Identifying Marginal Effects with Binary Instruments or by Regression Discontinuity∗

Carolina Caetano† Juan Carlos Escanciano‡

University of Rochester Indiana University

This draft: October 14, 2014.

Preliminary and Incomplete. Comments are welcome.

Abstract

This paper proposes a new strategy for the identification of marginal effects of an endogenous multi-valued variable (which can be continuous, or a vector) in a regression model with one binary instrumental variable. The model must be separable in the error term. Identification is achieved by exploiting heterogeneity of the “first stage” in covariates. The covariates themselves may be endogenous, and their endogeneity does not need to be modeled. With some modifications, the identification strategy is then extended to the Regression Discontinuity Design (RDD) with multi-valued endogenous variables. Contrary to conventional wisdom, this paper shows that adding covariates in RDD may improve identification. This paper also provides parametric, semiparametric and nonparametric estimators based on the identification strategy, discusses their asymptotic properties, and shows that the estimators have satisfactory performance in moderate samples sizes.

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 and Kenneth Chay for helpful discussions.
†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 ECO2012-33053.
1 Introduction

Instrumental variables (IV) methods are well established as one of the most useful approaches to identify causal effects in econometric models. A flexible extension of the classical linear model is given by the nonparametric structural regression model

\[ Y = g(X) + \varepsilon, \]  

(1)

where \( Y \) is the dependent variable, \( X \) is endogenous, in the sense that \( \mathbb{E}[\varepsilon|X] \neq 0 \) with positive probability, and \( \varepsilon \) is a structural unobservable error term.

In order to identify the structural regression function \( g(\cdot) \) the typical IV identification strategy requires the existence of an instrument \( T \) satisfying a validity condition: (i) \( T \) is exogenous in the sense that \( \mathbb{E}[\varepsilon|T] = 0 \) almost surely (a.s.); and a relevance condition: (ii) the dependence between \( X \) and \( T \) is “strong enough.” The relevance condition (ii) requires the instrument \( T \) be at least as “complex” as the endogenous variable \( X \), meaning that, for instance, if \( X \) is continuous then \( T \) should be continuous, or if \( X \) is discrete with \( q \) points of support, then \( T \) needs to have at least \( q \) points of support, and if \( X \) is a vector, then \( T \) must have at least as many components as \( X \); see Newey and Powell (2003). For identification and inference results in nonparametric IV see, e.g., Newey and Powell (2003), Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Darolles, Fan, Flores and Renault (2011), Horowitz (2011) and Chen and Pouzo (2012), to mention just a few.

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 the traditional IV’s identification fails. Our validity condition allows the instrument to be mean independent of the error term only conditional on a covariate \( Z \) \( (\mathbb{E}[\varepsilon|T,Z] = \mathbb{E}[\varepsilon|Z]) \), which may be endogenous.\(^1\) Additionally, we do not require that \( T \) be at least as complex as \( X \). 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 a simple experiment or a natural experiment.

Furthermore, when \( X \) is continuous or discrete with more than two points of support, the state of the art in the Regression Discontinuity Design (RDD) literature cannot identify \( X \)’s marginal effects (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.) However, we show that since we can think of the RDD as a limit case of the binary instrumental variable case, we can apply our methodology to RDD (with a few adaptations) and achieve the identification of \( X \)’s marginal effects.

Our identification strategy is based on the observation that the exogenous variation that \( T \) induces on the distribution of \( X \), although discrete, may be different for different sub-populations defined by

\(^1\)An instrument satisfying our validity condition is referred to as a Conditional IV (CIV) in the literature, see, e.g. Frolich (2007) and Kasy (2009), among others. None of these papers discussed identification in the nonparametric regression model (1).
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 Consider a linear model with three endogenous variables \((X_1, X_2, Z)\) such that

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.
\]

Standard IV methods are unable to identify \(\beta_1\) and \(\beta_2\) when \(T\) is binary. Under the conditional exogeneity of \(T\) defined by \(E[\varepsilon|Z, T] = E[\varepsilon|Z]\), we can write

\[
E[Y|T = 1, Z] - E[Y|T = 0, Z] = \beta_1 [E[X_1|T = 1, Z] - E[X_1|T = 0, Z]] + \beta_2 [E[X_2|T = 1, Z] - E[X_2|T = 0, Z]].
\]

If the effects of the binary instrument on the endogenous variables \(X_1\) and \(X_2\) are different, in the sense that there is no \(\lambda \neq 0\) such that

\[
E[X_1|T = 1, Z] - E[X_1|T = 0, Z] = \lambda [E[X_2|T = 1, Z] - E[X_2|T = 0, Z]] \quad \text{a.s.}
\]

then, \(\beta_1\) and \(\beta_2\) are identified. Our identification strategy exploits the heterogeneity of the “first stages” in covariates to separate the effects of \(X_1\) and \(X_2\).

Based on our identification results, we propose new parametric, semiparametric and nonparametric estimators of \(g\) (up to a constant). The proposed estimators are all two-step least squares estimators. For parametric and semiparametric estimators of linear structural models we establish their asymptotic normality and discuss simple ways to obtain consistent standard errors. These estimators overcome the “curse of dimensionality” problem present in nonparametric methods when the dimension of \(Z\) is moderate or large. We consider a semiparametric varying coefficients specification under RDD, which provides a practically convenient way to incorporate covariates’ heterogeneity in applications, including in the well-studied case of binary \(X\) (see Hahn, Todd and van der Klaauw (1990)). But the strength of the new semiparametric estimator is best seen when \(X\) is continuous or multi-valued. While there are currently no methods available for identification in this case, the semiparametric estimator under RDD provides a method to identify marginal effects. So, contrary to conventional wisdom, we show that adding covariates in RDD may improve identification. We also provide theoretical and Monte Carlo evidence supporting that both parametric and semiparametric estimators possess certain robustness properties to misspecification of the first step when the structural equation is correctly specified. 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).

Our approach has some connections with the local average treatment effect literature (see Imbens and Angrist (1994)) 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 instrument, whereas here we are interested in continuous endogenous variables and a binary instrument. Our results also complement alternative identification strategies.
for binary instruments and continuous endogenous variables in Chesher (2003), D’Haultfoeuille and Fevrier (2012), Torgovitsky (2012), and Masten and Torgovitsky (2014). None of these papers exploit the heterogeneity of the “first stage” in covariates.

The rest of the paper is organized as follows. Section 2 develops an example in detail which shows explicitly what are the fundamental requirements of our method, as well as why it works. Section 3 focuses on the binary IV case. It presents the identification results, and it also proposes an estimator for the slope of a vector of endogenous variables in a model where all the regression functions are linear in $Z$. Section 4 focuses on the RDD case. It extends the identification results of the binary IV case to the RDD, and it also proposes an estimator for a semiparametric varying coefficient specification which deals with the curse of dimensionality problem when the vector $Z$ is of moderate or high dimension. We 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 in house prices, based on Chay and Greenstone (2005). This example is particularly interesting 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. We conclude on Section 7.

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) for a survey of the literature on this problem.) The variable of interest, “smoking,” is naturally prone to endogeneity, given that there are many pre-existing selection factors associated both with smoking and with birth weight. Examples of such factors include the mother’s education level, marital status, age, etc. There is a notorious shortage of instruments for smoking. Existing options explore cigarette tax changes or variations across localities, but these instruments tend to be weak, meaning that pregnant women tend to change their smoking behavior very little because of the changes in cigarette prices. Experimental approaches fare better, with the intervention causing smoking cessation on average in 6% of the population (cite Lumley et al. (2011)). However, experiments provide only a binary instrument, so in the standard approach it is impossible to uncover the effects of each marginal cigarette.

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 hopefully quit smoking, while the women in the control group are merely observed. $T$ 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, experimental data does not allow us to 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 where we do this. Column (I) represents the years of education,

<table>
<thead>
<tr>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
<th>(IV)</th>
<th>(V)</th>
<th>(VI)</th>
<th>(VII)</th>
<th>(VIII)</th>
<th>(IX)</th>
<th>(X)</th>
<th>(XI)</th>
<th>row #</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>$\bar{X}_{0.Z}$</td>
<td>$\bar{X}_{1.Z}$</td>
<td>$\Delta_Y(Z)$</td>
<td>$P_{0.Z}(0)$</td>
<td>$P_{0.Z}(1)$</td>
<td>$P_{0.Z}(3)$</td>
<td>$P_{1.Z}(0)$</td>
<td>$P_{1.Z}(1)$</td>
<td>$P_{1.Z}(3)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/5</td>
<td>1/5</td>
<td>3/5</td>
<td></td>
<td>(2)</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>2</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/3</td>
<td>0</td>
<td>2/3</td>
<td></td>
<td>(3)</td>
</tr>
</tbody>
</table>

Table 1: Identification Idea

and we are at first considering only 3 possibilities: 6, 10 and 17. Columns (II) and (III) show the average amount smoked among 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 there is on average no variation at all in the “first stages.” Column (IV) shows the average difference (in grams) in the birth weight between the treatment and the control groups for that level of education ($\Delta_Y(Z) = E[Y|T = 1, Z] - E[Y|T = 0, Z]$).

