KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No.910

"The Effect of Measurement Error
in the Sharp Regression Discontinuity Design"

Takahide Yanagi

December 2014

KYOTO UNIVERSITY

KYOTO, JAPAN

The Effect of Measurement Error in the Sharp Regression Discontinuity Design*

Takahide Yanagi†

December 15, 2014

Abstract

This paper develops a nonparametric analysis for the sharp regression discontinuity (RD) design in which the continuous forcing variable may contain measurement error. We show that if the observable forcing variable contains measurement error, this error causes severe identification bias for the average treatment effect given the "true" forcing variable at the discontinuity point. The bias is critical in the sense that even if there is a significant causal effect, researchers are misled to the incorrect conclusion of no causal effect. Furthermore, the measurement error leads the conditional probability of the treatment to be continuous at the threshold. To investigate the average treatment effect using the mismeasured forcing variable, we propose an approximation using the small error variance approximation (SEVA) originally developed by Chesher (1991). Based on the SEVA, the average treatment effect is approximated up to the order of the variance of the measurement error using an identified parameter when the variance is small. We also develop an estimation procedure for the parameter that approximates the average treatment effect based on local polynomial regressions and the kernel density estimation. Monte Carlo simulations reveal the severity of the identification bias caused by the measurement error and demonstrate that our approximate analysis is successful.

Keywords: Regression discontinuity designs; classical measurement error; approximation; nonparametric methods; local polynomial regressions

JEL Classification: C13; C14; C21

*The author gratefully appreciates Yoshihiko Nishiyama, Ryo Okui, Kohtaro Hitomi, Hisaki Kono, and Naoya Sueishi for their helpful comments and discussions. The author would also like to thank seminar participants at Kyoto University. The author acknowledges financial support from the JSPS Grant-in-Aid for JSPS Fellows No.252035. All remaining errors are mine.

†Graduate School of Economics, Kyoto University, Yoshida-Hommachi, Sakyo, Kyoto, Kyoto, 606-8501, Japan; Research Fellow of Japan Society for the Promotion of Science. Email: yanagi.takahide.87w@st.kyoto-u.ac.jp

1 Introduction

This paper develops a nonparametric analysis for the sharp regression discontinuity (RD) design in which the continuous forcing variable may contain classical measurement error. We show that the measurement error causes severe bias for identifying the average treatment effect given the "true" forcing variable at the discontinuity point. We then examine to what extent the average treatment effect can be studied from observed data containing the measurement error. The average treatment effect is approximated using an identified parameter based on the small error variance approximation (SEVA) originally proposed by Chesher (1991). We also develop an estimation procedure for the approximating parameter based on local polynomial regressions and the kernel density estimation.

The RD design was first introduced by Thistlethwaite and Campbell (1960) and has been substantially studied in theoretical econometrics. It is known as a quasi-experimental design, which is a powerful design for treatment effect analyses and program evaluation. Examples of theoretical studies include research on identification (Hahn, Todd, and van der Klaauw, 2001; Lee, 2008; Frandsen, Frölich, and Melly, 2012), estimation (Porter, 2003; Imbens and Kalyanaraman, 2012; Arai and Ichimura, 2014), and inference methods (Lee and Card, 2008; McCrary, 2008; Calonico, Cattaneo, and Titiunik, 2014). In addition, much of the empirical literature has developed analyses based on RD designs because of its utility. For example, many studies have been conducted for education (e.g., Angrist and Lavy, 1999) and health economics (e.g., Card and Shore-Sheppard, 2004). Useful surveys on the RD literature can be seen in Imbens and Lemieux (2008) and Lee and Lemieux (2010).

Despite the vast body of RD literature, studies on RD designs with measurement error are scant (see the paragraph "Related literature" below). In RD designs, a treatment is completely or partly determined by whether a forcing variable[1] is greater than a known threshold. In the sharp RD design in which the treatment is completely determined by the forcing variable, the average treatment effect at the threshold is identified by the difference in the means of the outcome marginally above and below the threshold (Hahn et al., 2001). See Figure 1 for an intuitive understanding of this. However, if the observed forcing variable contains measurement error, we cannot observe the "true" forcing variable that determines the treatment. Identification analyses for the average treatment effect with the observed mismeasured forcing variable have not been developed enough.

