 2 to hold is that the differences between the distributions of X under "treatment" ($T = 1$) and "control" ($T = 0$) groups vary sufficiently with Z. The following examples illustrate the meaning of Assumption 2 in some important special cases.

Example 3.2 *(Gaussian variables) Suppose that (X, Z) is jointly normal conditionally on T, i.e.*

$$(X, Z)|T \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_T \\ \rho_T & 1 \end{pmatrix}\right).$$

Following Dunker, Florens, Hohage, Johannes and Mammen (2014), we can compute

$$\mathbb{E}[g(X)|Z = z, T = t] = (2\pi)^{-3/4} \exp\left(-\frac{z^2}{2}\right) \sum_{j=0}^{\infty} \mu^j(\rho_t) \mathbb{E}[g(X)p_j(X)] \frac{z^j}{\sqrt{j!}},$$

where p_j are the Hermite functions, $p_j(x) = (j!2\pi)^{-1/2} \exp\left(-0.5x^2\right) H_j(x)$, with H_j the j-th Hermite polynomial, and $\mu(\rho) = \rho/\sqrt{1-\rho^2}$. Therefore,

$$Ag(z) = (2\pi)^{-3/4} \exp\left(-\frac{z^2}{2}\right) \sum_{j=0}^{\infty} \left\{\mu^j(\rho_1) - \mu^j(\rho_0)\right\} \mathbb{E}[g(X)p_j(X)] \frac{z^j}{\sqrt{j!}}.$$

By the completeness of the Hermite polynomials, Assumption 2 in this context translates into $\rho_1 \neq \rho_0$ (with $\mathcal{G} = L_2(X)$). Notice that if $f_{X|T,Z}$ denotes the density of X conditional on T and Z, then

$$f_{X|T=t,Z=z}(x) = \frac{1}{\sqrt{2\pi(1-\rho_t^2)}} \exp\left(-\frac{(x-\rho_t z)^2}{2(1-\rho_t^2)}\right).$$

Therefore, the condition $\rho_1 \neq \rho_0$ is equivalent to saying that the difference between the distribution of X between "treatment" ($T = 1$) and "control" ($T = 0$) groups varies with Z.

Example 3.3 *(Linear multivariate model) Suppose that g is linear, so that the model is*

$$Y = \beta' X + h(Z) + \varepsilon, \quad \mathbb{E}[\varepsilon|Z,T] = \mathbb{E}[\varepsilon|Z], \tag{8}$$

where X is a d-dimensional vector which does not include a constant. Note that ε is not required to have zero mean, so it may include an intercept and/or functions of Z. In this model, the integral equation can be written as

$$\mathbb{E}[Y|Z,T=1] - \mathbb{E}[Y|Z,T=0] = \beta' \left(\mathbb{E}[X|Z,T=1] - \mathbb{E}[X|Z,T=0]\right),$$

or in short (using the generic notation $\Delta_V = \mathbb{E}[V|Z,T=1] - \mathbb{E}[V|Z,T=0]$)

$$\Delta_Y = \beta' \Delta_X. \tag{9}$$

Hence, in this example we take $\mathcal{G} = \{b'X : b \in \mathbb{R}^d\}$ and Assumption 2 is equivalent to

$$\mathbb{E}[\Delta_X \Delta_X'] \text{ is positive definite.}$$

It is straightforward to see why β is identifiable under this condition, since from equation (9) $\beta = (\mathbb{E}[\Delta_X \Delta_X'])^{-1} \mathbb{E}[\Delta_X \Delta_Y]$. In practice, this condition requires that the "first stages" of the several elements in the vector X vary with Z in a linearly independent manner. Notice that in linear models we can relax the conditions on the complexity of Z. For example, even though X is multivariate, Z may be univariate (though it must assume at least $d-1$ different values).

The following interesting examples are special cases of the linear multivariate model.

The discrete case: *In Example 3.1 we discussed the case where the endogenous variable is discrete. However, it is well known that is possible to analyze the discrete case within the framework of a linear model. Suppose that the endogenous variable D has support $\mathcal{S}_D := \{d_1, ..., d_q\}$, then*

$$\begin{aligned} g(D) &= g(d_1) + (g(d_2) - g(d_1))1(D = d_2) + \cdots + (g(d_q) - g(d_1))1(D = d_q) \\ &\equiv \alpha + \beta' X, \end{aligned}$$

where $\alpha = g(d_1)$, $X = (1(D = d_2), ..., 1(D = d_q))'$, and $\beta = (g(d_2) - g(d_1), ..., g(d_q) - g(d_1))'$ denotes the increment effects of interest.

Nonlinear cases: *Several nonlinear models cans still be written as linear-in-parameters models, such as, for example, the piecewise linear model. The simplest case is*

$$Y = \beta_1 D + \beta_2 D1(D > 0) + h(Z) + \varepsilon,$$

where D is an endogenous variable. In the application Section 6 we discuss connected piecewise linear models with more than two pieces, which can also be modeled similarly. We can also include some models with infinite, but parametric, variation in the marginal effects, such as

$$Y = \beta_1 D + \beta_2 D^2 + h(Z) + \varepsilon.$$

Classic IV identification conditions require two instruments in all examples above. In contrast, our identification assumption can still be applied when only T is used as instrument. These models, although nonlinear in variables, are linear in parameters, and therefore they can be treated as the linear model (8) above, simply defining $X = (D, D1(D > 0))'$ and $X = (D, D^2)'$, respectively.

Remark 3.1 *(Relation to panel data models) Consider the panel data model for $t = 0, 1$*

$$Y_t = g(X_t) + \eta + u_t$$

where $\mathbb{E}[u_t|X_0, X_1] = 0$. Then,

$$Y_1 - Y_0 = g(X_1) - g(X_0) + u_1 - u_0,$$

and we can identify g up to a constant from $\mathbb{E}[Y_1 - Y_0|X_0, X_1]$. We can understand this heuristically as an IV model with $T = t$ as the IV. Even though the IV is binary, we can identify g's derivatives in this case because we observe the counterfactuals X_0 and X_1 for the same observation. In the standard IV setting we cannot observe the counterfactuals. Our idea is to generate artificial counterfactuals by classifying observations according to a covariate Z so that we can apply a similar strategy.

Remark 3.2 *(Separability and the choice of Z) The classification variables Z are chosen by the researcher, and often there can be many options. The main consideration in the choice of Z is that it must be separable from X in the structural equation. It is important to notice that not all covariates in the model need to be separable from X, just those which will be used as Z in our identification approach. To better understand how to make this choice, consider the extended model*

$$Y = g(X, Z_c) + h(Z) + \varepsilon,$$

where $\mathbb{E}[\varepsilon|T, Z_c, Z] = \mathbb{E}[\varepsilon|Z]$. In this model the researcher chose to separate the variables into two groups. The variables Z_c are included into the model as exogenous controls. The variables Z are used as the classification variable.

The first question is why would the researcher include Z_c in the first place. In some cases it may be important to include such controls either because of suspected nonseparable effects or because it is hard to argue the validity of the IV unless it is conditional on Z_c as well.

The next question is which variables should be used as Z and which should be used as Z_c. There is a trade-off: on the one hand, Z can be endogenous, while Z_c must be exogenous. On the other hand, Z must be separable from X, while Z_c may interact with X in arbitrary ways.

Notice that Z needs to be only as complex as X, and so, for example if X is one continuous variable, it suffices to find just one continuous covariate which can be argued to be separable from X. The application Section 6 provides an explicit example of such considerations in an empirical problem. Note that if we assume a semiparametric structure for g, for example if $g(X, Z_c) = \beta(Z_c)'X$, then we can often drop the exogeneity requirement for Z_c (see Remark 3.3 below).

Remark 3.3 *(Non-separability between X and Z) Although our method cannot identify fully nonparametric non-separable heterogeneous effects in X and Z when X is not binary, i.e. g cannot be a fully nonparametric function of both (see Newey and Powell (2003)), we can extend our results to some non-separable models. For example, consider the following examples.*

Example 3.4 *(Linear model with heterogeneous effects in Z) Consider the varying coefficient model*

$$Y = \beta(Z)'X + h(Z) + \varepsilon, \quad \mathbb{E}[\varepsilon|Z,T] = \mathbb{E}[\varepsilon|Z],$$

where X is a d-dimensional vector. In this model, we can write

$$\Delta_Y = \beta(Z)'\Delta_X,$$

which can be used to identify $\beta(Z)$ provided $\Delta_X\Delta_X'$ is non-singular with positive probability. Under this assumption, we can estimate $\beta(\cdot)$ nonparametrically from local least squares regressions, similar to those carried out for the RRD below. Alternatively, we could specify $\beta(Z)$ as a linear function of Z, say $\beta(Z) = \beta_0 + \beta_1'Z$, which results in a linear-in-parameters model with endogenous variables X and interactions between X and Z, which can be dealt with as in Example 3.3 above. By comparing $\beta(\cdot)$ with the estimator obtained from $\beta = (\mathbb{E}[\Delta_X\Delta_X'])^{-1}\mathbb{E}[\Delta_X\Delta_Y]$ we can test for heterogenous marginal effects in Z. In the linear specification, this can be done by simply testing if $\beta_1 = 0$. For alternative models with random coefficients of endogenous variables see e.g. Hoderlein, Holzmann and Meister (2015) and references therein.

Example 3.5 *(non-separability in unobservables) Consider the model*

$$Y = g(X, h(Z) + \varepsilon), \tag{10}$$

where $\mathbb{E}[\varepsilon|X] \neq 0$ with positive probability, and ε is a structural unobservable error term. Assume g is monotonic in the second argument and denote by g^{-1} its inverse, i.e. $g^{-1}(Y,X) = h(Z) + \varepsilon$. Then, by the exclusion restriction $\mathbb{E}[\varepsilon|Z,T] = \mathbb{E}[\varepsilon|Z]$, it holds

$$\mathbb{E}[g^{-1}(Y,X)|Z,T=1] - \mathbb{E}[g^{-1}(Y,X)|Z,T=0] = 0 \ a.s. \tag{11}$$

Under suitable conditions, including the scale normalization $g^{-1}(y,\bar{x}) = y$ for some known $\bar{x}$ in the support of X (see Matzkin (2003)), we can identify g^{-1} from (11). For example, in the discrete setting of Example 3.1 if Y also has a discrete support with q_Y points, then the order condition for identification of g^{-1} under the normalization is $l \geq q_Y \cdot (q-1)$, where l and q are as in Example 3.1.

Remark 3.4 *It should be noted that Assumption 2 is different from the completeness between X and Z, but it can be understood as a weighted completeness or covariance completeness assumption between these two variables. To see this, define the propensity score $p(z) = \mathbb{E}[T|Z=z]$, assume $0 < p(z) < 1$ a.s., and note that Assumption 2 is equivalent to the implication: $\mathbb{E}[\lambda(X)(T - p(Z))|Z] = 0$ a.s. $\Longrightarrow \lambda(X) = \mathbb{E}[\lambda(X)]$ a.s. In contrast, the classical L_2 completeness between X and Z is equivalent to $\mathbb{E}[\lambda(X)|Z] = 0$ a.s. $\Longrightarrow \lambda(X) = 0$ a.s. (see Newey and Powell (2003)). Furthermore, our identification*

and estimation can be cast as nonparametric IV with "instrument" Z in the transformed model $Y(T - p(Z)) = g(X)(T - p(Z)) + \varepsilon(T - p(Z))$. That is, our identifying assumption requires that we can uniquely solve g (up to a constant) from the equation $Cov(Y, T|Z) = Cov(g(X), T|Z)$. This explains our terminology of covariance completeness.

Remark 3.5 *We present our method not as an alternative to the classic IV, but as a way to explore separability to identify quantities which cannot be identified with classical IV methods. However, in some instances, even when it is possible to use a classical IV approach, our method may have some advantages over classic IV. For example, in a linear model with a single endogenous variable and a single binary IV, a classical IV approach may be used. However, the IV may be weak, in the sense that $\mathbb{E}[X|T = 1] - \mathbb{E}[X|T = 0]$ could be small, but $\mathbb{E}[\Delta_X^2]$ could still be large. Consider the following example. Suppose that Z is a binary variable which can assume values $\{z_1, z_2\}$, with $\mathbb{P}(Z = z_1) = 1/2$. Suppose that $\mathbb{E}[X|T = 1, Z = z_1] - \mathbb{E}[X|T = 0, Z = z_1] = 10$, and $\mathbb{E}[X|T = 1, Z = z_2] - \mathbb{E}[X|T = 0, Z = z_2] = -10$. In this example, the instrument had a strong, but opposed effect on the subpopulations defined by Z. This is an extreme case in which there is no "first-stage" in the classical IV approach, since $\mathbb{E}[X|T = 1] - \mathbb{E}[X|T = 0] = 0$. However, the strength of our method (in the same units) is given by $\sqrt{\mathbb{E}[\Delta_X^2]}) = \sqrt{\mathbb{E}[(10 \cdot 1(Z = z_1) - 10 \cdot 1(Z = z_2))^2]} = 10$.*

3.2 Estimation

In this section we discuss estimation when g and $h(Z) + \mathbb{E}[\varepsilon|Z]$ are linear in parameters. To see the discussion of the estimation in the nonparametric case, refer to Appendix A.2.1. As shown in the Appendix, the nonparametric sieve estimator is also linear in parameters, therefore the implementation of this section is also relevant for the nonparametric case.