The first fundamental assumption of our method is a validity condition, which requires that $E[\varepsilon|T, Z] = E[\varepsilon|Z]$. 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 ($\Delta_Y(Z) = E[g(X)|T = 1, Z] - E[g(X)|T = 0, Z]$). In this example, the smoking variable $X$ takes the values 0, 1 or 3 (generalization for a more complex $X$ will become clear later). 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, and we will try to use all that we
observe to our advantage. Columns (V) to (VII) show the smoking distribution in the control group. 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 co-opted 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 varied. 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 effects on average 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 outcomes differences 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 assumption by the combination of the model in (1), which is separable in the error term, with the validity condition. Together these conditions yield $E[Y|T = 1, Z] − E[Y|T = 1, Z] = E[g(X)|T = 1, Z] − E[g(X)|T = 0, Z].$ 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 outcome between treatment and control are the result of the combination of the possible behaviors with the differences in probabilities. For example, row (1) gives us $10 = E[Y|T = 1, Z = 6] − E[Y|T = 0, Z = 6] = 0 \cdot g(0) + 0.5 \cdot g(1) + 0.5 \cdot g(3) − 0 \cdot g(0) + 0 \cdot g(1) + 1 \cdot g(3)] = 0.5 \cdot g(1) − 0.5 \cdot g(3).$ 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:

\[
\begin{align*}
0.5g(1) - 0.5g(3) &= 10 \quad (2) \\
0.2g(0) + 0.2g(1) - 0.4g(3) &= 22 \quad (3) \\
0.33g(0) - 0.33g(3) &= 30 \quad (4)
\end{align*}
\]

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 of equations has 3 equations and 3 unknowns. However, only two 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 numbers 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 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 linear combination of these. In particular, we can recover the value of any increment. It is straightforward to see in this example that, from equation (4), \(g(3) - g(1) = -20\), from equation (2), \(g(3) - g(0) = -90\), 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 explained 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

We consider a vector of observed variables \((Y, X, Z, T)\) satisfying the model

\[ Y = g(X) + \varepsilon, \]  

(5)

and the exclusion restriction \(\mathbb{E}[\varepsilon|Z, T] = \mathbb{E}[\varepsilon|Z]\). Then, the exclusion restriction yields the equality

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

(6)

Identifying any functional of \(g\) 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** (\(X\) discrete) Denote by \(S_X := \{x_1, \ldots, x_q\}\) and \(S_Z := \{z_1, \ldots, z_l\}\) the supports of the distributions of \(X\) and \(Z\), respectively. The classic nonparametric IV identification strategy is based on the equation

\[ \mathbb{E}[Y|T] = \mathbb{E}[g(X)|T], \]

which in this case translates into the equation \((\mathbb{E}[Y|T = 0], \mathbb{E}[Y|T = 1])' = Pg\), where \(P\) is the \(2 \times q\) matrix with column \(j\) equal to \((\mathbb{P}(X = x_j|T = 0), \mathbb{P}(X = x_j|T = 1))'\), \(j = 1, \ldots, q\), and \(g = (g(x_1), \ldots, g(x_q))'\) (\(A'\) denotes the transpose of \(A\)). In this classic setting the matrix \(P\) has a rank of at most 2, and so this equation is not identified if \(q > 2\) (see Newey and Powell (2003)). In fact, we can only identify linear functionals \(c'g\) where \(c\) is spanned by the two rows of \(P\) (see Severini and Tripathi (2006, 2012)), which are not necessarily of interest.

Our identification strategy consists of inverting equation (6), which in this context can be written as \((m(z_1), \ldots, m(z_l))' = Ag\), where \(m(z) := \mathbb{E}[Y|T = 1, Z = z] - \mathbb{E}[Y|T = 0, Z = z], z \in S_Z\). The matrix \(A := P_1 - P_0\), where \(P_t = (p_{tij})\) is the \(l \times q\) matrix which has entries \(p_{tij} = \mathbb{P}(X = x_j|T = t, Z = z_i)\), \(i = 1, \ldots, l, j = 1, \ldots, q\) and \(t = 0, 1\). Notice that since \(P_0\) and \(P_1\) are matrices of probabilities, \(A t = 0\), where \(t\) denotes the \(q \times 1\) vector of ones. Since \(A\) is not full-rank, \(g\) is actually not identified.

However, in this context we can identify linear functionals \(c'g\) with \(c\) in a space of dimension \(\text{rank}(A)\). In particular, if \(\text{rank}(A) = q - 1\), then all linear functionals \(c'g\) with \(c' t = 0\) 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 \(l \geq q - 1\), so \(Z\) needs to assume at least \(q - 1\) different values.

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

\[ m = Ag, \]  

(7)

where now \(Ag := \mathbb{E}[g(X)|Z, T = 1] - \mathbb{E}[g(X)|Z, T = 0]\) denotes a continuous (i.e. bounded) linear operator \(A : G \subseteq L_2(X) \to L_2(Z)\), where henceforth, for generic random vector \(\zeta\), \(L_2(\zeta)\) denotes
the Hilbert space with squared-integrable functions with respect to the distribution of \( \zeta \). 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 \( A \) be composed exclusively of the constant functions:

**Assumption 1** \( \mathcal{N}(A) = \{ c : c \in \mathbb{R} \} \).

Notice that the identification condition in Example 3.1 that \( \text{rank}(A) = q - 1 \) is equivalent to \( \mathcal{N}(A) = \{ cu : c \in \mathbb{R} \} \). In a general case, Assumption 1 is the analogue to the completeness condition required in nonparametric IV (see Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Andrews (2011) and D'Haultfoeuille (2011) for discussion on completeness). To compare these two identification assumptions, consider for simplicity the univariate case. The classical completeness with binary IV (see Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Andrews (2011) and D'Haultfoeuille (2011) for discussion on completeness). To compare these two identification assumptions, consider for simplicity the univariate case. The classical completeness with binary IV requires that for each non-constant \( \lambda(\cdot) \in L_2(X) \), \( \lambda(X) \) and \( T \) are correlated, that is, \( \beta IV \) in the least squares regression \( \lambda(X) = \alpha IV + \beta IV T + \varepsilon \), is non-zero, which is not possible when e.g. \( X \) is continuous. On the other hand, Assumption 1 requires that for such \( \lambda(\cdot) \) there exists a \( \phi(\cdot) \in L_2(Z) \) such that \( \alpha_1 \lambda \) or \( \alpha_3 \lambda \) are non-zero in the least squares regression \( \lambda(X) = \alpha_0 + \alpha_1 T + \alpha_2 \phi(Z) + \alpha_3 \phi(Z) T + u_\lambda \). In fact, Assumption 1 is equivalent to the latter statement, which can be shown along the lines of Lemma 2.1 in Severini and Tripathi (2006). That Assumption 1 implies that \( g \) is identified up to a constant follows from the fact that \( Ag = m \), with \( g(X) = g(X) - E[g(X)] \), and \( A \) is invertible on the orthocomplement of \( \mathcal{N}(A) \), say \( \mathcal{N}^\perp = \{ \lambda \in \mathcal{G} : E[\lambda(X)] = 0 \} \).

Necessary conditions for Assumption 1 are that \( X \) and \( Z \) have the same level of complexity (e.g. both are continuous) and that \( g \) is not a nonparametric function of \( Z \) and \( X \) (separability). Example 3.4 below shows that we can allow for parametric or semiparametric non-separability in \( Z \) and \( X \). Intuitively, what is needed for Assumption 1 to hold is that the differences between the distributions of \( X \) under treatment and control groups varies sufficiently with \( Z \). The following examples show the exact meaning of Assumption 1 in some special cases.