There are many empirical situations based on RD designs in which the observed forcing variable may be mismeasured. Empirical studies based on RD designs often use survey data. For example, Hullegie and Klein (2010) analyze the effect of private insurance coverage on individual health performance (e.g., the number of doctor visits) using a unique public insurance system in Germany. In Germany, employees whose income is below a threshold cannot buy private insurance, so that this unique system provides an RD design. Their forcing variable is

[1] It is also referred to as the "assignment" or "running" variable in the literature.

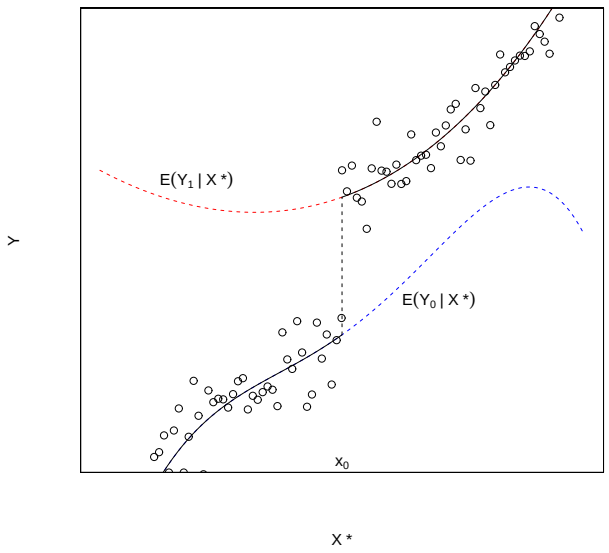


Figure 1: The dots indicate pairs of the observed outcome Y_i and the "true" forcing variable X_i^*. The black solid line is the conditional mean of Y_i given X_i^*, which is identified if (Y_i, X_i^*) is observed. The upper dotted red and lower dotted blue lines are the conditional means for the treated and untreated, respectively. The length of the dotted vertical line indicates the average treatment effect at the threshold x_0, $E(Y_{1i} - Y_{0i}|X_i^* = x_0)$, which is identified by the difference in the conditional means of Y_i marginally above and below x_0.

individual income, which is found using data from the German Socio-Economic Panel. They also indicate that their forcing variable seems to contain measurement error, because some people buy private insurance despite that their income is below the threshold (i.e., despite their supposed ineligibility to buy private insurance). There are many other applied studies based on RD designs that use survey data to conduct causal analyses, such as Card, Dobkin, and Maestas (2008), Battistin, Brugiavini, Rettore, and Weber (2009), Schanzenbach (2009), and Koch (2013). In such situations, there is always a risk that the forcing variable may be mismeasured, as in other literature in econometrics (see Bound, Brown, and Mathiowetz, 2001 and Schennach, 2013 for surveys on the literature of measurement error in econometrics).

This study first investigates the effect of the forcing variable with classical measurement error in the sharp RD design. We demonstrate that the difference in the conditional means of the outcome given the mismeasured forcing variable marginally above and below the threshold has an identification bias for the average treatment effect given the true forcing variable at the threshold. The identification bias is critical in the sense that even if there is a significant treatment effect, the bias misleads the researchers to the incorrect result of no treatment effect. Furthermore, the measurement error leads the conditional probability of the treatment to be continuous at the threshold. We derive the specific form of the identification bias caused by the measurement error.