Our identification strategy in the linear multivariate model is based on the identity $\Delta_Y = \beta' \Delta_X$ as derived in Example 3.3. This identity suggests a three-step estimator for β : Step 1, estimate Δ_X by $\hat{\Delta}_X$ using regression methods; Step 2, estimate Δ_Y by $\hat{\Delta}_Y$ using regression methods; and Step 3, run a regression of $\hat{\Delta}_Y$ on $\hat{\Delta}_X$ by ordinary least squares (OLS) to obtain an estimate $\widehat{\beta}_{3Step}$.

It turns out that this estimation strategy can be easily implemented as a TSLS. Specifically, given a random sample $\{(Y_i, X_i, Z_i, T_i)\}_{i=1}^n$ of (Y, X, Z, T), we propose to estimate β with the coefficient of the X_i on a TSLS regression of Y_i onto X_i and Z_i, using T_i and $T_i Z_i$ as instruments for X_i, and treating Z_i as exogenous. This estimator, which we denote by $\widehat{\beta}$, can be implemented with off-the-shelf econometric software. Additionally, the standard errors are correctly estimated as the standard errors of the TSLS regression proposed above, without the need for any correction.

To see why $\widehat{\beta}_{3Step} = \widehat{\beta}$, suppose that $g(X) = \beta' X$ and $h(Z) + \mathbb{E}[\varepsilon|Z] = \alpha + \gamma' Z$. Then, we can write the model (1), by adding and subtracting $\mathbb{E}[\varepsilon|Z]$, as

$$Y = \alpha + \beta' X + \gamma' Z + u, \tag{12}$$

where $u = \varepsilon - \mathbb{E}[\varepsilon|Z]$. We can see from this representation that Z is different from a classical "control" variable, since X is still correlated with u even after controlling for Z. Note that our validity condition,

$\mathbb{E}[\varepsilon|Z,T] = \mathbb{E}[\varepsilon|Z]$, can be equivalently written as

$$\mathbb{E}[u|Z,T] = 0 \text{ a.s.},$$

and thus β in (12) can be identified as in a standard instrumental variable model, and estimated with a TSLS regression as $\hat{\beta}$ described above, provided the relevance condition holds; see Theorem 3.2 below.

Explicitly, let the "first-stage" regression fitted value be $\widehat{\mathbb{E}}[X|Z,T] = \widehat{\alpha}_{0X} + \widehat{\alpha}_{1X}T + \widehat{\alpha}'_{2X}Z + \widehat{\alpha}'_{3X}ZT$ and the "reduced-form" fitted value be $\widehat{\mathbb{E}}[Y|Z,T] = \widehat{\alpha}_{0Y} + \widehat{\alpha}_{1Y}T + \widehat{\alpha}'_{2Y}Z + \widehat{\alpha}'_{3Y}ZT$, then it is well known that the TSLS estimators are related to the reduced form fits through the equation

$$\widehat{\mathbb{E}}[Y|Z,T] = \widehat{\alpha} + \widehat{\beta}'\widehat{\mathbb{E}}[X|Z,T] + \widehat{\gamma}'Z.$$

If we evaluate this empirical equation at $T = 1$ and $T = 0$ and subtract, we arrive at

$$\hat{\Delta}_Y = \widehat{\beta}'\hat{\Delta}_X.$$

Thus, by definition of OLS, $\widehat{\beta}_{3Step}$ must be equal to $\widehat{\beta}$.

The asymptotic distribution of $\widehat{\beta}$ is well-known, so the novelty of the next result is in the consistency argument under our identification assumptions. Its proof can be found on Appendix A.1.1.

Theorem 3.2 *Let Assumptions 1 and 2 hold with $\mathcal{G} = \{b'X : b \in \mathbb{R}^d\}$. Then,*

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d N(0, \Sigma),$$

where Σ is the classical TSLS asymptotic variance (which is assumed to be finite).

Remark 3.6 *The arguments above show that the TSLS will be consistent for β even when $\mathbb{E}[X|T,Z]$ and $h(Z) + \mathbb{E}(\varepsilon|Z)$ are non-linear, as long as $g(X)$ is a linear function of X and our identifying assumptions hold.*

Remark 3.7 *It can be shown that the TSLS $\widehat{\gamma}$ is a consistent estimator of γ. However, γ is generally not a structural parameter. The effect of Z on Y when Z is endogenous is given by $h(Z)$, which is not identified. Even if we suppose $h(Z) = \gamma'_z Z$, the parameter γ_z is generally not identified in our setting.*

Remark 3.8 *The choice of Z with regards to identification was discussed in Remark 3.2. However, when choosing Z, further consideration should be given to the variance of the estimator which would result from each possible choice. In the parametric setting of this section, the relevance condition is a standard TSLS rank condition that can be tested by traditional methods. The order condition here is that dimension of Z needs to be at least $d-1$ (the dimension of X minus one). Therefore, if there are several variables which satisfy the identification conditions, we recommend that Z be chosen as the variable (or variables) for which T and TZ make up the strongest instruments.*

4 The Regression Discontinuity Design Case

4.1 Identification

The notation in the Regression Discontinuity standard framework follows from that of the treatment effects literature (Imbens and Angrist (1994)). As such, it is usually concerned with the effects of a binary intervention. However, one can envision many applications of the RDD methodology where the endogenous variable of interest can take multiple values (see e.g. Chay and Greenstone (2005), Carpenter and Dobkin (2009) and Brollo, Nannicini, Perotti and Tabellin (2013)). We need to extend the notation to allow for this generalization. On that account we will generalize the notation of Hahn, Todd and van der Klaauw (1990) for the constant treatment effects case in the fuzzy design. The method can be extended to variable treatment effects analogously to the discussion in Remark 3.3. We study the fuzzy design because the endogenous variable in our setting may take potentially any value both below and above the threshold.

Let the potential outcome random variable $Y(x)$ satisfy the model[2]

$$Y_i(x) = g(x) + h(Z_i) + \varepsilon_i. \tag{13}$$

Let the "predictor" or "running" variable be denoted by W, which is univariate and continuously distributed, with threshold $\bar{w}$. The quantities of interest are $\mathbb{E}[Y_i(x) - Y_i(x')|W = \bar{w}]$, for $x \neq x'$. Given equation (13) and our assumptions below, these are the same as the increments $g(x) - g(x')$.

As in Section 3, we will use a covariate Z to classify the observations. The validity condition is the same as in the standard RDD setting, after conditioning on Z.

Assumption 3 *(validity)* $\mathbb{E}[\varepsilon|W = w, Z]$ *is continuous in* w *at* $\bar{w}$ *with probability one.*[3]

Then, assuming the limits involved are well defined, we obtain

$$\lim_{w\downarrow\bar{w}} \mathbb{E}[Y|W = w, Z] - \lim_{w\uparrow\bar{w}} \mathbb{E}[Y|W = w, Z] = \lim_{w\downarrow\bar{w}} \mathbb{E}[g(X)|W = w, Z] - \lim_{w\uparrow\bar{w}} \mathbb{E}[g(X)|W = w, Z].$$

The right hand side defines implicitly a linear operator of g. We denote it by Ag, with some abuse of notation. Assumption 2 can be interpreted in the context of the RDD analogously to the cases discussed in Section 3.1. Therefore, if Assumptions 2 (for the new A) and 3 hold, g is identified up to a constant (see Theorem 3.1 for the proof). The fundamental requirement is that the difference between the distribution of X conditional on W and Z at the limits from above and below $\bar{w}$ vary sufficiently with Z. The following example shows the translation of this condition for the multivariate linear case.

Example 4.1 *(Linear model) Consider the linear model (where* X *may be multivariate) under Assumption 3,*

$$Y = \alpha + \beta' X + h(Z) + \varepsilon,$$

[2] Hahn, Todd and van der Klaauw (1990) denote the potential outcome $y_i = \alpha_i + x_i \cdot \beta$. In our case, α is represented by $h(Z) + \varepsilon$ and $g(x)$ is the generalization of $x \cdot \beta$.

[3] Assumption (A1) in Hahn, Todd and van der Klaauw (1990).

so that

$$\lim_{w\downarrow\bar{w}} \mathbb{E}[Y|W=w,Z] - \lim_{w\uparrow\bar{w}} \mathbb{E}[Y|W=w,Z] = \beta' \left(\lim_{w\downarrow\bar{w}} \mathbb{E}[X|W=w,Z] - \lim_{w\uparrow\bar{w}} \mathbb{E}[X|W=w,Z] \right). \quad (14)$$

Denote, for a generic random vector V*,*

$$\delta_V = \lim_{w\downarrow\bar{w}} \mathbb{E}[V|W=w,Z] - \lim_{w\uparrow\bar{w}} \mathbb{E}[V|W=w,Z],$$

assuming the limits exist. Then equation (14) can be written shortly as $\delta_Y = \beta'\delta_X$*. Assumption 2 in this context is equivalent to* $\mathbb{E}[\delta_X\delta_X']$ *being positive definite. In other words, we need that the variation around the threshold in the mean of* X *for a given* Z *vary in a linearly independent way.*

To better understand the meaning of the relevance condition above, consider the following simplified example in which X *is univariate. In this case we can apply the classic RDD identification strategy as well as ours (in the general case above where* X *is multivariate identification with the classic RDD is not possible). Assume the specification*

$$X = \alpha_0 + \alpha_1 1(W \geq 0) + \alpha_2 Z + \alpha_3 1(W \geq 0) \cdot Z + \varepsilon_X,$$

where (ε_X, W, Z) *are independent and identically distributed as standard normals. Note that with this design and with* $\bar{w} = 0$*,*

$$\delta_X = \lim_{w\downarrow 0} \mathbb{E}[X|W=w,Z] - \lim_{w\uparrow 0} \mathbb{E}[X|W=w,Z] = \alpha_1 + \alpha_3 Z,$$

while since $\mathbb{E}[Z|W] = 0$*,*

$$\lim_{w\downarrow 0} \mathbb{E}[X|W=w] - \lim_{w\uparrow 0} \mathbb{E}[X|W=w] = \alpha_1.$$

If $\alpha_1 \neq 0$ *both classic RDD and our new RDD identify* β*. If in addition* $\alpha_3 \neq 0$*, our RDD overidentifies* β*, which may result in an increase of precision relative to classic RDD. However, if* $\alpha_1 = 0$ *the classical RDD identification fails, but our method identifies* β *if* $\alpha_3 \neq 0$*, since* $\mathbb{E}[\delta_X^2] = \alpha_3^2 > 0$*. These theoretical results are confirmed in our simulations below.*

4.2 Estimation

In this section we discuss estimation for the RDD case when g is linear, as in Example 4.1. To see the discussion of the estimation in the nonparametric case, refer to Appendix A.2.2. We write

$$Y = \alpha + \beta' X + h(Z) + \varepsilon, \quad (15)$$

where $\mathbb{E}[\varepsilon|W=w,Z]$ is continuous in w at $\bar{w}$ with probability one, and satisfies further conditions below (see Assumption 4 in the Appendix). Without loss of generality we take hereinafter $\bar{w} = 0$.

We assume that the first stage is given by the semiparametric varying coefficient models

$$\mathbb{E}[X|Z,W] := \alpha_{0X}(W) + \alpha_{1X}'(W)Z, \quad (16)$$

where $\alpha_{0X}(\cdot)$ and $\alpha_{1X}(\cdot)$ are unknown functions of W. This specification overcomes the "curse of dimensionality" problem when the dimension of Z is large, as can be the case in applications.

Model (16) provides a natural extension of the model with interactions that we used in the binary IV case to the continuous case. In fact, if we approximate $\alpha_{0X}(\cdot)$ and $\alpha_{1X}(\cdot)$ locally at $W = 0$ by constants to each side of the threshold, the resulting model is exactly a model with interactions where the binary IV is $T_i = 1(W_i \geq 0)$. This suggests that a local version of the proposed TSLS estimator in the IV case (see Section 3.2) will provide valid inference on β.

However, it is well known that the local constant kernel estimator has generally worse bias properties than the local linear kernel estimator at discontinuity points; see Fan and Gijbels (1996). For this reason, we suggest implementing our identification strategy for the RDD with a local linear estimator instead. This corresponds to the use of linear (as opposed to just constants) approximations for $\alpha_{0X}(\cdot)$ and $\alpha_{1X}(\cdot)$ around each side of the threshold $\bar{w} = 0$. More generally, we can apply higher order approximations, resulting in local polynomial estimators of a certain degree. All these estimators can be implemented as TSLS estimators by defining a suitable vector of instruments, which can be seen similarly to the binary IV case.

Specifically, consider the uniform kernel

$$k_{h_n i} = 1(-h_n \leq W_i \leq h_n), \tag{17}$$

where h_n is a bandwidth parameter satisfying that $h_n \to 0$ as $n \to \infty$ and other conditions below. Then, our estimator for RDD can be easily implemented by restricting the sample $\{(Y_i, X_i, Z_i, W_i)\}_{i=1}^n$ to only those observations such that $-h_n \leq W_i \leq h_n$, and running a TSLS regression of Y_i onto X_i and Z_i, treating Z_i as exogenous, and using

$$V_i = (T_i, Z_i \cdot T_i, W_i, T_i \cdot W_i, Z_i \cdot W_i, Z_i \cdot T_i \cdot W_i)' \tag{18}$$

as a vector of "instruments" for X_i. Denote by $\tilde{\beta}$ the corresponding TSLS estimator of the coefficient of X, then the asymptotic behavior of $\tilde{\beta}$ is described in the following theorem.