**Example 3.2** \( (X,Z) \text{ jointly normal} \) Suppose that

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

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

\[
E[g(X)|Z = z, T = t] = (2\pi)^{-3/4} \exp \left( -\frac{z^2}{2} \right) \sum_{j=0}^{\infty} \mu^j(\rho_1)E[g(X)p_j(X)] \frac{z^j}{\sqrt{j!}},
\]

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,

\[
Ag(z) = (2\pi)^{-3/4} \exp \left( -\frac{z^2}{2} \right) \sum_{j=0}^{\infty} \left\{ \mu^j(\rho_1) - \mu^j(\rho_0) \right\} E[g(X)p_j(X)] \frac{z^j}{\sqrt{j!}}.
\]
By the completeness of the Hermite polynomials, Assumption 1 holds with \( G = L_2(X) \), provided \( \rho_1 \neq \rho_0 \). Notice that if \( f_{X|T,Z} \) denotes the density of \( X \) conditional on \( T \) and \( Z \), then
\[
f_{X|T=t,Z=z}(x) = \frac{1}{\sqrt{2\pi(1-\rho^2_t)}} \exp\left(-\frac{(x-\rho_t z)^2}{2(1-\rho^2_t)}\right).
\]
Therefore, the condition \( \rho_1 \neq \rho_0 \) guarantees that the difference between the distribution of \( X \) between treatment and control groups varies with \( Z \).

**Example 3.3** (linear model) Suppose that \( g \) is linear, so that the model is
\[
Y = \alpha + \beta' X + \epsilon, \quad \mathbb{E}[\epsilon|Z,T] = \mathbb{E}[\epsilon|Z],
\]
where \( X \) is a \( d \)-dimensional vector. In this model, the integral equation can be written as
\[
\mathbb{E}[Y|Z,T = 1] - \mathbb{E}[Y|Z,T = 0] = \beta' (\mathbb{E}[X|Z,T = 1] - \mathbb{E}[X|Z,T = 0]),
\]
or in short (using the notation \( \Delta_V = \mathbb{E}[V|Z,T = 1] - \mathbb{E}[V|Z,T = 0] \))
\[
\Delta_V = \beta' \Delta_X.
\]
Hence, in this example \( G = \{a + b'X : a \in \mathbb{R}, b \in \mathbb{R}^d\} \) and Assumption 1 is equivalent to
\[
\mathbb{E}[\Delta_X \Delta'_X] \text{ is positive definite.}
\]
It is straightforward to see why \( \beta \) 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. It is interesting to 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.

The linear multivariate model contains many interesting examples as special cases. For instance, the discrete case discussed above can be written in this framework since for a discrete endogenous variable \( D \) with support \( S_D := \{d_1, \ldots, d_q\} \),
\[
g(D) = g(d_1) + (g(d_2) - g(d_1))1(D = d_2) + \cdots + (g(d_q) - g(d_1))1(D = d_q)
\equiv \alpha + \beta' X,
\]
where \( \alpha = g(d_1) \), \( X = (1(D = d_2), \ldots, 1(D = d_q))' \), \( 1(\cdot) \) denotes the indicator function, and \( \beta = (g(d_2) - g(d_1), \ldots, g(d_q) - g(d_1))' \) denotes the increment effects of interest. This setting also includes piecewise linear models, with the simplest case being:
\[
Y = \alpha + \beta_1 D + \beta_2 D 1(D > 0) + \epsilon,
\]
where \( D \) is an endogenous variable, or models with infinite but parametric variation in the marginal effects, such as
\[
Y = \alpha + \beta_1 S + \beta_2 S^2 + \epsilon.
\]
Classic IV identification conditions require two instruments in this example. 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 $X = (S, S^2)'$, respectively.

**Remark 3.1** It is important to notice that although Z can be endogenous, and thus not excludable from the structural equation (1), this is not necessary. However, if Z is exogenous, then Z is also an instrumental variable in the classical sense. In this case the choice of approach depends mostly on the strength of the identification with each method. In the classic IV method, the identification depends on the distribution of $X$ conditional on $Z = z$ varying sufficiently for different values $z$. In our method, the identification depends on the differences in the distribution on $X$ under treatment and control conditional on $Z = z$ varying sufficiently for different values of $z$. We also note that both approaches can be combined. For instance, we can use our approach to identify marginal effects, and the standard IV to identify the location of $g$.

**Remark 3.2** Although the validity condition $E[U|T,Z] = E[U|Z]$ in our approach is the same used in the classical IV method with covariates (see Frolich (2007) and Kasy (2009)), in our method it is never necessary to specify how $Z$ enters the structural equation. Hence, our method is robust to the misspecification of the functional form of $E[ε|Z]$.

**Remark 3.3** It should be noted that Assumption 1 is different from the completeness between $X$ and $Z$, but it can be understood as a weighted completeness assumption between these two variables. To see this, define the propensity score $p(z) = E[T|Z = z]$, assume $0 < p(z) < 1$ a.s., and note that Assumption 1 is equivalent to: $E[λ(X)(T − p(Z))|Z] = 0$ a.s. $⇒ λ(X) = E[λ(X)]$ a.s. In contrast, the classical $L_2$ completeness between $X$ and $Z$ is equivalent to $E[λ(X)|Z] = 0$ a.s. $⇒ λ(X) = 0$ 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 $Y(T − p(Z)) = g(X)(T − p(Z)) + ε(T − p(Z))$.

**Remark 3.4** To apply our method in a parametric setting one may want to fit $E[Y|T = t, Z]$ and $E[X|T = t, Z]$ for $t = 0, 1$, by, e.g., linear projections. Our simulations (Section 5) show that the method is robust to misspecifications from linearity in the first step.

**Remark 3.5** Although our method cannot identify nonparametric non-separable heterogeneous effects in $X$ and $Z$, i.e. $g$ cannot be a nonparametric function of both, we can extend our results to some parametric and semiparametric models with marginal effects of $X$ varying with $Z$, as the following example illustrates.

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

$$Y = \alpha(Z) + \beta(Z_1)'X + ε, \quad E[ε|Z, T] = E[ε|Z],$$

where $X$ is a $d$-dimensional vector and $Z = (Z_1', Z_2')'$. In this model, we can write

$$E[ΔY|Z_1] = \beta(Z_1)'E[ΔX|Z_1],$$
which can be used to identify \( \beta(z_1) \) provided

\[
E[E[\Delta X | Z_1]E[\Delta X' | Z_1]] \text{ is positive definite.}
\]

Under this assumption, we can estimate \( \beta(\cdot) \) from local least squares regressions, similar to those carried out for the RRD below. By comparing \( \beta(\cdot) \) with the estimator obtained from \( \beta = (E[\Delta X \Delta X'])^{-1} E[\Delta X \Delta Y] \) we can test for heterogenous in \( Z_1 \) effects.

### 3.2 Estimation

In this section we discuss estimation when \( g \) is linear, as in (8). To see the discussion of the estimation in the nonparametric case, refer to Appendix A.2.1. Suppose that we have a random sample \( \{(Y_i, X_i, Z_i, T_i)\}_{i=1}^n \) of \( (Y, X, Z, T) \). The estimation strategy consists of two steps. First estimate the conditional expectations \( E[Y|T = t, Z] \) and \( E[X|T = t, Z] \) for \( t = 0, 1 \). Second, plug the estimated conditional expectations into equation (9) and estimate \( \beta \) by ordinary least squares (OLS).

**Step 1**

We assume that the conditional means are linear in \( Z \). More precisely, we assume that for both \( V = Y \) and \( V = X \),

\[
E[V|T, Z] = \alpha_{0V} + \alpha_{1V}T + \alpha_{2V}'Z + \alpha_{3V}'ZT = \alpha_V'Z,
\]

where \( \alpha_V = (\alpha_{0V}, \alpha_{1V}, \alpha_{2V}, \alpha_{3V})' \) and \( S = (1, T, Z', Z'T)' \). Denote by \( \hat{\alpha}_V \) the OLS estimator in (10).

The assumption of linearity is made here for convenience. In fact, our Monte Carlo simulations show that our estimator is robust to the misspecification of the linear functional form assumption in (10). Nevertheless, if desired one could estimate the terms \( E[V|T, Z] \) nonparametrically and still preserve the \( \sqrt{n} \)-consistency and asymptotic normality of the estimator (although the asymptotic variance will be different).