To examine the average treatment effect using the mismeasured forcing variable, we then

suggest approximating it based on the SEVA originally developed by Chesher (1991). The accuracy of the approximation depends on the magnitude of the variance of the measurement error, σ^2. We show that the average treatment effect is approximated up to the order $O(\sigma^2)$ based on the SEVA when σ is small. In other words, the smaller standard deviation of the measurement error implies a more precise approximation for the average treatment effect. We thus consider that our approximate analysis is appropriate for empirical studies based on survey data in which the forcing variable may contain measurement error caused from incorrect entry or a memory lapse. Such measurement error can be classical and the variance can be small. Importantly, while σ^2 is generally unknown, σ^2 can be extrapolated or forecast. Additional data for the true forcing variable, such as public data or census data, allow us to conduct an extrapolation under the classical measurement error assumption. Importantly, our approach does not require additional variables such as instruments or repeated measurements, which are often unavailable in empirical situations.

We also provide a nonparametric estimation procedure for the parameter that approximates the average treatment effect based on local polynomial regressions and the kernel density estimation. We derive the consistency and asymptotic normality of the nonparametric estimator. Combining the asymptotic properties with our approximate analysis, the average treatment effect is approximately estimated up to the order $O(\sigma^2)$.

We conduct Monte Carlo simulations to investigate the practical effects of the measurement error on identification for the average treatment effect. The simulations also demonstrate the performance of our approximate analysis. We find that the measurement error critically affects identification of the average treatment effect in our simulation designs. The results of the Monte Carlo simulations also corroborate that our approximate analysis can function even when the magnitude of σ^2 accounts for 20 percent of the variance of a mismeasured forcing variable.

Related literature: As a study in RD literature, this paper is closely related to Battistin et al. (2009), Hullegie and Klein (2010), and Yu (2012).

Battistin et al. (2009) develop a fuzzy RD analysis in which the forcing variable contains measurement error. Their analysis is based on the non-differential measurement error and certain smoothness conditions on the joint distribution of the true and mismeasured forcing variables. They show that under these conditions, the average treatment effect on the treated is identified based on the mismeasured forcing variable using the fuzzy RD estimand, that is, the Wald estimand. However, their analysis does not function under continuous measurement error, because the conditional probability of the treatment given the mismeasured forcing variable is not discontinuous owing to the continuous measurement error (see Remark 3). Indeed, our Monte Carlo simulations demonstrate that the Wald estimator is highly unstable, because the measurement error leads the conditional probability of the treatment to be continuous.

Hullegie and Klein (2010) develop a fuzzy RD analysis with a continuous mismeasured forcing variable. Their analysis is based on the linear functional-form specifications and the

normally distributed measurement error independent of the observable forcing variable (such measurement error is not classical). They use the parametric specifications to identify and estimate a local average treatment effect in an RD design. In contrast, the present study focuses on a sharp RD design with classical measurement error in a nonparametric manner.

Yu (2012) studies RD designs with an observable continuous forcing variable containing classical measurement error. Although the model studied in the present paper is similar to that in his paper,[2] the approaches developed in both papers differ. He focuses on the conditions under which the average treatment effect at the threshold can be consistently estimated based on the difference in the mean outcomes given the mismeasured forcing variable and treatment at the threshold (see Remark 4). He shows that if the measurement error shrinks to zero depending on the sample size under some rate conditions, a local polynomial estimator for the difference is consistent for the average treatment effect. By contrast, this paper first approximates the average treatment effect in the population, which is based on the SEVA, and then develops an estimation for the approximating parameter. As a result, the identification and estimation approaches in both papers differ. For a better understanding, we compare the performance of the analyses developed in both papers using our Monte Carlo simulations, which reveal that our approximate analysis is more successful.

The present paper is also related to Pei (2011) and Dong (2014), but the objectives in those studies differ from that in our paper. Pei (2011) studies a model in which both a true forcing variable and measurement error are discretely distributed with bounded support. He develops an identification analysis for the average treatment effect utilizing the discreteness of the forcing variable. Dong (2014) studies RD designs in which a true unobservable forcing variable is continuous but the observed forcing variable is discretized or rounded, such as age in years. She provides modified identification and estimation procedures for the average treatment effect based on parametric polynomial modeling. As she mentions, such a rounding error cannot be classical. In contrast, the present paper develops a nonparametric analysis for the problem of classical measurement error in a sharp RD design with a continuous forcing variable.