Theorem 4.1 *Let Assumption 2 hold with $\mathcal{G} = \{b'X : b \in \mathbb{R}^d\}$, as well as Assumption 3 and Assumption 5 in Section A.1.2 in the Appendix. Then*

$$\sqrt{nh_n}(\tilde{\beta} - \beta) \to_d N\left(0, Q_2 \Omega Q_2'\right),$$

where expressions for Q_2 and Ω are given in Section A.1.2.

The proof of this theorem for a slightly generalized version of $\tilde{\beta}$ in which a general kernel, not necessarily the uniform kernel (17), is used can be found in Section A.1.2 in the Appendix.

The asymptotic variance of $\tilde{\beta}$ can be consistently estimated by the standard TSLS asymptotic variance (see Imbens and Lemieux (2008) for related discussion). Implementation of $\tilde{\beta}$ requires the choice of a bandwidth parameter h_n. We can choose h_n along the lines suggested in Calonico, Cattaneo and Titiunik (2014), which can be adapted to the presence of covariates.

Note that the way in which we introduce covariates in our RDD approach is different from how is traditionally done, which does not account for interaction terms between the covariates and the running variable; see, e.g., Imbens and Lemieux (2008). Our arguments suggest that, provided the separability conditions that we impose are satisfied, including covariates in this way may allow us to identify quantities which would otherwise not be identifiable with the current state-of-the-art in the RDD literature.

5 Monte Carlo Simulations

5.1 The case with a binary IV

Consider the following Data Generating Process (DGP):

$$\begin{aligned} Y &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_z Z + u, \\ Z &= \alpha_z u + \varepsilon_z, \\ X_1 &= \alpha_{01} + \alpha_{11} T + \alpha_{21} Z + \alpha_{31} T \cdot Z + \varepsilon_1, \\ X_2 &= \alpha_{02} + \alpha_{12} T + \alpha_{22} Z + \alpha_{32} T \cdot Z + \varepsilon_2, \end{aligned}$$

where $(u, \varepsilon_z, \varepsilon_1, \varepsilon_2)$ are independent standard normal random variables independent of T, which is distributed as Bernoulli random variable with probability $p = 0.5$. This is a linear model with three endogenous variables and one binary instrument. The classical order condition of standard IV does not hold in this case, and hence, classical IV is unable to identify the marginal effects β_1 and β_2. For identification of the marginal effects, our method requires that Assumption 2 holds, which in this case is equivalent to

$$\det \begin{vmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{12} & \alpha_{32} \end{vmatrix} \neq 0.$$

The parameters in the structural equation are set at $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 2$ and $\gamma_z = 1$. We set $\alpha_{01} = \alpha_{02} = \alpha_{21} = \alpha_{22} = 0$, and $\alpha_z = 1$, so Z is endogenous. Table 2 provides the average bias and Mean Squared Error (MSE) based on 10,000 Monte Carlo simulations, sample sizes $n = 100$, 300, 500 and 1000, and several values for $(\alpha_{11}, \alpha_{31}, \alpha_{12}, \alpha_{32})$. There are three variables in the structural equation and three in each reduced form equation, and therefore, the TSLS estimator is an IV estimator that treats T and the interaction term $T \cdot Z$ as instruments.

We observe a satisfactory bias performance uniformly over all parameter values. For the first two cases $(\alpha_{11} = 1, \alpha_{31} = 0, \alpha_{12} = 0, \alpha_{32} = 1)$ and $(\alpha_{11} = 0, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 0)$ the sample variance of the estimators is already small for small sample sizes as 100, and it decreases to zero with the sample size, in accordance with the consistency of the estimator. For the parameter values $(\alpha_{11} = 1.25, \alpha_{31} = 1, \alpha_{12} = 1, \alpha_{32} = 1.25)$ identification is much weaker, and larger sample sizes are required for a good performance, as expected. Overall, these results provide supporting evidence of the robustness of our identification strategy to the endogeneity of Z.

Table 2: **IV Case**

α_{11}	α_{31}	α_{12}	α_{32}	n	$Bias(\beta_1)$	$MSE(\beta_1)$	$Bias(\beta_2)$	$MSE(\beta_2)$
1	0	0	1	100	-0.0041	0.0252	0.0013	0.0123
				300	0.0015	0.0070	0.0000	0.0034
				500	-0.0014	0.0041	0.0000	0.0020
				1000	0.0001	0.0020	0.0003	0.0010
0	1	1	0	100	-0.0002	0.0121	0.0009	0.0243
				300	-0.0004	0.0035	0.0002	0.0070
				500	0.0002	0.0020	0.0000	0.0041
				1000	-0.0001	0.0010	0.0004	0.0020
1.25	1	1	1.25	100	0.1247	259.7900	-0.0897	193.4900
				300	0.0528	11.1740	-0.0434	7.2709
				500	-0.0065	0.2693	0.0053	0.2085
				1000	0.0007	0.0157	-0.0003	0.0136

10000 Monte Carlo Simulations.

In the second set of simulations we show how the estimator performs when the first stages are not linear. Consider now the DGP:

$$\begin{aligned} Y &= \alpha + \beta_1 D + \beta_2 D1(D > 0) + \gamma_z Z + u, \\ Z &= \alpha_z u + \varepsilon_z, \\ D &= \alpha_d Z + \gamma_d T \cdot Z + \varepsilon_d, \end{aligned}$$

where $(u, \varepsilon_z, \varepsilon_d)$ are independent standard normal random variables, independent of T, which is again distributed as Bernoulli with probability $p = 0.5$. This corresponds to a linear model

$$Y = \alpha + \beta_0' X + \varepsilon,$$

where $\beta_0 = (\beta_1, \beta_2)'$, $X = (D, D1(D > 0))'$ and $\varepsilon = \gamma_z Z + u$. Here

$$\mathbb{E}[D|Z, T = 1] - \mathbb{E}[D|Z, T = 0] = \gamma_d Z,$$

so γ_d controls the identification strength. Since there is only one binary IV, T, standard IV methods cannot be applied in this example. Note also that under this DGP the difference of conditional means Δ_X is nonlinear in Z in its second component. However, it can be shown that our estimator based on the linearity assumption is still consistent, as we illustrate with some Monte Carlo experiments. This shows the robustness of our estimator to the failure of the linearity assumption in the conditional mean $\mathbb{E}[X|Z, T] = \alpha_{0X}(T) + \alpha_{1X}'(T)Z$.

Table 3 provides the average bias and MSE based on 10000 Monte Carlo simulations. In all cases $\alpha = 0$, $\gamma_z = 1$, $\beta_1 = 1$, $\beta_2 = 2$, $\alpha_z = 1$, $\alpha_d = 1$. We consider two levels of identification, "moderate" $\gamma_d = 1$ and "high" $\gamma_d = 2$.

Table 3: **IV Case - Misspecified Model**

γ_d	n	$Bias(\beta_1)$	$MSE(\beta_1)$	$Bias(\beta_2)$	$MSE(\beta_2)$
1	100	0.00107	0.48403	-0.01446	2.64655
	300	-0.00067	0.01124	0.00039	0.03006
	500	0.00001	0.00638	0.00042	0.01664
	1000	0.00003	0.00303	-0.00072	0.00802
2	100	-0.00160	0.00924	0.00320	0.02321
	300	-0.00065	0.00252	0.00089	0.00645
	500	0.00000	0.00148	0.00024	0.00385
	1000	0.00031	0.00074	-0.00019	0.00189

10000 Monte Carlo Simulations.

The reported results show that the estimator is still consistent even though the conditional means are not linear. Estimates of β_2 require larger sample sizes than those of β_1 to achieve the same level of precision and bias performance. There is an efficiency loss in estimating β_2 relative to β_1, probably due to the nonlinearities in $\mathbb{E}[D1(D > 0)|Z, T = j]$ for $j = 0, 1$. As expected, the results improve with the identification strength. In sum, these simulations provide finite sample evidence of a satisfactory performance of the TSLS estimator and its robustness to the endogeneity of the covariates and the nonlinearity of the first stages.

5.2 The RDD case

We consider now a third DGP to evaluate the new RDD estimator. The DGP is given by:

$$\begin{aligned} Y &= \alpha + \beta X + \gamma_z Z + u, \\ T &= 1(W \geq 0), \\ X &= \alpha_0 + \alpha_1 T + \alpha_2 Z + \alpha_3 T \cdot Z + \varepsilon_X, \end{aligned}$$

where (u, ε_X, W, Z) are independent and identically distributed as standard normals. Note that with this design

$$\lim_{w \downarrow 0} \mathbb{E}[X|W = w, Z] = \alpha_0 + \alpha_1 + (\alpha_2 + \alpha_3) Z$$

and

$$\lim_{w \uparrow 0} \mathbb{E}[X|W = w, Z] = \alpha_0 + \alpha_2 Z.$$

Therefore, $\delta_{Xi} = \alpha_1 + \alpha_3 Z_i$. On the other hand, since $\mathbb{E}[Z|W] = 0$,

$$\lim_{w \downarrow 0} \mathbb{E}[X|W = w] = \alpha_0 + \alpha_1$$

and

$$\lim_{w \uparrow 0} \mathbb{E}[X|W = w] = \alpha_0.$$

Therefore, in this model α_1 controls the level of identification of the classical RDD, whereas α_1 and α_3 control that for the new RDD estimator. We compare the classical RDD estimator ($\widehat{\beta}_{RDD}$) and the new RDD estimator ($\tilde{\beta}$). We consider the parameter values $(\alpha, \beta, \gamma_z) = (0, 1, 0)$, $(\alpha_0, \alpha_2, \alpha_3) = (0, 1, 1)$ and several values of α_1. We implement both estimators as in Section 4.2 with a uniform kernel (17) and a bandwidth $h_n = 2n^{-1/4}$. In particular, the classical RDD estimator $\widehat{\beta}_{RDD}$ is obtained as a TSLS

Table 4: **Average Bias and MSE: RDD**

α_1	n	*Bias*		*MSE*	
		$\widehat{\beta}_{RDD}$	$\tilde{\beta}$	$\widehat{\beta}_{RDD}$	$\tilde{\beta}$
0	100	-0.0048	0.0016	0.8018	0.0650
	300	0.0018	-0.0004	0.8084	0.0275
	500	0.0024	-0.0014	0.8077	0.0182
	1000	-0.0014	0.0016	0.6842	0.0107
1	100	0.0002	0.0006	0.0784	0.0348
	300	-0.0008	-0.0006	0.0292	0.0139
	500	-0.0018	-0.0016	0.0192	0.0094
	1000	0.0016	0.0015	0.0104	0.0053
2	100	0.0005	0.0007	0.0181	0.0141
	300	-0.0004	-0.0004	0.0071	0.0056
	500	-0.0008	-0.0010	0.0047	0.0038
	1000	0.0007	0.0009	0.0026	0.0021

10000 Monte Carlo Simulations.

estimator in the restricted sample $\{(Y_i, X_i, Z_i, W_i) : -h_n \leq W_i \leq h_n\}_{i=1}^n$, where the structural equation is given by

$$Y_i = \alpha + \beta X + \gamma Z + u,$$

and the first stage is given by

$$\mathbb{E}[X|Z, W] = \alpha_{0X} + \alpha_{1X} T + \alpha_{2X} Z + \alpha_{3X} W + \alpha_{4X} T \cdot W,$$

whereas $\tilde{\beta}$ uses the first stage

$$\mathbb{E}[X|Z,W] = \alpha_{0X} + \alpha_{1X}'V + \alpha_{2X}Z,$$

where V is defined in (18).

Table 4 reports the average bias and MSE based on 10000 Monte Carlo simulations. For $\alpha_1 = 0$, the classical RDD is not able to identify the parameter β, and, as expected, has a MSE that does not converge to zero for large n. In contrast, our estimator $\tilde{\beta}$ presents a satisfactory performance, with small MSEs that decrease with the sample size. The classic RDD requires a value of $\alpha_1 = 2$ to achieve comparable results to those of the new RDD. Unreported results with other bandwidths and with other parameter values for $(\alpha_0, \alpha_2, \alpha_3)$ show a similar behavior. This Monte Carlo exercise illustrates the potential benefits of our approach even in situations where the classical RDD is applicable.

6 An Application to the Estimation of the Effects of Air Pollution on House Prices

We apply our identification strategy to the problem of estimating the effects of pollution on house prices, as in Chay and Greenstone (2005). The concern with endogeneity in this problem is warranted, since counties may differ from each other in many ways which may not be accounted by their observable characteristics and amenities. Chay and Greenstone base their identification strategy on an instrumental variable approach, which takes advantage of the quasi-experiment generated by the Clean Air Act around the time it was first implemented. Because the non-attainment threshold is sharp, it allows Chay and Greenstone to also pursue identification in an RDD framework. Thus, this application can illustrate our approach both in the binary IV and the RDD cases.

As explained in Remark 3.2, it is not necessary that all the covariates be separable from X in the model. Let Y denote the change between 1970 and 1980 in the logs of the county's median property values, X is the change between 1970 and 1980 in the geometric mean total suspended particulates (TSP) across all monitors in the county, T is the county's attainment status in 1975 according to the Clean Air Act, and Z_c and Z denote vectors of further variables which are used by Chay and Greenstone as controls in their model specification (2) (Chay and Greenstone (2005), p. 411). We can estimate a model

$$Y = g(X, Z_c) + h(Z) + \varepsilon,$$

with the exclusion restriction $\mathbb{E}[\varepsilon|T, Z_c, Z] = \mathbb{E}[\varepsilon|Z]$, thus allowing a set of covariates Z_c to be non-separable from X as long as they are exogenous. We can generalize Chay and Greenstone (2005)'s approach in two directions. First, even though the instrument T is binary we are able to identify g when it is more general than simply linear. Second, since Z in our approach may be endogenous, the covariates which we choose to use as Z need no longer be assumed exogenous.