**Step 2**

In this setting, (9) applies with \( \Delta_Y = \alpha_{1Y} + \alpha_{3Y}'Z \) and \( \Delta_X = \alpha_{1X} + \alpha_{3X}'Z \). Then, \( \hat{\beta} \) can be obtained by the OLS regression of the \( \hat{\Delta}_Y \) on the \( \hat{\Delta}_X \), where \( \hat{\Delta}_X = \hat{\alpha}_{1X} + \hat{\alpha}_{3X}'Z_i \) and \( \hat{\Delta}_Y = \hat{\alpha}_{1Y} + \hat{\alpha}_{3Y}'Z_i \). \( \hat{\beta} \) has the closed-form expression

\[
\hat{\beta} = \left( \sum_{i=1}^n \hat{\Delta}_X_i \hat{\Delta}_X_i' \right)^{-1} \sum_{i=1}^n \hat{\Delta}_X_i \hat{\Delta}_Y_i.
\]

**Asymptotic behavior of \( \hat{\beta} \)**

To derive the asymptotic distribution theory for the proposed estimator, use \( \Delta_{Y_i} = \beta_0' \Delta_{X_i} \) and substitute \( \hat{\Delta}_{Y_i} = \beta_0' \hat{\Delta}_{X_i} + \hat{\Delta}_{Y_i} - \Delta_{Y_i} - (\hat{\Delta}_{X_i} - \Delta_{X_i})' \beta_0 \) in \( \hat{\beta} \) to get the expansion

\[
\hat{\beta} = \beta_0 + \left( \frac{1}{n} \sum_{i=1}^n \hat{\Delta}_{X_i} \hat{\Delta}_{X_i}' \right)^{-1} \left( \sum_{i=1}^n \hat{\Delta}_{X_i}(\hat{\Delta}_{Y_i} - \Delta_{Y_i}) - \sum_{i=1}^n \hat{\Delta}_{X_i}(\hat{\Delta}_{X_i} - \Delta_{X_i})' \beta_0 \right).
\]
Then, using standard OLS theory we can further write
\[
\hat{\beta} - \beta_0 = (\mathbb{E}[\Delta X_i \Delta' X_i])^{-1} \mathbb{E}[\Delta X_i \hat{Z}_i] B_1 (\hat{\alpha} - \alpha) + o_P(n^{-1/2}),
\]
where \( \hat{Z}_i = (1, Z_i')', \hat{\alpha} = (\hat{\alpha}_Y, vec(\hat{\alpha}_X)'), \alpha = (\alpha_Y, vec(\alpha_X)')' \) (vec is the vectorization operator) and \( B_1 \equiv B_1(\beta_0) \) is given by
\[
B_1 = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & -\beta_0' & 0 & 0 \\
0 & 0 & 0 & 0 & -\beta_0'
\end{bmatrix},
\]
with \( 0 \) a \( 1 \times 4 \) vector of zeros. The asymptotic normality for \( \hat{\beta} \) then follows from that of the joint OLS estimator \( \hat{\alpha} \) and the expression (11). Define
\[
Q_1 = (\mathbb{E}[\Delta X_i \Delta' X_i])^{-1} \mathbb{E}[\Delta X_i \hat{Z}_i] B_1 \quad \text{and} \quad \Omega = \lim_{n \to \infty} \text{Var} \left( \sqrt{n}(\hat{\alpha} - \alpha) \right).
\]
The have the following result. The proof is omitted because it is standard.

**Theorem 3.1** Assume: (i) \( \mathbb{E}[\varepsilon | Z, T] = \mathbb{E}[\varepsilon | Z] \) and (ii) \( \mathbb{E}[\Delta X_i \Delta' X_i] \) and \( \Omega \) are finite and non-singular, and \( \mathbb{E}[\Delta X_i \hat{Z}_i] \) is finite. Then,
\[
\sqrt{n}(\hat{\beta} - \beta) \to_d N(0, \Sigma),
\]
where \( \Sigma = Q_1 \Omega Q_1' \).

Estimation of standard errors can be easily constructed from the previous expressions. Any software that provides consistent standard errors for Seemingly Unrelated Equations (SUR) with the same covariates can be used to obtain a consistent estimate of \( \Omega \), say \( \hat{\Omega} \). Then, we estimate \( \Sigma \) by \( \hat{Q}_1 \hat{\Omega} \hat{Q}_1' \), where
\[
\hat{Q}_1 = \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\Delta} X_i \hat{\Delta}' X_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\Delta} X_i \hat{Z}_i' \right) B_1(\hat{\beta}).
\]

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, some examples among the applied works of this literature are interested in the effects of a variable which can take multiple values. One such example can be seen in our application in Section 6. We need to extend the notation to allow for this generalization of the endogenous variable of interest \( X \). 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. Our choice is justified because our model cannot allow for nonparametric heterogeneous effects, since the separability of the covariate is fundamental to our approach. Additionally, just as in the fuzzy design, the endogenous variable in our setting may take potentially any value both below and above the threshold, though it must do so with different probabilities.
Let the potential outcome random variable $Y(x)$ satisfy the model\textsuperscript{2}

$$Y(x) = g(x) + \varepsilon.$$  

(12)

Let the “predictor” variable be denoted $W$, which is univariate and continuously distributed, with threshold $\bar{w}$. The quantities of interest are $E[Y(x_i) - Y(x_j)|W = \bar{w}]$, for $x_i \neq x_j$. Given equation (12), these are the same as the increments $g(x_i) - g(x_j)$.

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.

\textbf{Assumption 2} $E[\varepsilon|W = w, Z]$ is continuous in $w$ at $\bar{w}$ with probability equal to one.\textsuperscript{3}

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

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

The right hand side defines implicitly a linear operator of $g$, say $Ag$, and Assumption 1 applied to $A$ implies that $g$ is identified up to a constant. The interpretation of this assumption in the context of the RDD is analogous to the cases discussed in Section 3.1. 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.

\textbf{Example 4.1} \textit{Consider the linear model under Assumption 2,}

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

\textit{so that}

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

(13)

\textit{or using the short notation }\delta_Y = \beta'\delta_X, \textit{where we denote for a generic random vector }V

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

\textit{assuming the limits exist. Assumption 1 in this context is equivalent to }$E[\delta_X \delta'_X]$ \textit{being positive definite.}

\textsuperscript{2}Hahn, Todd and van der Klaauw (1990) denote the potential outcome $y_i = \alpha + x_i \cdot \beta$. In our case, $\alpha$ is represented by $\varepsilon$ and $g(x)$ is the generalization of $x \cdot \beta$.

\textsuperscript{3}Assumption (A1) in Hahn, Todd and van der Klaauw (1990).
4.2 Estimation

In this section we discuss estimation 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. As in the binary IV case, the estimation strategy consists of two steps. First, estimate $\lim_{w \uparrow \bar{w}} \mathbb{E}[Y|W = w, Z]$ for both $V = Y$ and $V = X$, which provides a natural extension of (10) to the continuous case. Without loss of generality assume the cutoff point is $\bar{w} = 0$. To simplify notation denote

$$\lim_{w \uparrow \bar{w}} \alpha_{0V}(w) = \alpha_{0V}^+ \quad \lim_{w \downarrow \bar{w}} \alpha_{0V}(w) = \alpha_{0V}^-$$

and

$$\lim_{w \uparrow \bar{w}} \alpha_{1V}(w) = \alpha_{1V}^+ \quad \lim_{w \downarrow \bar{w}} \alpha_{1V}(w) = \alpha_{1V}^-.$$

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

$$\hat{\alpha}_V^+ = \arg\min_{\alpha_V^+} \sum_{i=1}^n (V_i - \alpha_{0V}^+ - \alpha_{1V}^+Z_i - \alpha_{2V}^+X_i - \alpha_{3V}^+Z_iW_i)^2 k_h(W_i) 1(W_i \geq 0),$$

where $k_h(W) = k(W/h)$, $k$ is a kernel function and $h$ is a bandwidth parameter satisfying some conditions in Section A.1 in the Appendix (Assumption RDD). The estimation of $\hat{\alpha}_{0Y}^-$ and $\hat{\alpha}_{1Y}^-$ for $V = Y$ and $V = X$ is analogous, replacing $1(W_i \geq 0)$ in the minimization problem by $1(W_i < 0)$.