As a study in the literature of measurement error, the present paper is related to Chesher (1991), Chesher and Schluter (2002), and Battistin and Chesher (2014). These papers use the SEVA to investigate the effects of measurement error in separate settings. However, to our knowledge, no study applies the SEVA to the problem of measurement error in RD designs. Our study builds on the literature by showing the conditions under which the average treatment effect in the RD design is approximated based on the SEVA.

Organization of rest of the paper: Section 2 introduces our setting and the parameter of interest. Section 3 examines the effect of the measurement error for identifying the average treatment effect. Section 4 develops our approximate analysis based on the SEVA. Section 5

[2]More specifically, the setting considered in the present paper is almost the same as "Case 2" in Yu (2012). He also considers other settings in which the treatment is determined using the mismeasured forcing variable and/or in which the treatment is unobservable.

develops an estimation method for the approximating parameter. Section 6 describes the Monte Carlo simulations. Section 7 concludes. All proofs are provided in the Appendix.

Notations: For generic random variables Z_i and W_i, we denote the conditional density (distribution function) of Z_i given $W_i = w$ as $f_{Z|W}(\cdot|w)$ ($F_{Z|W}(\cdot|w)$). We denote the support of Z_i as $supp(Z)$. For a generic function $g(\cdot)$, we denote the left and right limits of $g(x)$ at x_0 as $g(x_0-) \coloneqq \lim_{e\uparrow 0} g(x_0+e)$ and $g(x_0+) \coloneqq \lim_{e\downarrow 0} g(x_0+e)$, respectively. For $s \in \mathbb{N}$, we denote the s-th order (partial) derivative of $g(x)$ with respect to x as $g^{(s)}(x)$. The indicator function $\mathbf{1}(E)$ is 1 if the event E is true and 0 otherwise. We denote a $K \times L$ matrix B with the (k,l) entry $b_{k,l}$ as $B = (b_{k,l})_{(k,l)}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, L$.

2 Settings

We observe independent and identically distributed (i.i.d.) random variables $\{(Y_i, D_i, X_i)\}_{i=1}^n$, where $Y_i \in supp(Y) \subset \mathbb{R}$ is an outcome, $D_i \in \{0,1\}$ is a treatment that depends on an unobservable "true" continuous forcing variable $X_i^* \in supp(X^*) \subset \mathbb{R}$, and $X_i \in supp(X) \subset \mathbb{R}$ is the observable continuous forcing variable that may contain measurement error. If unit i is treated, $D_i = 1$, otherwise, $D_i = 0$. We can write $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$, where Y_{1i} is the potential outcome when unit i is treated and Y_{0i} is that when untreated. Both Y_{1i} and Y_{0i} cannot be observed for any unit, because no units can be both treated and untreated. This is the standard potential outcome notation.

Suppose that D_i is completely determined using X_i^* as follows:

$$D_i = \mathbf{1}(X_i^* \geq x_0), \tag{1}$$

where $x_0 \in supp(X^*)$ is a known fixed threshold. The relationship is commonly referred as the sharp RD design in RD literature (see Lee and Lemieux, 2010). Equation (1) means that all units with $X_i^* \geq x_0$ are treated and all units with $X_i^* < x_0$ are untreated. Here, $E(D_i|X_i^* = x) = \mathbf{1}(x \geq x_0)$ is the deterministic function of x, and it is discontinuous at x_0.

If we can observe the true forcing variable X_i^*, the average treatment effect given $X_i^* = x_0$ is identified under the continuity of $E(Y_{0i}|X_i^* = x)$ (Hahn et al., 2001):

$$E(Y_{1i} - Y_{0i}|X_i^* = x_0) = \tau^*, \tag{2}$$

where

$$\tau^* \coloneqq E(Y_i|X_i^* = x_0+) - E(Y_i|X_i^* = x_0-) \tag{3}$$

$$= E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0-, D_i = 0). \tag{4}$$

The left-hand side of (2) is the average treatment effect at the threshold, which is the common parameter of interest in the sharp RD design. τ^* is the difference in the conditional means of Y_i given X_i^* (and D_i) at the threshold. The right-hand side of (3) is equal to that of (4) because

of (1). If (Y_i, X_i^*) is observed, the right-hand sides of (3) and (4) are identified, so the average treatment effect is also identified.