In choosing which covariates are part of Z, we must be concerned with their separability from X. The issue is that counties' preference for pollution may differ as a function of some of the covariates. Because of this, it is unadvisable to use covariates such as, for example, the percent change in income,

education levels, racial composition and unemployment in the county population, as these could be reasonably assumed to influence the population's taste in pollution. The covariates which we believe are most likely to be separable are the county's changes between 1970 and 1980 in the percent spending in highways, health and education. We considered estimators of our method with each of those variables separately and also together, as can be seen in Table 5.

The estimation is done using the same data set as in Chay and Greenstone (2005) as well as the exact same covariate specification. The first column in Table 4 shows the results of a standard IV estimation of the effects of pollution change, which is what is done in Chay and Greenstone (2005). The replicated results are, not surprisingly, identical to that paper. Columns A to D show the results of our estimation approach using different variables as the separable classification covariate Z. Row I uses a specification without exogenous control variables. Although in specification I our results are of similar

Table 5: **Estimation Results - Linear Case, Binary IV**

	IV	A	B	C	D
I	-.347	-.340	-.327	-.317	-.327
	(.140)	(.138)	(.136)	(.134)	(.135)
II	-.203	-.208	-.202	-.203	-.208
	(.093)	(.094)	(.093)	(.093)	(.093)

Table 5: Columns A to D use our approach with Z equal to the change from 1970 to 1980 in the % spending on highways (A), health (B), and education (C). In column (D), Z is the vector of all three variables. Specification I has no exogenous covariates Z_c. Specification II uses as exogenous covariates Z_c the exact same specification as in Chay and Greenstone (2005) excluding the covariates which we are using as Z.

magnitude to the standard IV, they are slightly smaller and vary depending on the covariate. They are particularly smaller when all three covariates are used. We believe that this happens because when covariates Z_c are not used, the classic IV operates under the assumption that the IV is unconditionally valid, while our estimator operates under the assumption that the IV is valid only conditional on the separable covariate ($\mathbb{E}[\varepsilon|T,Z] = \mathbb{E}[\varepsilon|Z]$). That said, Chay and Greenstone never suppose that their IV is valid, but only that it is valid conditional on controls. Row II shows the results conditional on controls. There the identifying assumption of the classic IV approach is that $\mathbb{E}[\varepsilon|T,Z_c,Z] = 0$, while our identifying assumption is that $\mathbb{E}[\varepsilon|T,Z_c,Z] = \mathbb{E}[\varepsilon|Z]$. Nevertheless, our results generally confirm the estimates found by Chay and Greenstone.

As a robustness check of the separability of the variables we chose as Z, we ran the same regressions using each of the other covariates in the model as Z instead. The results are extremely similar. The 5 covariates which yielded the most different results are the number of houses built between 1970 and 1980, the rate of vacancies in 1980, change in income per-capita, the change in government revenue per-capita and the change in the fraction of the population with at least a college degree. However,

we should point out that even for these covariates, the differences between our estimates and -.203 was still less than .1.

The advantages of our method are better showcased in the nonlinear case, where the classical instrumental variables methods cannot identify marginal effects with a single instrumental variable. In order to compare our results to Chay and Greenstone's we maintain their specification in all aspects, but allow X to have richer marginal effects on Y. The model is thus $Y = g(X) + Z_c'\gamma_c + Z'\gamma + \varepsilon$, assuming that $\mathbb{E}[\varepsilon|T, Z_c, Z] = \mathbb{E}[\varepsilon|Z]$, where g is a connected piecewise linear function and Z_c is the entire specification of controls in Chay and Greenstone (2005) except for the variables in Z. Figure 1 has three linear pieces, which connect at the terciles of the distribution of X. Hence, we can write $g(x) = a_1\psi_1^3(x) + a_2\psi_2^3(x) +$

Figure 1: 3 pieces

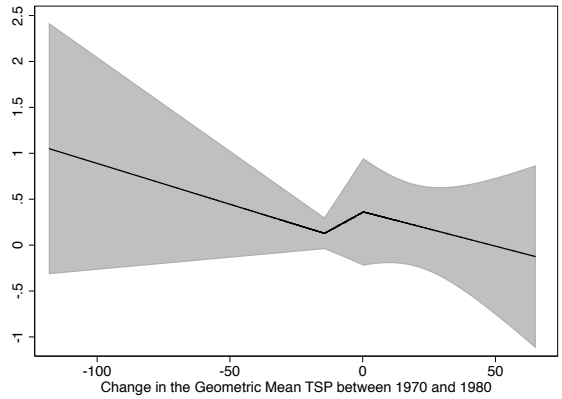


Figure 2: 4 pieces

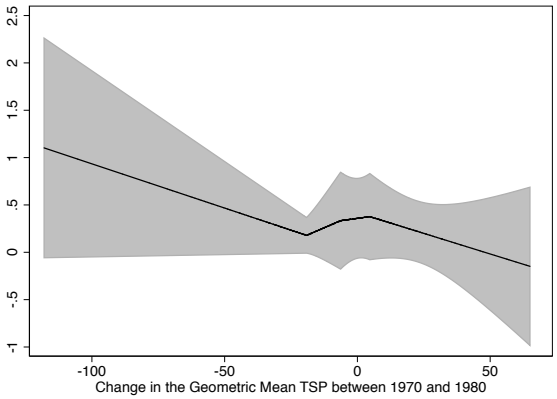


Figure 3: 5 pieces

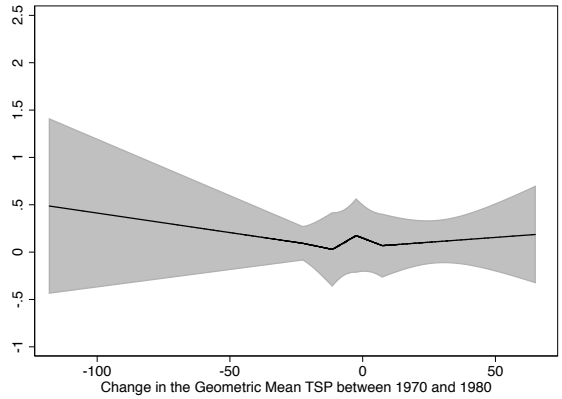
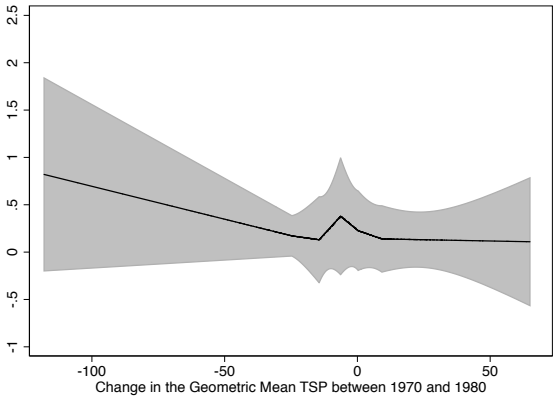


Figure 4: 6 pieces



Figures 1 to 4: Nonlinear case – IV approach. Curves are the results of our approach with exogenous covariates, and Z as in column D in Table 5. The domain of each piece is determined by the quantiles. For instance, Figure 2's pieces connect at the 25th, 50t and 75th quantiles of the change in the geometric Mean TSP between 1970 and 1980.

$a_3\psi_3^3(x)$, where the $\psi_j^3(x)$ are the elements of a B-spline basis of degree 1 and smoothness 0 with knots at the terciles just described. In practice this is the same as if we had three endogenous variables $\psi_1^3(X)$, $\psi_2^3(X)$ and $\psi_3^3(X)$. For Z we used the three variables in column D in Table 5 (call them $High$, $Health$ and $Educ$) expanded into the elements of the B-spline basis. So, with some abuse of notation $Z =$

$(\psi_1^3(High), \psi_2^3(High), \psi_3^3(High), \psi_1^3(Health), \psi_2^3(Health), \psi_3^3(Health), \psi_1^3(Educ), \psi_2^3(Educ), \psi_3^3(Educ))'$. Figures 2 to 4 are obtained analogously.

In Figure 1 the standard errors are calculated as the standard error of the predicted $\hat{g}(x) = \hat{a}_1\psi_1^3(x) + \hat{a}_2\psi_2^3(x) + \hat{a}_3\psi_3^3(x)$, so $SE(\widehat{g}(x)) = (\psi_1^3(x), \psi_2^3(x), \psi_3^3(x))'\Omega(\psi_1^3(x), \psi_2^3(x), \psi_3^3(x))$, where Ω is the estimated covariance matrix of $(\hat{a}_1, \hat{a}_2, \hat{a}_3)'$. In a pre-packaged software the standard errors can be programmed directly as the standard errors of the predicted g. Standard errors in Figures 2 to 4 are obtained analogously.

Our results qualitatively confirm the findings of the linear case, but it is important to notice that the effect may be even more negative than predicted in the linear case in the majority of the domain. Also, interestingly, for an important part of the domain the effect seems to go in the opposite direction. In fact, for small reductions in pollution, all regressions show derivatives which are not only positive, but rather high.

Figures 1 to 4 are representative of the patterns we found when we tried other strategies. For example, we also used as Z each of the elements used in columns A to C in Table 5 separately (i.e. for g with three pieces we did $Z = \left(\psi_1^3(High), \psi_2^3(High), \psi_3^3(High)\right)'$, and we also expanded the elements of Z in other ways, for example into two piece B-splines ($Z = (\psi_1^2(High), \psi_2^2(High), \psi_1^2(Health), \psi_2^2(Health), \psi_1^2(Educ), \psi_2^2(Educ))'$), four piece B-splines, etc., all with very similar results. The standard errors get substantially larger as we increase the number of pieces in g, but they are not affected (and seem in fact to decrease) as we increase the number of elements in Z.

For the RDD approach, Chay and Greenstone (2005) report the results with bandwidth $h_n = 25$ and local quadratic estimators in the paper, which can be seen in the fourth row of results in Table 6 for direct comparisons. The results show some variability and dependence on the bandwidth size, but qualitatively they confirm the findings of the standard IV case. For the smallest bandwidth $h_n = 15$ our results are considerably smaller than the standard RDD, and they show large variations depending on the chosen Z. However, for the larger bandwidths they become more stable and similar to the standard RDD, although still smaller (and generally closer to the numbers found in the IV case).

We can also apply our method for the RDD to identify the response curve in the nonlinear case, which cannot be achieved with the standard RDD identification method. Figures 5 to 8 below are very similar to the curves estimated in the IV case, and they still show the curious pattern of positive derivatives for part of the support. However, the slopes are substantially smaller in the entire domain. The standard errors are also smaller, which is in part due to the fact that there are no covariates Z_c.

These curves are representative of the patterns we find when we try other strategies. For example, Figures 5 to 8 use the local quadratic instead of the local linear estimator because we wanted to compare our results to Chay and Greenstone (2005)'s which use the local quadratic estimator in their approach. However, the results with the local linear are nearly identical. We also experimented using different combinations of variables to form Z, just as described in the standard IV case. The results are again very similar, with one interesting difference: when we used the % spending on education (column C) the curve slopes downward for the highest quantiles.

Table 6: **Estimation Results - Linear Case, RDD**

h	p	RDD	A	B	C	D
15	1	-.305	-.155	-.145	-.173	-.074
		(.246)	(.206)	(.171)	(.175)	(.148)
	2	-.292	-.019	-.120	-.137	.034
		(.240)	(.179)	(.157)	(.163)	(.130)
25	1	-.249	-.233	-.192	-.234	-.182
		(.142)	(.137)	(.122)	(.131)	(.117)
	2	-.275	-.169	-.212	-.253	-.112
		(.141)	(.119)	(.120)	(.132)	(.106)
35	1	-.374	-.305	-.307	-.332	-.227
		(.156)	(.138)	(.132)	(.145)	(.118)
	2	-.330	-.208	-.267	-.289	-.151
		(.142)	(.114)	(.121)	(.131)	(.101)
45	1	-.419	-.405	-.364	-.379	-.311
		(.139)	(.134)	(.124)	(.131)	(.115)
	2	-.391	-.373	-.342	-.355	-.310
		(.136)	(.128)	(.119)	(.126)	(.115)

Table 6: Columns A to D use our approach with Z equal to the change from 1970 to 1980 in the % spending on highways (A), health (B), and education (C). In column (D), Z is the vector of all three variables. h is the bandwidth and p is the degree of the local polynomial. We used local linear and local quadratic estimators because Chay and Greenstone (2005) reported estimates use the local quadratic.

Our application on the effect of air pollution on house prices confirms the findings of Chay and Greenstone (2005) when a constant marginal effect model is considered, but also uncovers substantial heterogeneity in the effect of pollution on house prices when richer marginal effects are entertained. The impact of air quality on house prices is much larger for counties that significantly change their behaviour as a result of the Clean Air Act than for other counties that experience a minor decrease or an increase in pollution during the 1970-1980 period. Thus, our results are consistent with a nonparametric local average treatment effect interpretation where for the population of compliers the marginal effect is larger than the overall average marginal effect (averaged over the whole population) documented in Chay and Greenstone (2005).