Step 2

We can now apply the same approach that we adopted in the linear model to equation (13). Notice that in this setting $\delta_{Xi} = \alpha_{0X}^+ - \alpha_{0X}^- + (\alpha_{1X}^+ - \alpha_{1X}^-)'Z_i$ and $\delta_{Yi} = \alpha_{0Y}^+ - \alpha_{0Y}^- + (\alpha_{1Y}^+ - \alpha_{1Y}^-)'Z_i$, and thus we can obtain $\hat{\delta}_{Yi}$ and $\hat{\delta}_{Xi}$ by substituting the estimated $\alpha$’s from the previous step. Then, $\hat{\beta}$ is obtained from the OLS regression of the $\hat{\delta}_{Yi}$ on the $\hat{\delta}_{Xi}$, and has the closed-form expression

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

Asymptotic behavior of $\hat{\beta}$

The estimator $\hat{\beta}$ is a two-step semiparametric estimator, whose asymptotic normality is established in the following result. For simplicity we only consider a univariate $X$ and $Z$. The multivariate case can be obtained analogously.
Define the vectors 
\[ \hat{\delta}_h = (\hat{\delta}_0, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3, \hat{\delta}_0 X, \hat{\delta}_1 X, \hat{\delta}_2 X, \hat{\delta}_3 X)', \delta_h = (\delta_0, \delta_1, \delta_2, \delta_3, \delta_0 X, \delta_1 X, \delta_2 X, \delta_3 X)', \]
where \( \hat{\delta}_j = \hat{\alpha}_j + \alpha_j - \hat{\alpha}_j - \alpha_j \) for \( j = 0, 1, 2, 3 \) and \( V = Y \) or \( X \). Similarly to the arguments for the parametric case, it can be shown from Section A.1 in the Appendix that
\[ \hat{\beta} - \beta_0 = (E[\delta^2])^{-1} E[\delta X Z_i] B_2 (\hat{\delta}_h - \delta_h) + o_P((nh)^{-1/2}), \]
where \( Q_2 = (E[\delta^2])^{-1} E[\delta X Z_i] B_2 \), and
\[ B_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & -\beta_0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -\beta_0 & 0 & 0 \end{bmatrix}. \]
Following similar arguments to those of Hahn, Todd and van der Klaauw (1990) we show in Section A.1 in the Appendix (cf. Lemma A.8) that
\[ \sqrt{nh} (\hat{\delta}_h - \delta_h) \to_d N(\mu, \Omega), \]
and provide expressions for \( \mu \) and \( \Omega \). The non-zero bias \( \mu \) arises from the use of the bandwidth \( h = \rho n^{-1/5} \), for a positive constant \( \rho \). If instead, \( nh^5 \to 0 \) as \( n \to \infty \), then \( \mu = 0 \). The following result follows. The proof can be found in Section A.1 in the Appendix.

**Theorem 4.1** Under Assumption RRD in Section A.1 in the Appendix,
\[ \sqrt{nh}(\hat{\beta} - \beta_0) \to_d N(Q_2 \mu, Q_2 \Omega Q_2'). \]

5 Monte Carlo Simulations

Consider the following data generating process (DGP)
\[ Y = \alpha + \beta D + \beta_1 D1(D > 0) + \gamma Z + U, \]
\[ Z = \alpha Z + \varepsilon, \]
\[ D = \alpha T \times Z + \varepsilon, \]
where \((U, \varepsilon, \varepsilon, \varepsilon)\) are independent standard normal random variables, independent of \( T \), which is distributed as Bernoulli with probability \( p = 0.5 \). This corresponds to a linear model
\[ Y = \alpha + \beta_0 X + \varepsilon, \]
where \( \beta_0 = (\beta_1, \beta_2)' \) and \( X = (D, D1(D > 0))' \). Here
\[ \mathbb{E}[D|Z, T = 1] - \mathbb{E}[D|Z, T = 0] = \gamma Z, \]
so \( \gamma \) controls the identification strength. Note that when \( \alpha_z \neq 0 \) \( Z \) is endogenous. Since there is only one binary IV, \( T \), standard IV methods do not work in this example. Note also that under this
DGP the difference of conditional means $\Delta_{X_1}$ is nonlinear in $Z$. 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 linearity assumption in the conditional means $E[V|Z,T] := \alpha_0V(T) + \alpha_1V(T)Z$ for $V = Y$ and $X$.

Table 2 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 three levels of identification, “low” $\gamma_d = 0.5$, “moderate” $\gamma_d = 1$ and “high” $\gamma_d = 2$.

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>Bias($\beta_1$)</th>
<th>MSE($\beta_1$)</th>
<th>Bias($\beta_2$)</th>
<th>MSE($\beta_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_d = 0.5$</td>
<td>50</td>
<td>-0.82761</td>
<td>17986.7</td>
<td>1.12719</td>
<td>29447.5</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.32646</td>
<td>1445.03</td>
<td>0.29936</td>
<td>2082.79</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.00934</td>
<td>1.73402</td>
<td>-0.01057</td>
<td>6.04405</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.00293</td>
<td>0.03323</td>
<td>-0.00717</td>
<td>0.08917</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.00009</td>
<td>0.01420</td>
<td>-0.00160</td>
<td>0.03732</td>
</tr>
<tr>
<td>$\gamma_d = 1$</td>
<td>50</td>
<td>0.00228</td>
<td>20.0298</td>
<td>0.01411</td>
<td>45.1597</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.00107</td>
<td>0.48403</td>
<td>-0.01446</td>
<td>2.64655</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-0.00067</td>
<td>0.01124</td>
<td>0.00039</td>
<td>0.03006</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.00001</td>
<td>0.00638</td>
<td>0.00042</td>
<td>0.01664</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.00003</td>
<td>0.00303</td>
<td>-0.00072</td>
<td>0.00802</td>
</tr>
<tr>
<td>$\gamma_d = 2$</td>
<td>50</td>
<td>0.00532</td>
<td>0.76305</td>
<td>-0.00327</td>
<td>1.09976</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.00160</td>
<td>0.00924</td>
<td>0.00320</td>
<td>0.02321</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-0.00065</td>
<td>0.00252</td>
<td>0.00089</td>
<td>0.00645</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.00000</td>
<td>0.00148</td>
<td>0.00024</td>
<td>0.00385</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.00031</td>
<td>0.00074</td>
<td>-0.00019</td>
<td>0.00189</td>
</tr>
</tbody>
</table>

The reported results show that when identification is “weak” ($\gamma_d = 0.5$), sample sizes of $n = 300$ or larger are needed to attain a small bias, and larger than 300 to observe a low variance. Estimates of $\beta_2$ require larger sample sizes than those of $\beta_2$ to achieve the same level of precision and bias performance. The MSE is small in both cases when $n = 500$, and it decreases, as expected, for $n = 1000$. The results improve when identification is “moderate” ($\gamma_d = 1$), and are excellent when identification is “strong” ($\gamma_d = 1$), obtaining very good precisions already for $n = 100$. In all cases, $\beta_1$ is estimated at a much better precision than $\beta_2$, which can be explained by the nonlinearities in $E[D_1(D > 0)|Z,T = j]$ for $j = 0,1$, and the additional uncertainty from estimating its linear approximation. The price we pay in terms of precision is small relative to the gains in robustness to misspecification of the linear specification when identification is not weak and sample sizes are moderate or large.
6 Empirical Application

TO BE COMPLETED

7 Conclusions

TO BE COMPLETED

A Appendix

A.1 Theory for Semiparametric RDD

In this section we establish the asymptotic normality for the proposed semiparametric OLS estimator in the RDD case. For simplicity, we only consider the case where $X$ and $Z$ are univariate. The extension to the multivariate case is trivial. Let $\varepsilon_{V_i} = V_i - E[V_i|W_i, Z_i]$ denote the regression errors for $V = Y$ and $V = X$. We require the following regularity conditions, which parallel those of HTV:

**Assumption RRD:**

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. 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 $\mu_j(w) = E[Z_j|W = w], q_{Yj}(w) = E[Z^2_j \varepsilon^2_{Y_i}|W = w], q_{Xj}(w) = E[Z^2_j \varepsilon^2_{X_i}|W = w], r_j(w) = E[Z^2_j \varepsilon_{Y_i}\varepsilon_{X_i}|W = w], s_{Yj}(w) = E[Z^3_j \varepsilon^3_{Y_i}|W = w]$ and $s_{Xj}(w) = E[Z^3_j \varepsilon^3_{X_i}|W = w]$ 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 $h = \rho n^{-1/5}$.