Because of the presence of measurement error, we cannot observe the true forcing variable X_i^*, so the right-hand sides of (3) and (4) cannot be identified. We instead observe X_i that contains measurement error, as follows:

$$X_i = X_i^* + \sigma U_i, \tag{5}$$

where the random variable σU_i is continuous measurement error with $E(U_i) = 0$ and $var(U_i) = 1$, and $\sigma \geq 0$ indicates the standard deviation. This additive representation is commonly used in the literature of measurement error (see, e.g., Schennach, 2013). We assume $x_0 \in supp(X)$.

We introduce the following assumptions for the measurement error.

Assumption 1. *(i)* U_i *is independent of* $(Y_{1i}, Y_{0i}, D_i, X_i^*)$. *(ii)* $f_U(\cdot)$ *is continuous on bounded support. (iii)* $E(U_i) = 0, var(U_i) = 1, E|U_i|^3 < \infty$.

Assumption 1 (i) is the classical measurement error assumption (see Bound et al., 2001 and Schennach, 2013 for the interpretation). This assumption requires joint independence between the measurement error and the other variables. Assumption 1 (i) is identical to the independence between U_i and (Y_{1i}, Y_{0i}, X_i^*) in the sharp RD design, because D_i is the deterministic function of X_i^*. Assumption 1 (ii) ensures the continuity of U_i with bounded support. The bounded support is required to guarantee the establishment of the approximation developed in Section 4. This condition may be restrictive in the theoretical view, but it can be satisfied in many empirical situations. Assumption 1 (iii) is a mild moment condition. The existence of the third-order moment is unrestrictive under the bounded support of U_i.

Remark 1. We can allow the outcome to contain measurement error. In other words, we can allow a situation in which we observe not the "true" outcome Y_i^* but the mismeasured outcome Y_i. The analysis in this paper remains unchanged if the measurement error contained in Y_i is independent of $(Y_{1i}, Y_{0i}, D_i, X_i^*, U_i)$ and has mean zero. Thus, we do not explicitly consider that the outcome is mismeasured in this paper.

Remark 2. We do not allow a situation in which D_i is also mismeasured, which would require other approaches to analyze the problem caused by the measurement error.

3 Identification bias caused by measurement error

This section investigates the effect of the measurement error for identifying the average treatment effect at the threshold. We show that the measurement error leads the difference in the mean outcomes just above and below the threshold and the discontinuity of the conditional probability of the treatment to vanish. We then discuss possible approaches to examine the average treatment effect using the mismeasured forcing variable.

Although the parameter of interest in the RD design is the average treatment effect at the threshold, we focus on studying the effect of the measurement error for identifying τ^*. This is because τ^* equals the average treatment effect under the continuity of $E(Y_{0i}|X_i^*)$, as discussed in the previous section.

Observing (Y_i, D_i, X_i), it is not uncommon to consider the following parameters:

$$\tau_X \coloneqq E(Y_i|X_i = x_0+) - E(Y_i|X_i = x_0-),$$
$$\tau_{XD} \coloneqq E(Y_i|X_i = x_0+, D_i = 1) - E(Y_i|X_i = x_0-, D_i = 0).$$

τ_X replaces the true forcing variable X_i^* in (3) with the mismeasured forcing variable X_i. Similarly, τ_{XD} replaces X_i^* in (4) with X_i. τ_X and τ_{XD} are identified using the observable data. Importantly, τ_X and τ_{XD} generally differ, because D_i is not generally a deterministic function of X_i, that is, $D_i \neq \mathbf{1}(X_i \geq x_0)$.