Figure 5: 3 pieces

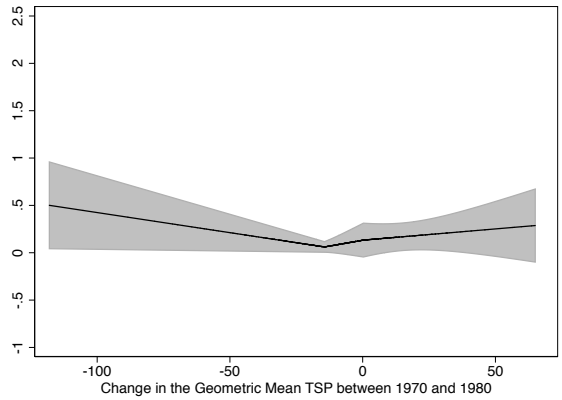


Figure 6: 4 pieces

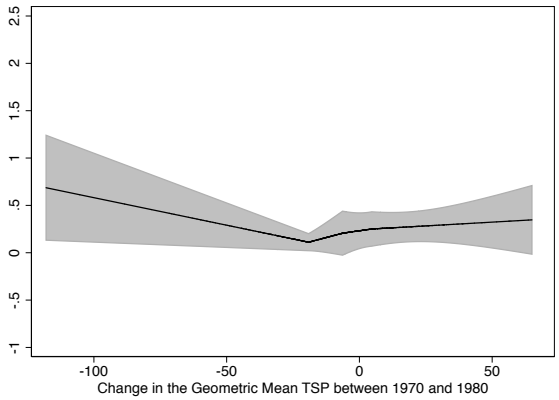


Figure 7: 5 pieces

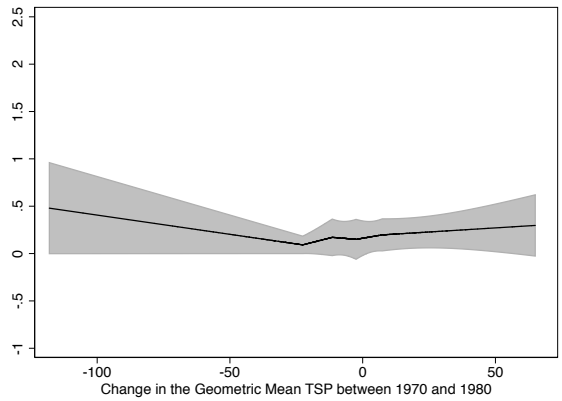
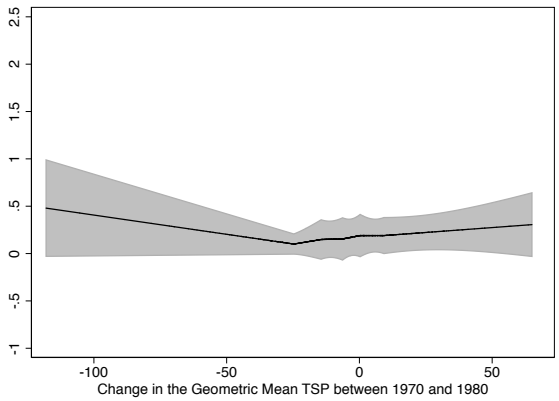


Figure 8: 6 pieces



Figures 5 to 8: Nonlinear case – RDD approach. Curves are the results of our approach using Z as in column D in Table 5 and the local quadratic (bandwidth is $h = 25$). The domain of each piece is determined by the quantiles. For instance, Figure 6's pieces connect at the 25th, 50t and 75th quantiles of the change in the geometric Mean TSP between 1970 and 1980.

7 Conclusions

In this paper we have proposed a strategy to identify marginal effects of "complex" variables using a binary IV in the presence of other possible endogenous covariates. The strategy hinges on the heterogeneity of the "first stages" and separability of some covariates, and it can be extended to the RDD. It can be applied to parametric, semiparametric and nonparametric settings. In models that are linear in parameters (which also include nonparamatric models estimated by sieves), the identification strategy can be implemented with a simple TSLS estimator that treats the covariates as if they were exogenous, and runs a first stage with interactions between the Binary IV and the covariates. Thus, our identification strategy can be implemented with off-the-shelf econometric software. Monte Carlo simulations show that this TSLS estimator performs well in practice, and it is robust to endogeneity of covariates and misspecification of the first stage linear conditional expectation.

In the RDD case, we propose a semiparametric estimator based on a linear specification of the

structural equation and a varying coefficients specification of the first stages that is linear in covariates but nonparametric in the running variable. We have provided an asymptotic analysis for this semiparametric estimator which parallels that of the classical RDD estimator for binary endogenous variables in Hahn, Todd and van der Klaauw (1990). The estimator can be also implemented as a simple TSLS.

There are several extensions of our methods that deserve further investigation. First, as shown in Remark 3.3 our identification strategy can be extended to some non-separable models in unobservables. It is of interest to develop estimation strategies for such non-separable models (e.g. sieve nonparametric estimation) and establish their asymptotic properties. With simple modifications, this strategy could also be extended to the some RDD models with heterogeneous treatment effects. Second, our method may lead to overidentification even in situations where the classic IV and RDD provide just-identification. In such scenario, one can use our methods to test for the overidentifying restrictions, which is not possible with classical IV or RDD.

A Appendix

A.1 Theory for Parametric and Semiparametric Estimators

A.1.1 Proof of Theorem 3.2

We prove the consistency of the TSLS estimator. Its asymptotic distribution follows standard arguments, which are therefore omitted. TSLS identifies the coefficients of the regression of Y on a constant, X^* and Z, where

$$X^* = \alpha_{0X} + \alpha_{1X}T + \alpha_{2X}'Z + \alpha_{3X}'ZT.$$

By the Frisch-Waugh-Lowell Theorem, the slope of X^*, say β^*, is given by

$$\beta^* = \left(\mathbb{E}[\Pi_z\Pi_z']\right)^{-1}\mathbb{E}[\Pi_z Y],$$

where $\Pi_z = X^* - \mathbb{E}[X^*|Z] = \mathbb{E}[X|T,Z] - \mathbb{E}[X|Z]$. We prove that is legitimate to take the inverse of $\mathbb{E}[\Pi_z\Pi_z']$ under our conditions, by showing that $\mathbb{E}[\Pi_z\Pi_z']$ is invertible if and only if $\mathbb{E}[\Delta_X\Delta_X']$ is invertible. To see that, suppose $\mathbb{E}[\Delta_X\Delta_X']$ is singular. Then, there exists a $\lambda \neq 0$ such that, a.s.

$$\mathbb{E}[\lambda'X|T=1,Z] = \mathbb{E}[\lambda'X|T=0,Z].$$

Then,

$$\mathbb{E}[\lambda'X|T,Z] = \mathbb{E}[\lambda'X|Z],$$

and therefore, $\lambda'\Pi_z = 0$ a.s. (i.e. $\mathbb{E}[\Pi_z\Pi_z']$ is singular). The reciprocal follows the same arguments.

Then, substituting (12) into β^*, yields that $\beta^* = \beta$, thereby proving the consistency of the TSLS for β.

A.1.2 Theory for Semiparametric RDD

In this section we establish the asymptotic normality for a generalized version of the TSLS $\tilde{\beta}$ for RDD where a general kernel, not necessarily the uniform kernel (17), is used. For simplicity, we only consider

the case where X and Z are univariate. Without loss of generality assume hereinafter that $\bar{w} = 0$. We introduce some notation. Let $\varepsilon_{V_i} = V_i - \mathbb{E}[V_i|W_i, Z_i]$ denote the regression errors for $V = Y$ and $V = X$. We assume that

$$\mathbb{E}[V|Z, W] := \alpha_{0V}(W) + \alpha'_{1V}(W)Z \tag{19}$$

for both $V = Y$ and $V = X$. To simplify notation denote

$$\lim_{w\downarrow 0} \alpha_{0V}(w) = \alpha^+_{0V} \qquad \lim_{w\downarrow 0} \alpha_{1V}(w) = \alpha^+_{1V}$$

and

$$\lim_{w\uparrow 0} \alpha_{0V}(w) = \alpha^-_{0V} \qquad \lim_{w\uparrow 0} \alpha_{1V}(w) = \alpha^-_{1V}.$$

We can then estimate these quantities using local linear regression. Precisely, for $V = X$ or $V = Y$, let $\alpha^+_V = (\alpha^+_{0V}, \alpha^{+\prime}_{1V}, \alpha^+_{2V}, \alpha^{+\prime}_{3V})'$, and

$$\widehat{\alpha}^+_V = arg\min_{\alpha^+_V} \sum_{i=1}^n \left(V_i - \alpha^+_{0V} - \alpha^{+\prime}_{1V} Z_i - \alpha^+_{2V} W_i - \alpha^{+\prime}_{3V} Z_i W_i\right)^2 k_{h_n}(W_i) 1(W_i \geq 0),$$

where $k_{h_n}(W) = k(W/h_n)$, k is a kernel function and h_n is a bandwidth parameter satisfying some conditions below (Assumption 4). The estimation of $\widehat{\alpha}^-_{0V}$ and $\widehat{\alpha}^-_{1V}$ for $V = Y$ and $V = X$ is analogous, replacing $1(W_i \geq 0)$ in the minimization problem by $1(W_i < 0)$.

We can now apply the arguments of equation (14). Notice that in this setting $\delta_{Xi} = \alpha^+_{0X} - \alpha^-_{0X} + \left(\alpha^+_{1X} - \alpha^-_{1X}\right)' Z_i$ and $\delta_{Yi} = \alpha^+_{0Y} - \alpha^-_{0Y} + \left(\alpha^+_{1Y} - \alpha^-_{1Y}\right)' Z_i$, and thus we can obtain $\widehat{\delta}_{Yi}$ and $\widehat{\delta}_{Xi}$ by substituting the estimated α's from the previous step. Then, $\tilde{\beta}$ is obtained from the OLS regression of the $\widehat{\delta}_{Yi}$ on the $\widehat{\delta}_{Xi}$ (without intercept), and has the closed-form expression

$$\tilde{\beta} = \left(\sum_{i=1}^n \widehat{\delta}_{Xi}\widehat{\delta}'_{Xi}\right)^{-1} \sum_{i=1}^n \widehat{\delta}_{Xi}\widehat{\delta}_{Yi}.$$

We investigate the asymptotic properties of $\tilde{\beta}$ under the following assumptions, which parallel those of Hahn, Todd and van der Klaauw (1990):

Assumption 4 *Suppose that*

1. *For $w > 0$ and $V = Y$ and X, $\alpha_{0V}(w)$ and $\alpha_{1V}(w)$ are twice continuously differentiable. There exists some $M > 0$ such that $\dot{\alpha}^+_{jV}(w) = \lim_{u\downarrow w} \partial\alpha_{jV}(u)/\partial u$ and $\ddot{\alpha}^+_{jV}(w) = \lim_{u\downarrow w} \partial^2\alpha_{jV}(u)/\partial u^2$ are uniformly bounded on $(0, M]$, for $j = 0, 1$. Similarly, $\dot{\alpha}^-_{jV}(w) = \lim_{u\uparrow w} \partial\alpha_{jV}(u)/\partial u$ and $\ddot{\alpha}^-_{jV}(w) = \lim_{u\uparrow w} \partial^2\alpha_{jV}(u)/\partial u^2$ are uniformly bounded on $[-M, 0)$, for $j = 0, 1$.*

2. *Equation (19) holds. The limits $(\alpha^+_{jV}, \dot{\alpha}^+_{jV}, \ddot{\alpha}^+_{jV}, \alpha^-_{jV}, \dot{\alpha}^-_{jV}, \ddot{\alpha}^-_{jV})$ are well-defined and finite for $j = 0, 1$ and $V = Y$ and X.*

3. *The density of W, $f(w)$, is continuous and bounded near $w = 0$. It is also bounded away from zero near $w = 0$.*

4. *The kernel k is continuous, symmetric and nonnegative-valued with compact support.*

5. *The functions $\mu_j(w) = \mathbb{E}[Z^j|W=w]$, $q_{Yj}(w) = \mathbb{E}[Z^{2j}\varepsilon_{Y_i}^2|W=w]$, $q_{Xj}(w) = \mathbb{E}[Z^{2j}\varepsilon_{X_i}^2|W=w]$, $r_j(w) = \mathbb{E}[Z^{2j}\varepsilon_{Y_i}\varepsilon_{X_i}|W=w]$, $s_{Yj}(w) = \mathbb{E}[Z^{3j}\varepsilon_{Y_i}^3|W=w]$ and $s_{Xj}(w) = \mathbb{E}[Z^{3j}\varepsilon_{X_i}^3|W=w]$ are uniformly bounded near $w=0$, with well-defined and finite left and right limits to $w=0$, for $j=0,1$ and 2.*

6. *The bandwidth satisfies $nh_n^5 \to 0$.*

Define the vector of estimates $\widehat{\delta}_{h_n} = (\widehat{\delta}_{0Y}, \widehat{\delta}_{1Y}, h_n\widehat{\delta}_{2Y}, h_n\widehat{\delta}_{3Y}, \widehat{\delta}_{0X}, \widehat{\delta}_{1X}, h_n\widehat{\delta}_{2X}, h_n\widehat{\delta}_{3X})'$ and true parameters $\delta_{h_n} = (\delta_{0Y}, \delta_{1Y}, h_n\delta_{2Y}, h_n\delta_{3Y}, \delta_{0X}, \delta_{1X}, h_n\delta_{2X}, h_n\delta_{3X})'$, where $\widehat{\delta}_{jV} = \widehat{\alpha}_{jV}^+ - \widehat{\alpha}_{jV}^-$ and $\delta_{jV} = \alpha_{jV}^+ - \alpha_{jV}^-$ for $j=0,1,2,3$ and $V=Y$ or X. There are two main steps in the proof of asymptotic normality of $\tilde{\beta}$. First, we show that

$$\begin{aligned}\tilde{\beta} - \beta_0 &= \left(E[\delta_{Xi}^2]\right)^{-1} E[\delta_{X_i}\tilde{Z}_i']B_2\left(\widehat{\delta}_{h_n} - \delta_{h_n}\right) + o_P((nh_n)^{-1/2}) \\ &\equiv Q_2\left(\widehat{\delta}_{h_n} - \delta_{h_n}\right) + o_P((nh_n)^{-1/2}),\end{aligned} \tag{20}$$

where $\tilde{Z}_i = (1, Z_i)'$, $Q_2 = \left(E[\delta_{Xi}^2]\right)^{-1} E[\delta_X\tilde{Z}_i']B_2$, and

$$B_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & -\beta_0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -\beta_0 & 0 & 0 \end{bmatrix}.$$

Second, following similar arguments to those of Hahn, Todd and van der Klaauw (1990) we show in Lemma A.8 that

$$\sqrt{nh_n}\left(\widehat{\delta}_{h_n} - \delta_{h_n}\right) \to_d N(0, \Omega), \tag{21}$$

and provide an expression for Ω. The asymptotic normality then follows from (20) and (21).