The estimator can be written as

$$\hat{\beta} = \frac{\sum_{i=1}^n \delta_{X_i} \delta_{Y_i}}{\sum_{i=1}^n \delta^2_{X_i}}.$$
To investigate the asymptotic properties of $\hat{\beta}$, using $\delta Y_i = \beta_0 \delta X_i$ substitute $\hat{\delta} Y_i = \beta_0 \hat{\delta} X_i + \delta Y_i - \beta_0 (\hat{\delta} X_i - \delta X_i)$ in $\hat{\beta}$ to get

$$\hat{\beta} = \beta_0 + \sum_{i=1}^{n} \hat{\delta} X_i (\hat{\delta} Y_i - \delta Y_i) \sum_{i=1}^{n} \hat{\delta}^2 X_i - \beta_0 \frac{\sum_{i=1}^{n} \hat{\delta} X_i (\hat{\delta} X_i - \delta X_i).}{\sum_{i=1}^{n} \hat{\delta}^2 X_i}.$$ 

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

$$\hat{\beta} - \beta_0 = (E[\delta^2 X_i])^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \delta X_i (\delta Y_i - \delta Y_i) - \frac{1}{n} \sum_{i=1}^{n} \delta X_i (\hat{\delta} X_i - \delta X_i)' \beta_0 \right) + o_P((nh)^{-1/2}).$$

Introduce the notation $\hat{\delta}^j V = \hat{\alpha}^j V - \hat{\alpha}^j V$ and $\delta^j V = \alpha^j V - \alpha^j V$ for $j = 0, 1, 2$ and 3 and $V = Y$ and $X$. The local linear estimator for $\alpha^+ V$ solves the least squares problem

$$\min_{\alpha} \sum_{i=1}^{n} (V_i - \alpha' S_i)^2 k_h(W_i)1(W_i \geq 0),$$

for $V = Y$ and $X$, where $S_i = (1, Z_i', W_i, W_i Z_i')'$ and similarly the estimator for $\alpha^- V$ solves

$$\min_{\alpha} \sum_{i=1}^{n} (V_i - \alpha' S_i)^2 k_h(W_i)1(W_i < 0).$$

Henceforth we focus on the analysis for $\alpha^+ V$, since that of $\alpha^- V$ is symmetric.

Introduce the notation $f(0^+) = \lim_{w\to0} f(w)$,

$$V_i^+ = V_i - \alpha^+ V S_i, \quad k_{ih}^+ = k_h(W_i)1(W_i \geq 0)$$

and

$$S_{ih} = (1, Z_i, W_i/h, Z_i W_i/h)^'.$$

Then, with this notation we can write the OLS problem for $\alpha^+ V$ as

$$\min_{\alpha} \sum_{i=1}^{n} (V_i^+ - (\alpha - \alpha^+ V)' S_i)^2 k_{ih}^+,$$

whose first order condition yields

$$\begin{bmatrix} \hat{\alpha}_0^+ V - \alpha_0^+ V \\ \hat{\alpha}_1^+ V - \alpha_1^+ V \\ h (\hat{\alpha}_2^+ V - \alpha_2^+ V) \\ h (\hat{\alpha}_3^+ V - \alpha_3^+ V) \end{bmatrix} = \left( \sum_{i=1}^{n} S_{ih} S_{ih}' k_{ih}^+ \right)^{-1} \left( \sum_{i=1}^{n} S_{ih} V_i^+ k_{ih}^+ \right).$$

Based on this expression we analyze the asymptotic properties of the local linear estimators. All the Lemmas below make use of Assumption RDD.
Lemma A.1 (Denominator)

\[ \frac{1}{nh} \sum_{i=1}^{n} S_{ih} S'_{ih} k_{ih}^{+} \rightarrow f(0^{+}) \Gamma_d \]

where

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

\[ \gamma_l = \int_{0}^{\infty} u^l k(u)du \quad \text{and} \quad \mu_j^{+} = \lim_{w \downarrow 0} E[Z^j | W = w]. \]

**Proof.** Let

\[ \theta_{lj} = \frac{1}{nh} \sum_{i=1}^{n} \left( \frac{W_i}{h} \right)^l Z_i^j k_{ih}^{+}, \quad l, j = 0, 1, 2. \]

Then, by the change of variables \( u = w/h, \)

\[ E[\theta_{lj}] = h^{-1} E \left[ \left( \frac{W_i}{h} \right)^l Z_i^j k_{ih}^{+} \right] \]

\[ = \int_{0}^{\infty} u^l k(u) \mu_j(uh) f(uh) du \]

\[ = \mu_j^{+} f(0^{+}) \gamma_l + o(1), \]

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

\[ \text{Var}(\theta_{lj}) \leq (nh)^{-1} E \left[ \left( \frac{W_i}{h} \right)^{2l} Z_i^{2j} k_{ih}^{+2} \right] \]

\[ = (nh)^{-1} \int_{0}^{\infty} u^{2l} k^2(u) \mu_{2j}(uh) f(uh) du \]

\[ = o(1), \]

again by the Dominated Convergence theorem.

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

\[ \zeta_V(w, z) = \alpha_{0V}(w) + \alpha_{1V}(w) z - \alpha_{0V}^{+} + \alpha_{1V}^{+} z \]

\[ - (\dot{\alpha}_{0V}^{+} + \dot{\alpha}_{1V}^{+} z) w - \frac{1}{2} (\ddot{\alpha}_{0V}^{+} + \ddot{\alpha}_{1V}^{+} z) w^2, \]

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

\[ \sup_{0 < w < Mh} |\zeta_V(w, z)| = o(h^2)(1 + |Z|). \]
Lemma A.2 (Numerator: Expectation)

\[
E \left[ \frac{1}{nh} \sum_{i=1}^{n} S_{ih} V_i^+ k_{ih}^+ \right] \to_p \frac{1}{2} f(0^+) h^2 (b_V^+ + o(1)),
\]

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

\[
u_{lj} = \frac{1}{nh} \sum_{i=1}^{n} \left( \frac{W_i}{h} \right)^l Z_i^j V_i^+ k_{ih}^+ , \quad l, j = 0, 1.
\]

Then, write

\[
E[u_{lj}] = h^{-1} E \left[ \left( \frac{W_i}{h} \right)^l Z_i^j V_i^+ k_{ih}^+ \right]
= h^{-1} E \left[ \left( \frac{W_i}{h} \right)^l Z_i^j \left( \frac{1}{2} \ddot{\alpha}_{0V}(0^+) + \ddot{\alpha}_{1V}(0^+) Z_i \right) W_i^2 + \zeta_V(W_i, Z_i) \right] k_{ih}^+
= h^{-1} \frac{1}{2} \ddot{\alpha}_{0V}(0^+) E \left[ \left( \frac{W_i}{h} \right)^l Z_i^j W_i^2 k_{ih}^+ \right] + h^{-1} \frac{1}{2} \ddot{\alpha}_{1V}(0^+) E \left[ \left( \frac{W_i}{h} \right)^l Z_i^{j+1} W_i^2 k_{ih}^+ \right]
+ h^{-1} E \left[ \left( \frac{W_i}{h} \right)^l Z_i^j \zeta_V(W_i, Z_i) k_{ih}^+ \right].
\]

By the change of variables \(u = w/h\),

\[
h^{-1} E \left[ \left( \frac{W_i}{h} \right)^l Z_i^j W_i^2 k_{ih}^+ \right] = h^2 \int_{0}^{\infty} u^{l+2} k(u) \mu_j(uh) f(uh) du
= h^2 \mu_j^+ f(0^+) \gamma_{l+2} + o(1),
\]

and similarly

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

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

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

Lemma A.3 (Numerator: Conditional Expectation)

\[
\frac{1}{nh} \sum_{i=1}^{n} E[S_{ih} V_i^+ k_{ih}^+ | W_i, Z_i] = \frac{1}{nh} \sum_{i=1}^{n} E[S_{ih} V_i^+ k_{ih}^+] + o_p(h^2).
\]
Proof. We have

\[
\frac{1}{nh} \sum_{i=1}^{n} \mathbb{E}[S_{ih} V_i^+ k_{ih}^+ | W_i, Z_i] = \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ \left( \frac{1}{2} (\bar{a}_{0V}(0^+) + \bar{a}_{1V}(0^+)) Z_i \right) W_i^2 + \zeta_V(W_i, Z_i)
\]

\[
= \frac{1}{2} \bar{a}_{0V}(0^+) \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ W_i^2 + \frac{1}{2} \bar{a}_{1V}(0^+) \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ Z_i W_i^2
\]

\[
+ \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ \zeta_V(W_i, Z_i).
\]

Observe that

\[
\mathbb{V}ar \left( \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ W_i^2 \right) = (nh)^{-1} \mathbb{V}ar \left( S_{ih} k_{ih}^+ W_i^2 \right)
\]

\[
\leq C (nh)^{-1} h^{-1} \mathbb{E} \left[ S_{ih} S_{ih}^t k_{ih}^+ W_i^4 \right]
\]

\[
= O \left( (nh)^{-1} h^4 \right)
\]

\[
= o(1),
\]

since for \( l, j = 0, 1, 2 \)

\[
h^{-1} \mathbb{E} \left[ \left( \frac{W_i}{h} \right)^l Z_i^j k_{ih}^+ W_i^4 \right] = h^4 \int_0^{\infty} u^l k^2(u) \mu_j(uh) f(uh) du
\]

\[
= h^4 \mu_j^+ f(0^+) v_l + o(1),
\]

where

\[
v_l = \int_0^{\infty} u^l k^2(u) du.
\]