We can guess that τ_X has a severe identification bias for τ^* and that τ_X is close to zero. We observe that $E(Y_i|X_i = x_0+)$ is computed based on a subset of units with $X_i \geq x_0$ (i.e., the right half of Figure 2). Units with $X_i \geq x_0$ but $X_i^* < x_0$ may lead $E(Y_i|X_i = x_0+)$ to substantially differ from $E(Y_i|X_i^* = x_0+)$, because the conditional distribution of Y_i can be discontinuous at $X_i^* = x_0$ owing to the RD structure, that is, because the realization of Y_i for $X_i^* < x_0$ can differ from those for $X_i^* \geq x_0$ (see Figure 1). Similarly, $E(Y_i|X_i = x_0-)$ is computed based on a subset with the remaining units (i.e., the left half of Figure 2) and it can substantially differ from $E(Y_i|X_i^* = x_0-)$ by the influence of the units with $X_i < x_0$ but $X_i^* \geq x_0$. As a result, τ_X could have a severe bias for identifying τ^*. As demonstrated in a later theorem, τ_X is equal to zero because of the bias.

In contrast, we can guess that τ_{XD} does not substantially differ from τ^*. We observe that $E(Y_i|X_i = x_0+, D_i = 1)$ is computed based on a subset of units with $X_i \geq x_0$ and $X_i^* \geq x_0$ (i.e., the upper right in Figure 2). Because $E(Y_i|X_i = x_0+, D_i = 1)$ is the conditional mean for a subset of units with $X_i^* \geq x_0$, unlike $E(Y_i|X_i = x_0+)$, it is not affected by the units with $X_i \geq x_0$ and $X_i^* < x_0$. Similarly, $E(Y_i|X_i = x_0-, D_i = 0)$ is computed based on a subset of units with $X_i < x_0$ and $X_i^* < x_0$ (i.e., the lower left in Figure 2). Thus, $E(Y_i|X_i = x_0+, D_i = 1)$ and $E(Y_i|X_i = x_0-, D_i = 0)$ may not substantially differ from $E(Y_i|X_i^* = x_0+)$ and $E(Y_i|X_i^* = x_0-)$, respectively. Then, the identification bias of τ_{XD} may be smaller than that of τ_X for τ^*.

Nonetheless, both τ_X and τ_{XD} have an identification bias for τ^* because of the measurement error. To evaluate the identification biases, we introduce the following assumption, which ensures the existence of limits and the use of the dominated convergence theorem. Let $m(x^*) \coloneqq E(Y_{0i}|X_i^* = x^*) + \mathbf{1}(x^* \geq x_0)(E(Y_{1i} - Y_{0i}|X_i^* = x^*) - \tau^*)$.

Assumption 2. *(i)* $E(Y_i|X_i^* = x_0+)$, $E(Y_i|X_i^* = x_0-)$, $E(Y_i|X_i = x_0+)$, $E(Y_i|X_i = x_0-)$, $E(Y_i|X_i = x_0+, D_i = 1)$, *and* $E(Y_i|X_i = x_0-, D_i = 0)$ *exist. (ii)* $f_{X^*|X}(x^*|x_0+)$ *and* $f_{X^*|X}(x^*|x_0-)$ *exist for any* $x^* \in \mathbb{R}$. *(iii)* $1 - F_{X^*|X}(x_0|x_0+)$ *and* $F_{X^*|X}(x_0|x_0-)$ *exist and are non-zero. (iv)* $E(Y_{1i}|X_i^* = x^*)f_{X^*|X}(x^*|x)$, $E(Y_{0i}|X_i^* = x^*)f_{X^*|X}(x^*|x)$, *and* $m(x^*)f_{X^*|X}(x^*|x)$ *are dominated by some integrable functions in* $x^* \in \mathbb{R}$ *for* x *near* x_0.