To prove (20), use $\delta_{Yi} = \beta_0\delta_{Xi}$ and substitute $\widehat{\delta}_{Yi} = \beta_0\widehat{\delta}_{Xi} + \widehat{\delta}_{Yi} - \delta_{Yi} - \beta_0(\widehat{\delta}_{Xi} - \delta_{Xi})$ in $\tilde{\beta}$ to get

$$\tilde{\beta} = \beta_0 + \frac{\sum_{i=1}^n \widehat{\delta}_{Xi}(\widehat{\delta}_{Yi} - \delta_{Yi})}{\sum_{i=1}^n \widehat{\delta}_{Xi}^2} - \beta_0\frac{\sum_{i=1}^n \widehat{\delta}_{Xi}(\widehat{\delta}_{Xi} - \delta_{Xi})}{\sum_{i=1}^n \widehat{\delta}_{Xi}^2}.$$

By Lemmas A.1-A.8 below, we can further write

$$\tilde{\beta} - \beta_0 = \left(E[\delta_{Xi}^2]\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \delta_{Xi}(\widehat{\delta}_{Yi} - \delta_{Yi}) - \frac{\beta_0}{n}\sum_{i=1}^n \delta_{Xi}(\widehat{\delta}_{Xi} - \delta_{Xi})\right) + o_P((nh_n)^{-1/2}).$$

With some simple manipulations this leads to (20).

Henceforth we focus on the analysis for α_V^+, since that of α_V^- is symmetric. Introduce the notation $f(0^+) = \lim_{w\downarrow 0} f(w)$,

$$V_i^+ = V_i - \alpha_V^{+\prime} S_i \qquad \text{and} \qquad k_{ih_n}^+ = k_{h_n}(W_i)1(W_i \geq 0),$$

where $S_i := (1, Z_i, W_i, W_iZ_i)'$ and $S_{ih_n} := (1, Z_i, W_i/h_n, Z_iW_i/h_n)'$. Then, with this notation we can write the OLS problem for α_V^+ as

$$\min_\alpha \sum_{i=1}^n \left(V_i^+ - (\alpha - \alpha_V^+)'S_i\right)^2 k_{ih_n}^+,$$

whose first order condition yields

$$\begin{bmatrix} \widehat{\alpha}_{0V}^{+} - \alpha_{0V}^{+} \\ \widehat{\alpha}_{1V}^{+} - \alpha_{1V}^{+} \\ h_n\left(\widehat{\alpha}_{2V}^{+} - \alpha_{2V}^{+}\right) \\ h_n\left(\widehat{\alpha}_{3V}^{+} - \alpha_{3V}^{+}\right) \end{bmatrix} = \left(\sum_{i=1}^{n} S_{ih_n} S_{ih_n}' k_{ih_n}^{+}\right)^{-1} \left(\sum_{i=1}^{n} S_{ih_n} V_i^{+} k_{ih_n}^{+}\right).$$

Based on this expression we analyze the asymptotic properties of the local linear estimators. All the Lemmas below make use of Assumptions 2, 3 and 4.

Lemma A.1 (Denominator)

$$\frac{1}{nh_n}\sum_{i=1}^{n} S_{ih_n} S_{ih_n}' k_{ih_n}^{+} \to_p f(0^{+})\Gamma_d$$

where

$$\Gamma_d = \begin{bmatrix} \gamma_0 & \gamma_0\mu_1^{+} & \gamma_1 & \gamma_1\mu_1^{+} \\ \gamma_0\mu_1^{+} & \gamma_0\mu_2^{+} & \gamma_1\mu_1^{+} & \gamma_1\mu_2^{+} \\ \gamma_1 & \gamma_1\mu_1^{+} & \gamma_2 & \gamma_2\mu_1^{+} \\ \gamma_1\mu_1^{+} & \gamma_1\mu_2^{+} & \gamma_2\mu_1^{+} & \gamma_2\mu_2^{+} \end{bmatrix},$$

$$\gamma_l = \int_0^{\infty} u^l k(u)du \qquad \textit{and} \qquad \mu_j^{+} = \lim_{w\downarrow 0}\mathbb{E}[Z^j|W=w].$$

Proof. *Let*

$$\theta_{lj} = \frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^{l} Z_i^j k_{ih_n}^{+}, \qquad l,j = 0,1,2.$$

Then, by the change of variables $u = w/h_n$,

$$\begin{aligned} \mathbb{E}[\theta_{lj}] &= h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^{l} Z_i^j k_{ih_n}^{+}\right] \\ &= \int_0^{\infty} u^l k(u)\mu_j(uh_n)f(uh_n)du \\ &= \mu_j^{+} f(0^{+})\gamma_l + o(1), \end{aligned}$$

where $\mu_j(w) = \mathbb{E}[Z^j|W=w]$ *and the convergence follows by the Dominated Convergence theorem. As for the variance*

$$\begin{aligned} Var(\theta_{lj}) &\leq \left(nh_n^2\right)^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^{2l} Z_i^{2j} k_{ih_n}^{+2}\right] \\ &= (nh_n)^{-1}\int_0^{\infty} u^{2l}k^2(u)\mu_{2j}(uh_n)f(uh_n)du \\ &= o(1), \end{aligned}$$

again by the Dominated Convergence theorem. ■

We now consider the asymptotic behaviour of the numerator. Define the function

$$\begin{aligned}\zeta_V(w,z) = \alpha_{0V}(w) + \alpha_{1V}(w)z - \alpha_{0V}^+ + \alpha_{1V}^+ z \\ - \left(\dot{\alpha}_{0V}^+ + \dot{\alpha}_{1V}^+ z\right) w - \frac{1}{2}\left(\ddot{\alpha}_{0V}^+ + \ddot{\alpha}_{1V}^+ z\right) w^2,\end{aligned}$$

where $\dot{\alpha}_{0V}^+ = \lim_{w\downarrow 0} \partial\alpha_{0V}(w)/\partial w$ and $\ddot{\alpha}_{0V}^+ = \lim_{w\downarrow 0} \partial^2\alpha_{0V}(w)/\partial w^2$, and similarly for α_{1V}. Note that $\dot{\alpha}_{0V}^+ = \alpha_{2V}^+$ and $\ddot{\alpha}_{0V}^+ = \alpha_{3V}^+$ and observe that

$$\sup_{0<w<Mh_n} |\zeta_V(w,z)| = o(h_n^2)(1+|Z|).$$

Lemma A.2 (Numerator: Expectation)

$$\mathbb{E}\left[\frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} V_i^+ k_{ih_n}^+\right] \to_p \frac{1}{2} f(0^+) h_n^2 (b_V^+ + o(1)),$$

where

$$b_V^+ = \begin{bmatrix} \gamma_2(\ddot{\alpha}_{0V}^+ + \ddot{\alpha}_{0V}^+\mu_1^+) \\ \gamma_2(\ddot{\alpha}_{0V}^+\mu_1^+ + \ddot{\alpha}_{0V}^+\mu_2^+) \\ \gamma_3(\ddot{\alpha}_{0V}^+ + \ddot{\alpha}_{0V}^+\mu_1^+) \\ \gamma_3(\ddot{\alpha}_{0V}^+\mu_1^+ + \ddot{\alpha}_{0V}^+\mu_2^+) \end{bmatrix}.$$

Proof. *Let*

$$u_{lj} = \frac{1}{nh_n}\sum_{i=1}^n \left(\frac{W_i}{h_n}\right)^l Z_i^j V_i^+ k_{ih_n}^+, \qquad l,j = 0,1.$$

Then, write

$$\begin{aligned}\mathbb{E}[u_{lj}] &= h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j V_i^+ k_{ih_n}^+\right] \\ &= h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j \left(\frac{1}{2}\left(\ddot{\alpha}_{0V}(0^+) + \ddot{\alpha}_{1V}(0^+)Z_i\right) W_i^2 + \zeta_V(W_i,Z_i)\right) k_{ih_n}^+\right] \\ &= h_n^{-1}\frac{1}{2}\ddot{\alpha}_{0V}(0^+)\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j W_i^2 k_{ih_n}^+\right] + h_n^{-1}\frac{1}{2}\ddot{\alpha}_{1V}(0^+)\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^{j+1} W_i^2 k_{ih_n}^+\right] \\ &+ h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j \zeta_V(W_i,Z_i) k_{ih_n}^+\right].\end{aligned}$$

By the change of variables $u = w/h_n$,

$$\begin{aligned}h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j W_i^2 k_{ih_n}^+\right] &= h_n^2 \int_0^\infty u^{l+2} k(u)\mu_j(uh_n) f(uh_n) du \\ &= h_n^2 \mu_j^+ f(0^+)\gamma_{l+2} + o(1),\end{aligned}$$

and similarly

$$h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^{j+1} W_i^2 k_{ih_n}^+\right] = h_n^2 \mu_{j+1}^+ f(0^+)\gamma_{l+2} + o(1).$$

On the other hand, assume without loss of generality that $[-M, M]$ contains the support of k, so that

$$h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j \zeta_V(W_i, Z_i) k_{ih_n}^+\right] = o(h_n^2).$$

■

Lemma A.3 (Numerator: Conditional Expectation)

$$\frac{1}{nh_n}\sum_{i=1}^n \mathbb{E}[S_{ih_n} V_i^+ k_{ih_n}^+ | W_i, Z_i] = \frac{1}{nh_n}\sum_{i=1}^n \mathbb{E}[S_{ih_n} V_i^+ k_{ih_n}^+] + o_p(h_n^2).$$

Proof. *We have*

$$\begin{aligned}
\frac{1}{nh_n}\sum_{i=1}^n \mathbb{E}[S_{ih_n} V_i^+ k_{ih_n}^+ | W_i, Z_i] &= \frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ \left(\frac{1}{2}\left(\ddot{\alpha}_{0V}(0^+) + \ddot{\alpha}_{1V}(0^+) Z_i\right) W_i^2 + \zeta_V(W_i, Z_i)\right) \\
&= \frac{1}{2}\ddot{\alpha}_{0V}(0^+)\frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ W_i^2 + \frac{1}{2}\ddot{\alpha}_{1V}(0^+)\frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ Z_i W_i^2 \\
&\quad + \frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ \zeta_V(W_i, Z_i).
\end{aligned}$$

Observe that

$$\begin{aligned}
Var\left(\frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ W_i^2\right) &= \left(nh_n^2\right)^{-1} Var\left(S_{ih_n} k_{ih_n}^+ W_i^2\right) \\
&\leq C\left(nh_n\right)^{-1} h_n^{-1}\mathbb{E}\left[S_{ih_n} S'_{ih_n} k_{ih_n}^{+2} W_i^4\right] \\
&= O\left(\left(nh_n\right)^{-1} h_n^4\right) \\
&= o(1),
\end{aligned}$$

since for $l, j = 0, 1, 2$

$$\begin{aligned}
h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^l Z_i^j k_{ih_n}^{+2} W_i^4\right] &= h_n^4 \int_0^\infty u^l k^2(u) \mu_j(uh_n) f(uh_n) du \\
&= h_n^4 \mu_j^+ f(0^+) v_l + o(1),
\end{aligned}$$

where

$$v_l = \int_0^\infty u^l k^2(u) du.$$

Similarly,

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ Z_i W_i^2\right) = o(1).$$

and

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^n S_{ih_n} k_{ih_n}^+ \zeta_V(W_i, Z_i)\right) = o(1).$$

■

Note that

$$\frac{1}{nh_n}\sum_{i=1}^{n} S_{ih_n}k^+_{ih_n}\left(V_i^+ - \mathbb{E}[V_i^+|W_i,Z_i]\right) = \frac{1}{nh_n}\sum_{i=1}^{n} S_{ih_n}k^+_{ih_n}\varepsilon_{V_i},$$

where $\varepsilon_{V_i} = V_i - \mathbb{E}[V_i|W_i,Z_i]$ denotes the regression error. Then, we have the following result.