Similarly,

\[
\mathbb{V}ar \left( \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ Z_i W_i^2 \right) = o(1).
\]

and

\[
\mathbb{V}ar \left( \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ \zeta_V(W_i, Z_i) \right) = o(1).
\]

Note that

\[
\frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ \left( V_i^+ - \mathbb{E}[V_i^+ | W_i, Z_i] \right) = \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ \varepsilon_{Vi},
\]

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

Lemma A.4 (Numerator: Conditional Variance)

\[
\mathbb{V}ar \left( \frac{1}{nh} \sum_{i=1}^{n} S_{ih} k_{ih}^+ \left( V_i^+ - \mathbb{E}[V_i^+ | W_i, Z_i] \right) \right) = \frac{1}{nh^2} f(0^+) \Sigma_{V^+} + o(1),
\]

22
where

\[
\Sigma_{V^+} = \begin{bmatrix}
v_0 & v_0 q_1^+ & v_2 & v_2 q_1^+ \\
v_0 q_1^+ & v_0 q_2^+ & v_2 q_1^+ & v_2 q_2^+ \\
v_2 & v_2 q_1^+ & v_4 & v_4 q_1^+ \\
v_2 q_1^+ & v_2 q_2^+ & v_4 q_1^+ & v_4 q_2^+
\end{bmatrix},
\]

\[v_l = \int_0^\infty u^l k^2(u)du\] and \[q_j^+ = \lim_{w \to 0} E[Z^{2j} \varepsilon_Y^2 | W = w].\]

**Proof.** Consider the generic term, for \(l, j = 0, 1\)

\[
\frac{1}{nh} \sum_{i=1}^n \left( \frac{W_i}{h} \right)^l Z_i^j k^+_{ih} \varepsilon_{Y_i},
\]

and its variance, which equals

\[
(nh)^{-1} h^{-1} \mathbb{E} \left[ \left( \frac{W_i}{h} \right)^{2l} Z_i^{2j} k^+_{ih} \varepsilon_{Y_i}^2 \right] = (nh)^{-1} \int_0^\infty u^{2l} k^2(u) q_j(uh) f(uh)du
\]

\[= (nh)^{-1} f(0^+) q_j^+ v_{2l} + o(1)
\]

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

**Lemma A.5 (Numerator: Conditional Covariance)**

\[
\text{Cov} \left( \frac{1}{nh} \sum_{i=1}^n S_{ih} k^+_{ih} \varepsilon_{Y_i}, \frac{1}{nh} \sum_{i=1}^n S_{ih} k^+_{ih} \varepsilon_{X_i} \right) = \frac{1}{nh} f(0^+) \Sigma_{YX^+} + o(1),
\]

where

\[
\Sigma_{YX^+} = \begin{bmatrix}
v_0 & v_0 r_1^+ & v_2 & v_2 r_1^+ \\
v_0 r_1^+ & v_0 r_2^+ & v_2 r_1^+ & v_2 r_2^+ \\
v_2 & v_2 r_1^+ & v_4 & v_4 r_1^+ \\
v_2 r_1^+ & v_2 r_2^+ & v_4 r_1^+ & v_4 r_2^+
\end{bmatrix}
\]

and

\[r_j^+ = \lim_{w \to 0} \mathbb{E}[Z^{2j} \varepsilon_Y \varepsilon_X | W = w].\]

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

**Lemma A.6 (Numerator: Conditional CLT)**

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

**Proof.** Consider a generic term for \(l, j = 0, 1\)

\[
\frac{1}{\sqrt{nh}} \sum_{i=1}^n \left( \frac{W_i}{h} \right)^l Z_i^j k^+_{ih} \varepsilon_{Y_i}.
\]

23
We apply Lyapounov with third absolute moment. By the lemma on the asymptotic variance, we need to establish
\[(nh)^{-1/2} h^{-1} \mathbb{E} \left[ \left( \frac{W_i}{h} \right)^3 Z_i^3 \right] = o(1).\]

But note that, defining \(s_j(w) = \mathbb{E}[Z_i^3 | W = w]\),
\[h^{-1} \mathbb{E} \left[ \left( \frac{W_i}{h} \right)^3 Z_i^3 \right] = \int_0^\infty u^3 k^3(u) s_j(u) f(u) du = O(1).\]

Lemma A.7 (Numerator: Unconditional CLT)
\[(nh)^{-1/2} \sum_{i=1}^n \left( \frac{S_i h^+_{ik} Y_i^+}{S_i h^+_{ik} X_i^+} \right) - \frac{(nh)^{1/2} h^2}{2} f(0^+) \left( \begin{array}{c} b_Y^+ \\ b_X^+ \end{array} \right) \to_d f^{1/2}(0^+) \mathcal{N} \left( 0, \begin{bmatrix} \Sigma_{Y^+} & \Sigma_{Y^+ X^+} \\ \Sigma_{Y^+ X^+} & \Sigma_{X^+} \end{bmatrix} \right).\]

Proof. It follows from previous Lemmas.

Denote
\[\hat{\alpha}_h^+ - \alpha_h^+ = \begin{bmatrix} \hat{\alpha}_0^+ - \alpha_0^+ \\ \hat{\alpha}_1^+ - \alpha_1^+ \\ h(\hat{\alpha}_2^+ - \alpha_2^+) \\ \hat{\alpha}_3^+ - \alpha_3^+ \end{bmatrix}, \quad \hat{\alpha}_h^- - \alpha_h^- = \begin{bmatrix} \hat{\alpha}_0^- - \alpha_0^- \\ \hat{\alpha}_1^- - \alpha_1^- \\ h(\hat{\alpha}_2^- - \alpha_2^-) \\ \hat{\alpha}_3^- - \alpha_3^- \end{bmatrix}\]
and
\[\Sigma_+ = \begin{bmatrix} \Sigma_{Y^+} & \Sigma_{Y^+ X^+} \\ \Sigma_{Y^+ X^+} & \Sigma_{X^+} \end{bmatrix}, \quad \Sigma_- = \begin{bmatrix} \Sigma_{Y^-} & \Sigma_{Y^- X^-} \\ \Sigma_{Y^- X^-} & \Sigma_{X^-} \end{bmatrix},\]
where the definition of \(\Sigma_-\) is like \(\Sigma_+\) but with limits to the left of \(w = 0\).

Lemma A.8 (Joint CLT)
\[\sqrt{nh} \left( \hat{\alpha}_h^+ - \alpha_h^+ - \hat{\alpha}_h^- + \alpha_h^- \right) - \frac{(nh)^{1/2} h^2}{2} \begin{bmatrix} \Gamma^{-1}_d & 0 \\ 0 & \Gamma^{-1}_d \end{bmatrix} \begin{bmatrix} b_Y^+ - b_Y^- \\ b_X^+ - b_X^- \end{bmatrix} \to_d f^{-1/2}(0^+) \begin{bmatrix} \Gamma^{-1}_d & 0 \\ 0 & \Gamma^{-1}_d \end{bmatrix} \mathcal{N} \left( 0, \Sigma_+ + \Sigma_- \right).\]

Proof. It follows from previous Lemmas and the asymptotic independence of \((nh)^{-1/2} (\hat{\alpha}_h^+ - \alpha_h^+)\) and \((nh)^{-1/2} (\hat{\alpha}_h^- - \alpha_h^-)\).
Define the selection matrix

\[
B_2 = \begin{bmatrix}
1 & 0 & 0 & 0 & -\beta_0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & -\beta_0 & 0
\end{bmatrix},
\]

and note that

\[
B_2 (\hat{\delta}_h - \delta_h) = \begin{bmatrix}
\hat{\delta}_0Y - \delta_0Y \\
\hat{\delta}_1Y - \delta_1Y
\end{bmatrix} - \begin{bmatrix}
\hat{\delta}_0X - \delta_0X \\
\hat{\delta}_1X - \delta_1X
\end{bmatrix} \beta_0.
\]

Define

\[
\mu = \rho \begin{bmatrix}
\Gamma_0^{-1} & 0 \\
0 & \Gamma_1^{-1}
\end{bmatrix} \begin{bmatrix}
b_Y^+ - b_Y^- \\
b_X^+ - b_X^- 
\end{bmatrix}
\]

and

\[
\Omega = (\Sigma_+ + \Sigma_-).
\]