Lemma A.4 (Numerator: Conditional Variance)

$$Var\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{ih_n}k^+_{ih_n}\left(V_i^+ - \mathbb{E}[V_i^+|W_i,Z_i]\right)\right) = \frac{1}{nh_n}f(0^+)\Sigma_{V^+} + o(1),$$

where

$$\Sigma_{V^+} = \begin{bmatrix} v_0 & v_0q_1^+ & v_2 & v_2q_1^+ \\ v_0q_1^+ & v_0q_2^+ & v_2q_1^+ & v_2q_2^+ \\ v_2 & v_2q_1^+ & v_4 & v_4q_1^+ \\ v_2q_1^+ & v_2q_2^+ & v_4q_1^+ & v_4q_2^+ \end{bmatrix},$$

$$v_l = \int_0^\infty u^l k^2(u)du \text{ and } q_j^+ = \lim_{w\downarrow 0}\mathbb{E}[Z^{2j}\varepsilon^2_{V_i}|W=w].$$

Proof. *Consider the generic term, for $l,j=0,1$*

$$\frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^l Z_i^j k^+_{ih_n}\varepsilon_{V_i},$$

and its variance, which equals

$$\begin{aligned}(nh_n)^{-1}h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^{2l} Z_i^{2j}k^{+2}_{ih_n}\varepsilon^2_{V_i}\right] &= (nh_n)^{-1}\int_0^\infty u^{2l}k^2(u)q_j(uh_n)f(uh_n)du \\ &= (nh_n)^{-1}f(0^+)q_j^+v_{2l} + o(1)\end{aligned}$$

where $q_j(w) = \mathbb{E}[Z^{2j}\varepsilon^2_{V_i}|W=w]$. ■

Lemma A.5 (Numerator: Conditional Covariance)

$$Cov\left(\frac{1}{nh_n}\sum_{i=1}^{n} S_{ih_n}k^+_{ih_n}\varepsilon_{Y_i}, \frac{1}{nh_n}\sum_{i=1}^{n} S_{ih_n}k^+_{ih_n}\varepsilon_{X_i}\right) = \frac{1}{nh_n}f(0^+)\Sigma_{YX^+} + o(1),$$

where

$$\Sigma_{YX^+} = \begin{bmatrix} v_0 & v_0r_1^+ & v_2 & v_2r_1^+ \\ v_0r_1^+ & v_0r_2^+ & v_2r_1^+ & v_2r_2^+ \\ v_2 & v_2r_1^+ & v_4 & v_4r_1^+ \\ v_2r_1^+ & v_2r_2^+ & v_4r_1^+ & v_4r_2^+ \end{bmatrix}$$

and

$$r_j^+ = \lim_{w\downarrow 0}\mathbb{E}[Z^{2j}\varepsilon_{Y_i}\varepsilon_{X_i}|W=w].$$

Proof. *The proof is analogous to the previous Lemma, and hence it is omitted.* ■

Lemma A.6 (Numerator: Conditional CLT)

$$(nh_n)^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} S_{ih_n}k^+_{ih_n}\varepsilon_{Y_i} \\ S_{ih_n}k^+_{ih_n}\varepsilon_{X_i}\end{pmatrix} \to_d f(0^+)N\left(0,\begin{bmatrix}\Sigma_{Y^+} & \Sigma_{YX^+}\\ \Sigma_{YX^+} & \Sigma_{X^+}\end{bmatrix}\right).$$

Proof. *Consider a generic term for $l,j=0,1$*

$$\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\left(\frac{W_i}{h_n}\right)^l Z_i^j k^+_{ih_n}\varepsilon_{V_i}.$$

We apply Lyapounov with third absolute moment. By the lemma on the asymptotic variance, we need to establish

$$(nh_n)^{-1/2}\,h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^{3l} Z_i^{3j}k^{+3}_{ih_n}\varepsilon^3_{V_i}\right]=o(1).$$

But note that, defining $s_j(w)=\mathbb{E}[Z^{3j}\varepsilon^3_{V_i}|W=w]$,

$$\begin{aligned}h_n^{-1}\mathbb{E}\left[\left(\frac{W_i}{h_n}\right)^{3l} Z_i^{3j}k^{+3}_{ih_n}\varepsilon^3_{V_i}\right] &= \int_0^\infty u^{3l}k^3(u)s_j(uh_n)f(uh_n)du\\ &= O(1).\end{aligned}$$

■

Lemma A.7 (Numerator: Unconditional CLT)

$$(nh_n)^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} S_{ih_n}k^+_{ih_n}Y_i^+ \\ S_{ih_n}k^+_{ih_n}X_i^+\end{pmatrix} - \frac{(nh_n)^{1/2}h_n^2}{2}f(0^+)\begin{pmatrix} b_Y^+\\ b_X^+\end{pmatrix} \to_d f^{1/2}(0^+)N\left(0,\begin{bmatrix}\Sigma_{Y^+} & \Sigma_{YX^+}\\ \Sigma_{YX^+} & \Sigma_{X^+}\end{bmatrix}\right).$$

Proof. *It follows from previous Lemmas.* ■

Denote

$$\widehat{\alpha}^+_{h_n}-\alpha^+_{h_n}=\begin{bmatrix}\widehat{\alpha}^+_{0Y}-\alpha^+_{0Y}\\ \widehat{\alpha}^+_{1Y}-\alpha^+_{1Y}\\ h_n\left(\widehat{\alpha}^+_{2Y}-\alpha^+_{2Y}\right)\\ h_n\left(\widehat{\alpha}^+_{3Y}-\alpha^+_{3Y}\right)\\ \widehat{\alpha}^+_{0X}-\alpha^+_{0X}\\ \widehat{\alpha}^+_{1X}-\alpha^+_{1X}\\ h_n\left(\widehat{\alpha}^+_{2X}-\alpha^+_{2X}\right)\\ h_n\left(\widehat{\alpha}^+_{3X}-\alpha^+_{3X}\right)\end{bmatrix}\qquad \widehat{\alpha}^-_{h_n}-\alpha^-_{h_n}=\begin{bmatrix}\widehat{\alpha}^-_{0Y}-\alpha^-_{0Y}\\ \widehat{\alpha}^-_{1Y}-\alpha^-_{1Y}\\ h_n\left(\widehat{\alpha}^-_{2Y}-\alpha^-_{2Y}\right)\\ h_n\left(\widehat{\alpha}^-_{3Y}-\alpha^-_{3Y}\right)\\ \widehat{\alpha}^-_{0X}-\alpha^-_{0X}\\ \widehat{\alpha}^-_{1X}-\alpha^-_{1X}\\ h_n\left(\widehat{\alpha}^-_{2X}-\alpha^-_{2X}\right)\\ h_n\left(\widehat{\alpha}^-_{3X}-\alpha^-_{3X}\right)\end{bmatrix}$$

and

$$\Sigma_+=\begin{bmatrix}\Sigma_{Y^+} & \Sigma_{YX^+}\\ \Sigma_{YX^+} & \Sigma_{X^+}\end{bmatrix}\qquad \Sigma_-=\begin{bmatrix}\Sigma_{Y^-} & \Sigma_{YX^-}\\ \Sigma_{YX^-} & \Sigma_{X^-}\end{bmatrix},$$

where the definition of Σ_- is like Σ_+ but with limits to the left of $w=0$.

Lemma A.8 (Joint CLT)

$$\sqrt{nh_n}\left(\widehat{\alpha}^+_{h_n} - \alpha^+_{h_n} - \widehat{\alpha}^-_{h_n} + \alpha^-_{h_n}\right) - \frac{(nh_n)^{1/2}h_n^2}{2}\begin{bmatrix}\Gamma_d^{-1} & 0\\ 0 & \Gamma_d^{-1}\end{bmatrix}\begin{pmatrix} b_Y^+ - b_Y^- \\ b_X^+ - b_X^-\end{pmatrix}$$
$$\to_d f^{-1/2}(0^+)\begin{bmatrix}\Gamma_d^{-1} & 0\\ 0 & \Gamma_d^{-1}\end{bmatrix} N\left(0, \Sigma_+ + \Sigma_-\right).$$

Proof. *It follows from previous Lemmas and the asymptotic independence of* $(nh_n)^{-1/2}\left(\widehat{\alpha}^+_{h_n} - \alpha^+_{h_n}\right)$ *and* $(nh_n)^{-1/2}\left(\widehat{\alpha}^-_{h_n} - \alpha^-_{h_n}\right)$. ■

Define

$$\Omega = (\Sigma_+ + \Sigma_-).$$

Proof of the Theorem 4.1:. From (20),

$$\sqrt{nh_n}\left(\tilde{\beta} - \beta_0\right) = Q_2\sqrt{nh_n}\left(\widehat{\delta}_{h_n} - \delta_{h_n}\right) + o_P(1).$$

Then, apply the previous Lemma and Slutsky's Lemma. ■

A.2 Nonparametric Estimators

A.2.1 Binary IV Case

We first introduce some notation that will be used throughout this Section. Henceforth, A', $rank(A)$, A^-, $tr(A)$ and $|A| := (tr(A'A))^{1/2}$ denote the transpose, rank, Moore-Penrose generalized inverse, trace and the Euclidean norm of a matrix A, respectively. For generic random vectors ζ and ξ, let F_ζ and $F_{\zeta/\xi}$ be the cumulative distribution function (cdf) and conditional cdf of ζ and ζ given ξ, respectively. Denote the corresponding densities with respect to a σ-finite measure $\mu(x)$ by f_ζ and $f_{\zeta/\xi}$. Unless otherwise stated, the underlying measure will be the Lebesgue measure. Let $\mathcal{S}_\zeta$ denote the support of ζ. Let $L_2(\zeta)$ denote the Hilbert space with inner product $\langle h, g\rangle := \int f(x)g(x)dF_\zeta(x)$ and the corresponding norm $\|g\|_2^2 := \langle g, g\rangle$. Henceforth, sometimes we drop the domain of integration for simplicity of notation. For a linear operator $K : L_2(X) \to L_2(Y)$, denote the subspaces $\mathcal{R}(K) := \{f \in L_2(Y) : \exists s \in L_2(X), Ks = f\}$ and $\mathcal{N}(K) := \{f \in L_2(X) : Kf = 0\}$. Let $\mathcal{D}(K)$ denote the domain of definition of K. Let K^* denote the adjoint operator of K. We will use some basic results from operator theory and Hilbert spaces. See Carrasco, Florens and Renault (2006) for an excellent review of these results.

Equation (7) provides an integral equation of the first kind that can be used for estimating g. Similar estimators have been proposed before in Newey and Powell (2003), Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Darolles, Fan, Florens and Renault (2011), Horowitz (2011), Chen and Pouzo (2012) and Santos (2012), to mention just a few. Here, we follow closely Blundell, Chen and Kristensen (2007). Although, strictly speaking, our model is not given by a conditional moment restriction on a unique set of covariates, we can easily adapt the existing results to make them applicable in our setting. For simplicity, we focus here on the univariate Z and X case.

There are many nonparametric methods that can be used to estimate m and A. Here we follow Blundell, Chen and Kristensen (2007) and use a sieve OLS estimator (SLS), see also Ai and Chen (2003) and Newey and Powell (2003). Optimally weighted estimators can be obtained applying ideas in Blundell, Chen and Kristensen (2007). We assume we have a random (i.e. independent and identically distributed, in short iid) sample $\{(Y_i, X_i, Z_i, T_i)\}_{i=1}^n$ of size $n \geq 1$, with the same distribution as the fourth-dimensional vector (Y, X, Z, T). We assume g is in a suitable space of smooth functions. Suppose $\mathcal{S}_X$ is a bounded interval of $\mathbb{R}$, with non-empty interior. For any smooth function $h : \mathcal{S}_X \subset \mathbb{R} \to \mathbb{R}$ and some $r > 0$, let $[r]$ be the largest integer smaller than r, and

$$\|h\|_{\infty,r} := \max_{j \leq \underline{\eta}} \sup_{x \in \mathcal{S}_X} \left|\nabla^j h(x)\right| + \sup_{x \neq x'} \frac{\left|\nabla^{[r]} h(x) - \nabla^{[r]} h(x')\right|}{|x - x'|^{r - [r]}}.$$

Further, let $C_c^r(\mathcal{S}_X)$ be the set of all continuous functions h with $\|h\|_{\infty,r} \leq c$. Since the constant c is irrelevant for our results, we drop the dependence on c and denote $C^r(\mathcal{S}_X)$. We shall assume that $g \in C^r(\mathcal{S}_X)$ for some r and approximate $C^r(\mathcal{S}_X)$ with a sieve space $\mathcal{G}_n$ satisfying some conditions below. Define $k_n = \dim(\mathcal{G}_n)$. Given an integer $s > 0$ define the Sobolev norm $\|h\|_s^2 := \sum_{l=0}^s \left\|h^{(s)}\right\|_2^2$, where $h^{(s)}(x) := \partial^s h(x)/\partial x^s$, with $h^{(0)} \equiv h$.