**Proof of the Theorem 4.1.** Write

\[
\sqrt{nh} (\hat{\beta} - \beta_0) = Q_2 \sqrt{nh} (\hat{\delta}_h - \delta_h) + o_P(1),
\]

and use the previous Lemma. ■

### A.2 Nonparametric Estimators

#### A.2.1 Binary Instrument Case

We first introduce some notation that will be used throughout this Section. Henceforth, \(A', \text{ rank}(A), A^-, \text{tr}(A) \) and \(|A| := (\text{tr}(A'A))^1/2\) denote the transpose, rank, Moore-Penrose generalized inverse, trace and the Euclidean norm of a matrix \(A\), respectively. The symbol := denotes definitional relation. For generic random vectors \(\zeta\) and \(\xi\), let \(F_{\zeta}\) and \(F_{\zeta/\xi}\) be the cumulative distribution function (cdf) and conditional cdf of \(\zeta\) and \(\zeta\) given \(\xi\), respectively. Denote the corresponding densities with respect to a \(\sigma\)-finite measure \(\mu(x)\) by \(f_{\zeta}\) and \(f_{\zeta/\xi}\). Unless otherwise stated, the underlying measure will be the Lebesgue measure. Let \(S_{\zeta}\) denote the support of \(\zeta\). 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 \(\|g\|^2 := \langle g, g \rangle\). 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 \(R(K) := \{f \in L_2(Y) : \exists s \in L_2(X), Ks = f\}\) and \(N(K) := \{f \in L_2(X) : Kf = 0\}\). Let \(D(K)\) 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) and Chen and Pouzo (2012), among others. 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.
We assume the sieve space $S_X$ with 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 $S_X$ is bounded interval of $\mathbb{R}$, with non-empty interior. For any smooth function $h : S_X \subset \mathbb{R} \to \mathbb{R}$ and some $r > 0$, let $[r]$ be the largest integer smaller than $r$, and

$$
\|h\|_{\infty, r} := \max_{1 \leq j \leq r} \sup_{x \in S_X} |\nabla^j h(x)| + \sup_{x \neq x'} \frac{|\nabla^r h(x) - \nabla^r h(x')|}{|x - x'|^{-[r]}}.
$$

Further, let $C^r(S_X)$ 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 $C^r(S_X)$. We shall assume that $g \in C^r(S_X)$ for some $r$ and approximate $C^r(S_X)$ with a sieve space $G_n$ satisfying some conditions below. Define $k_n = \dim G_n$. Given an integer $s > 0$ define the Sobolev norm $\|h\|_s := \sum_{i=0}^s \|h^{(i)}\|_2$, where $h^{(i)}(x) := \partial_i h(x)/\partial x^i$, with $h^{(0)} = h$.

We approximate $m(z) \equiv m(z, t)$ by the function $\hat{m}(z, t) : = \sum_{j \in J_n} m_{tj}p_{0j}(z, t)$, where $p_{0j}$ are some known basis functions and $J_n := \#(J_n) \to \infty$ as $n \to \infty$. We write $\hat{m}(z, t) = p^{J_n}(z, t)'A$, where $p^{J_n}(z, t) = (p_{01}(z, t), ..., p_{0J_n}(z, t))'$ and $A = (m_{t1}, ..., m_{tJ_n})$. Define $P := (p^{J_n}(Z_1, T_1), ..., p^{J_n}(Z_n, T_n))'$. Then, the SLS is

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

More precisely, we take $p^{J_n}(z, t) = (B^{J_2n}(z), t \times B^{J_2n}(z))$, where $B^{J_2n}(z)$ is a $J_2n \times 1$ vector of univariate B-splines or polynomial splines and $J_n = 2J_2n$. We define $\hat{m}(z) := \hat{m}(z, 1) - \hat{m}(z, 0)$.

Similarly, for a fixed $g$, we consider the sieve estimator of $Ag$ as $\hat{A}g = \hat{A}_1g - \hat{A}_0g$, where

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

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

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

We assume the sieve space $G_n$ is of the form

$$
G_n = \{ g_n : S_X \to \mathbb{R}, \sup_x |g_n(x)| < c, \sup_x \left| \nabla^r g_n(x) \right| < c \}
$$

$$
g_n(x) = \psi^{k_n}(x)'\Pi, \quad g_n(x_0) = 0 \}
$$

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 $\Pi$ is a $k_n \times 1$ vector of coefficients to be estimated. In the simulations and the application below we use B-splines for $\psi^{k_n}$. Blundell, Chen and Kristensen (2007) discussed practical ways to incorporate the
constraints into the computation of \( \hat{g}_n \). 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 \( S_X \).

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

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

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

**Assumption 3**: The data \( \{(Y_i, X_i, Z_i, T_i)\}_{i=1}^n \) are iid and Assumption 1 holds.

**Assumption 4**: (i) \( g \in \mathcal{C}^r(S_X) \) for \( r > \frac{1}{2} \); (ii) \( \mathbb{E}[|X|^{2a}] < \infty \) for some \( a > r \).

**Assumption 5**: For any \( t = 1, 2, m_t \in \mathcal{C}^{r_m}(S_Z) \) with \( r_m > 1/2 \) and \( \mathbb{E}[g_n(X)|Z = \cdot, T = t] \in \mathcal{C}^{r_m}(S_Z) \) for any \( g_n \in \mathcal{G}_n \).

**Assumption 6**: (i) The smallest and the largest eigenvalues of \( \mathbb{E}[B_{J_{2n}}^T(Z) \times B_{J_{2n}}(Z)'] \) 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 \( S_Z \), which is a compact interval with non-empty interior.

**Assumption 7**: (i) \( k_n \to \infty, J_{2n}/n \to 0 \); (ii) \( \lim_{n \to \infty} (J_{2n}/k_n) = c_0 > 1 \) and \( \lim_{n \to \infty} (k_n^2/n) = 0 \).

**Assumption 8**: 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 \( \|\hat{g}_n - g\| \). Its proof is essentially the same as that of Theorem 2 in Blundell, Chen and Kristensen (2007), hence it is omitted.

**Theorem A.9** Let Assumptions 3-8 hold. Then,

\[
\|\hat{g}_n - g\| = O_P \left( k_n^{-r} + \tau_n \times {\sqrt{k_n \over n}} \right).
\]

**A.2.2 RDD Case**

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

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

where now \( \hat{m} \) and \( \hat{A} \) 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 \( \hat{m}(z) := \hat{m}_+(z) - \hat{m}_-(z) \) and \( \hat{A}g = \hat{A}_+g - \hat{A}_-g \), where \( \hat{m}_+(z) = \hat{a} \) in the solution to the least squares problem

\[
(\hat{a}, \hat{b}_0, \hat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} (Y_i - a_0 - b_0W - b_1(Z_i - z))^2 k_h(W_i)k_h(Z_i - z)1(W \geq 0),
\]

where \( k_h(z_i) = h^{-1}k(z_i/h) \), \( k(\cdot) \) is a kernel function, and \( h \) denotes a bandwidth parameter satisfying regularity conditions described below. Similarly, \( \hat{m}_-(z) = \hat{a} \) in

\[
(\hat{a}, \hat{b}_0, \hat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} (Y_i - a_0 - b_0W - b_1(Z_i - z))^2 k_h(W_i)k_h(Z_i - z)1(W < 0),
\]

\( \hat{A}_+g = \hat{a} \) in

\[
(\hat{a}, \hat{b}_0, \hat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} (g(X_i) - a_0 - b_0W - b_1(Z_i - z))^2 k_h(W_i)k_h(Z_i - z)1(W \geq 0),
\]

and \( \hat{A}_-g = \hat{a} \) in

\[
(\hat{a}, \hat{b}_0, \hat{b}_1) = \arg \min_{a_0, b_0, b_1} \sum_{i=1}^{n} (g(X_i) - a_0 - b_0W - b_1(Z_i - z))^2 k_h(W_i)k_h(Z_i - z)1(W < 0).
\]

We can follow Theorem 2 in Blundell, Chen and Kristensen (2007) to obtain rates for \( \hat{g}_n \). In fact, from the proof of Theorem 2 in Blundell, Chen and Kristensen (2007) we obtain the same rates for \( \hat{g}_n \) in the RRD case as in the binary IV case,

\[
\|\hat{g}_n - g\| = O_P \left( k_n^{-r} + \tau_n \times \sqrt{\frac{k_n}{n}} \right),
\]

provided we show that

\[
\|\hat{m} - m\| = O_P \left( \sqrt{\frac{k_n}{n}} \right)
\]

and

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

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

\[
\|\hat{m} - m\| = O_P (d_n),
\]

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

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

Therefore, we require rates on the bandwidth \( h \) so that

\[
d_n = O_P \left( \sqrt{\frac{k_n}{n}} \right).
\]
This provides some flexibility in how to choose the bandwidth $h$. 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 $h_n \equiv h$ to obtain the desired rates for $\hat{g}_n - g$ under different scenarios on the rate for the measure of ill-posedness.

References