We approximate $m_t(z) \equiv m(z,t) := \mathbb{E}[Y|Z = z, T = t]$ by the function $\tilde{m}(z,t) := \sum_{j \in \mathcal{J}_n} m_{tj} p_{0j}(z,t)$, where p_{0j} are some known basis functions and $J_n := \#(\mathcal{J}_n) \to \infty$ as $n \to \infty$. We write $\tilde{m}(z,t) = p^{J_n}(z,t)' m^{J_n}(t)$, where $p^{J_n}(z,t) = (p_{01}(z,t), ..., p_{0J_n}(z,t))'$ and $m^{J_n}(t) = (m_{t1}, ..., m_{tJ_n})$. Define $P := (p^{J_n}(Z_1, T_1), ..., p^{J_n}(Z_n, T_n))'$. Then, the SLS is

$$\widehat{m}(z,t) = p^{J_n}(z,t)'(P'P)^- \sum_{i=1}^n p^{J_n}(Z_i, T_i) Y_i.$$

More precisely, we take $p^{J_n}(z,t) = (B^{J_{2n}}(z), t \cdot B^{J_{2n}}(z))$, where $B^{J_{2n}}(z)$ is a $J_{2n} \cdot 1$ vector of univariate B-splines or polynomial splines and $J_n = 2J_{2n}$. We define $\widehat{m}(z) := \widehat{m}(z,1) - \widehat{m}(z,0)$.

Similarly, for a fixed g, we consider the sieve estimator of Ag as $\widehat{A}g = \widehat{A}_1 g - \widehat{A}_0 g$, where

$$\widehat{A}_t g = p^{J_n}(z,t)'(P'P)^- \sum_{i=1}^n p^{J_n}(Z_i, T_i) g(X_i).$$

Finally, the SLS for g is given by the solution of

$$\widehat{g}_n = \arg\min_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^n \left(\widehat{m}(Z_i) - \widehat{A}g(Z_i)\right)^2.$$

We assume the sieve space $\mathcal{G}_n$ is of the form

$$\begin{aligned} \mathcal{G}_n = \{g_n : \mathcal{S}_X \to \mathbb{R},\ & \sup_x |g_n(x)| < c, \sup_x \left|\nabla^{[r]} g_n(x)\right| < c \\ & g_n(x) = \psi^{k_n}(x)'\Pi,\ g_n(x_0) = 0\}, \end{aligned}$$

where $\psi^{k_n}(\cdot)$ is a $k_n \times 1$ vector of known basis that are at least $\gamma = ([r] + 1)$ times differentiable and Π is a $k_n \times 1$ vector of coefficients to be estimated. In the application we use B-splines for ψ^{k_n}.

Blundell, Chen and Kristensen (2007) discussed practical ways to incorporate the constraints into the computation of $\widehat{g}_n$. For large samples the unconstrained estimator performs well. Note that $g_n(x_0) = 0$ is a normalization restriction (location), where x_0 is an arbitrary point in $\mathcal{S}_X$.

The following sieve measure of ill-posedness plays a crucial role in the asymptotic theory of sieve estimators, see Blundell, Chen and Kristensen (2007),

$$\tau_n := \sup_{g \in \mathcal{G}_n} \frac{\|g\|}{\left\|(A^*A)^{1/2} g\right\|}.$$

Consider the following assumptions, which are the same as in Blundell, Chen and Kristensen (2007), and therefore are discussed extensively there.

Assumption 5 *Suppose that*

1. *The data* $\{(Y_i, X_i, Z_i, T_i)\}_{i=1}^n$ *are iid and Assumption 2 holds.*
2. *(i)* $g \in C^r(\mathcal{S}_X)$ *for* $r > 1/2$ *and* $g(x_0) = 0$*; (ii)* $\mathbb{E}[|X|^{2a}] < \infty$ *for some* $a > r$*.*
3. *For* $t = 1, 2$*,* $m_t \in C^{r_m}(\mathcal{S}_Z)$ *with* $r_m > 1/2$ *and* $\mathbb{E}[g_n(X)|Z = \cdot, T = t] \in C^{r_m}(\mathcal{S}_Z)$ *for any* $g_n \in \mathcal{G}_n$*.*
4. *(i) The smallest and the largest eigenvalues of* $\mathbb{E}[B^{J_{2n}}(Z) \cdot B^{J_{2n}}(Z)']$ *are bounded and bounded away from zero for each* J_{2n}*; (ii)* $B^{J_{2n}}(Z)$ *is a B-spline basis of order* $\gamma > r_m > 1/2$*; (iii) the density of* Z *is continuous, bounded, and bounded away from zero over its support* $\mathcal{S}_Z$*, which is a compact interval with non-empty interior.*
5. *(i)* $k_n \to \infty$*,* $J_{2n}/n \to 0$*; (ii)* $\lim_{n\to\infty} (J_{2n}/k_n) = c_0 > 1$ *and* $\lim_{n\to\infty} \left(k_n^2/n\right) = 0$*.*
6. *There is* $g_n \in \mathcal{G}_n$ *such that* $\tau_n^2 \|A(g - g_n)\|^2 \leq C \|g - g_n\|^2$*.*

The following Theorem establishes rates for $\|\widehat{g}_n - g\|$. Its proof is the same as that of Theorem 2 in Blundell, Chen and Kristensen (2007), hence it is omitted.

Theorem A.9 *Let Assumption 5 hold. Then,*

$$\|\widehat{g}_n - g\| = O_P\left(k_n^{-r} + \tau_n \cdot \sqrt{\frac{k_n}{n}}\right).$$

A.2.2 RDD Case

The estimator in the RDD case is also a SLS for g given by the solution of

$$\widehat{g}_n = \arg\min_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^n \left(\widehat{m}(Z_i) - \widehat{A} g(Z_i)\right)^2,$$

where now $\widehat{m}$ and $\widehat{A}$ are estimated by local linear kernel estimators. For the sake of space, we only consider the univariate case for X and Z, and provide only a sketch of the arguments to avoid repetition

with the existing literature. Similarly as before, we write $\widehat{m}(z) := \widehat{m}_+(z) - \widehat{m}_-(z)$ and $\widehat{A}g = \widehat{A}_+g - \widehat{A}_-g$, where $\widehat{m}_+(z) = \widehat{a}$ in the solution to the least squares problem

$$(\widehat{a}, \widehat{b}_0, \widehat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} \left(Y_i - a_0 - b_0 W - b_1 (Z_i - z)\right)^2 k_{h_n}(W_i) k_{h_n}(Z_i - z) 1(W \geq 0),$$

where $k_{h_n}(z_l) = h_n^{-1} k(z_l / h_n)$, $k(\cdot)$ is a kernel function, and h_n denotes a bandwidth parameter satisfying regularity conditions described below. Similarly, $\widehat{m}_-(z) = \widehat{a}$ in

$$(\widehat{a}, \widehat{b}_0, \widehat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} \left(Y_i - a_0 - b_0 W - b_1 (Z_i - z)\right)^2 k_{h_n}(W_i) k_{h_n}(Z_i - z) 1(W < 0),$$

$\widehat{A}_+ g = \widehat{a}$ in

$$(\widehat{a}, \widehat{b}_0, \widehat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} \left(g(X_i) - a_0 - b_0 W - b_1 (Z_i - z)\right)^2 k_{h_n}(W_i) k_{h_n}(Z_i - z) 1(W \geq 0),$$

and $\widehat{A}_- g = \widehat{a}$ in

$$(\widehat{a}, \widehat{b}_0, \widehat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} \left(g(X_i) - a_0 - b_0 W - b_1 (Z_i - z)\right)^2 k_{h_n}(W_i) k_{h_n}(Z_i - z) 1(W < 0).$$

We can follow Theorem 2 in Blundell, Chen and Kristensen (2007) to obtain the rates for $\widehat{g}_n$,

$$\|\widehat{g}_n - g\| = O_P\left(k_n^{-r} + \tau_n \cdot \sqrt{\frac{k_n}{n}}\right),$$

provided we show that

$$\|\widehat{m} - m\| = O_P\left(\sqrt{\frac{k_n}{n}}\right)$$

and

$$\sup_{g \in \mathcal{G}_n} \left\| \left(\widehat{A} - A\right) g(Z_i) \right\| = O_P\left(\sqrt{\frac{k_n}{n}}\right);$$

see Claim 2 in p. 1658 of Blundell, Chen and Kristensen (2007). But Theorem 6 in Masry (1996) shows that

$$\|\widehat{m} - m\| = O_P\left(d_n\right),$$

where $d_n = \left(\log n / n h_n^2\right)^{1/2} + h_n^4$. Combining standard empirical processes arguments with the results of Masry (1996), we similarly obtain

$$\sup_{g \in \mathcal{G}_n} \left\| \left(\widehat{A} - A\right) g(Z_i) \right\| = O_P\left(d_n\right).$$

Therefore, we require rates on the bandwidth h_n so that

$$d_n = O_P\left(\sqrt{\frac{k_n}{n}}\right).$$

This provides some flexibility in how to choose the bandwidth h_n. Higher order polynomial estimation improves the bias of the first step estimates, and leads to wider set of possible bandwidths. The reader is referred to Blundell, Chen and Kristensen (2007) and Masry (1996) for discussion on rates permitted for k_n and h_n to obtain the desired rates for $\widehat{g}_n - g$ under different scenarios on the rate for the measure of ill-posedness.

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Almond, D., and J. Currie (2011): "Killing Me Softly: The Fetal Origins Hypothesis," *The Journal of Economic Perspectives*, 25(3), 153–172.

Andrews, D. (2011): "Examples of L2-Complete and Boundedly-Complete Distributions," Cowles Foundation for Research in Economics.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?, " *The Quarterly Journal of Economics*, 106(4), 979-1014.

Angrist, J. D., and W. N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88, 450-477.

Angrist, J., Graddy, K. and G. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499-527.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV Estimation of Shape-invariant Engel Curves, " *Econometrica*, 75, 1613–1670.

Brollo, F., T. Nannicini, R. Perotti and G. Tabellini (2013): "The Political Resource Curse" *American Economic Review*, 103(5): 1759-96.

Calonico, S., Cattaneo, M.D. and Titiunik, R. (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs", *Econometrica*, 82, 2295-2326.

Card, D. (1995): "The Wage Curve: A Review," *Journal of Economic Literature*, vol. 33(2), 285-299.

Carpenter, C. and Dobkin, C. (2009): "The Effect of Alcohol Access on Consumption and Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age", *American Economic Journal: Applied Economics*, Vol. 1, Issue 1, pp. 164-82.

Carrasco, M., J. P. Florens, and E. Renault (2006): "Linear Inverse Problem in Strucutral Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, vol. 6, ed. by J. J. Heckman and E. E. Leamer. Amsterdam: North-Holland, 5633–5751.

Chay, K., and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113(2), 376–424.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in Handbook of Econometrics (J. J. Heckman and E. E. Leamer, eds.) volume 6, 5549–5632. Elsevier, Amsterdam.

Chen, X., and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80(1), 277–321.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405-1441.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 1, 1-12.

D'Haultfoeuille, X. and P. Fevrier (2014): "Identification of Nonseparable Models with Endogeneity and Discrete Instruments," *Econometrica*, forthcoming.

D'Haultfoeuille, X., Hoderlein, S. and Y. Sasaki (2013): "Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatments", Boston College Working Paper wp839.

DiNardo, J., and D. S. Lee. (2011): "Program Evaluation and Research Designs," In Handbook of Labor Economics, ed. O. Ashenfelter and D. Card, vol. 4A, 463-536. Elsevier Science B.V.

Dunker, F., Florens, J-P., Hohage, T., Johannes, J. and Mammen, E. (2014): "Iterative Estimation of Solutions to Noisy Nonlinear Operator Equations in Nonparametric Instrumental Regression", *Journal of Econometrics*, 178, 444-455.

Fan, J., and Gijbels, I. (1996): *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

Frolich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

Hall, P., and J. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 2904–2929.

Hahn, J., Todd, P., and van der Klaauw, W. (1999): "Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design." National Bureau of Economic Research Working Paper 7131.

Hahn, J., Todd, P., and van der Klaauw, W. (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–09.

Hoderlein, S., Holzmann, H. and Meister, A. (2015): "The Triangular Model with Random Coefficients," unpublished manuscript.

Horowitz, J. (2007): "Asymptotic normality of a nonparametric instrumental variables estimator," *International Economic Review,* 48, 1329–1349.

Horowitz, J. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79(2), 347–394.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 61, 2, 467-476.

Imbens, G. W. and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615–35.

Kasy M. (2009): "Semiparametrically Efficient Estimation of Conditional Instrumental Variable Parameters," *International Journal of Biostatistics*, 5 (1).

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature* 48, 281-355.

Lumley, J., C. Chamberlain, T. Dowswell, S. Oliver, L. Oakley, and L. Watson (2009): "Interventions for Promoting Smoking Cessation During Pregnancy (Cochrane Review)," *The Cochrane Library*, 8(3).

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

Masten, M. and Torgovitsky, A. (2014): "Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model," CEMMAP working paper CWP02/14.

Matzkin, R.L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-13785.

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

Santos, A. (2012): "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, 80, 213–275.

Severini, T. A., and G. Tripathi (2006): "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors," *Econometric Theory*, 22(2), 258–278.

Severini, T. A., and G. Tripathi (2012): "Efficency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors," *Journal of Econometrics*, 170(2), 491-498.

Torgovitsky, A. (2014): "Identification of Nonseparable Models using Instruments with Small Support," *Econometrica*, forthcoming.

van der Klaauw, W. (2008): "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics." *Labour*, 22(2): 219–45.