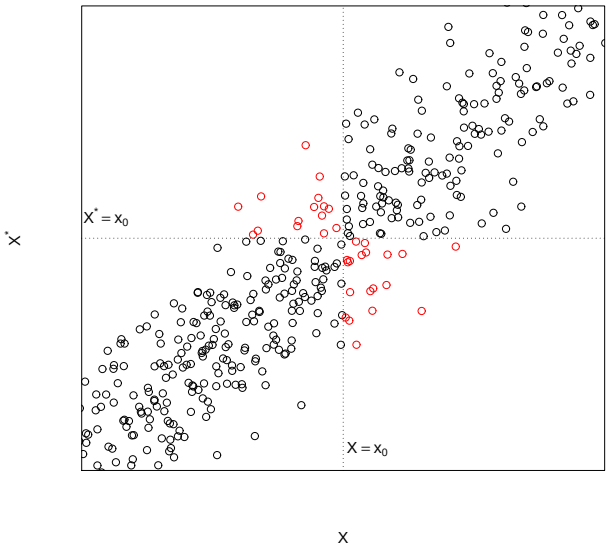


Figure 2: The dots are the pairs of (X_i, X_i^*) generated using the data-generating process in Monte Carlo simulations (Section 6). The horizontal and vertical axes are X_i and X_i^*, respectively. The dotted lines indicate the discontinuity point. $E(Y_i|X_i = x_0+)$ and $E(Y_i|X_i = x_0-)$ are based on the units in the right and left halves of the graph, respectively. By contrast, $E(Y_i|X_i = x_0+, D_i = 1)$ and $E(Y_i|X_i = x_0-, D_i = 0)$ are based on the units in the upper right and lower left, respectively. The units in the upper left and lower right (the red dots) severely affect identification for the average treatment effect.

The following theorem shows the specific form of the identification biases of τ_X and τ_{XD} for τ^*. The result (ii) in the theorem is shown in Yu (2012).

Theorem 1. *Suppose that Assumptions 1 and 2 hold and* $\sigma > 0$.

(i) It holds that

$$\begin{aligned}\tau_X = &\int_{x_0}^{\infty} E(Y_{1i}|X_i^* = x^*)\left(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)\right)dx^* \\ &+ \int_{-\infty}^{x_0} E(Y_{0i}|X_i^* = x^*)\left(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)\right)dx^* \\ &= 0.\end{aligned} \tag{6}$$

(ii) (Yu, 2012) It holds that

$$\tau_{XD} = \tau^* + \left(\frac{\int_{x_0}^{\infty} m(x^*)f_{X^*|X}(x^*|x_0+)dx^*}{1 - F_{X^*|X}(x_0|x_0+)} - \frac{\int_{-\infty}^{x_0} m(x^*)f_{X^*|X}(x^*|x_0-)dx^*}{F_{X^*|X}(x_0|x_0-)}\right). \tag{7}$$

The theorem demonstrates that τ_X and τ_{XD} have identification biases for τ^* and the average treatment effect at the threshold. As a result, we cannot precisely evaluate the causal effect of interest based on τ_X or τ_{XD}.

Importantly, the right-hand side of (6) vanishes because of the continuity of $f_{X^*|X}(x^*|x)$ at $x = x_0$, which is implied by the continuity of $f_U(\cdot)$ as shown in the proof of the theorem. In

other words, τ_X becomes zero because of the effect of the continuous measurement error. This result is remarkable, because this implies that even if there is a substantial causal effect, τ_X misleads researchers into the incorrect conclusion in which there is no causal effect. Indeed, our Monte Carlo simulations demonstrate that an estimator for τ_X is significantly close to zero.

The second term on the right-hand side of (7) is the identification bias of τ_{XD} for τ^*. The identification bias does not vanish in general. For example, this identification bias does not vanish even when the treatment effect is constant, that is, when $Y_{1i} - Y_{0i} = b$ for any i and a constant b. In addition, the bias cannot be nonparametrically identified, because it relates the joint distribution of $(Y_{1i}, Y_{0i}, X_i^*, X_i)$ including unobservables (Y_{1i}, Y_{0i}, X_i^*).

Remark 3. Battistin et al. (2009) show that in the fuzzy RD design with a mismeasured forcing variable, the average treatment effect for the treated is identified using the Wald estimand:

$$\frac{E(Y_i|X_i = x_0+) - E(Y_i|X_i = x_0-)}{E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-)}. \tag{8}$$

under the non-differential measurement error assumption and certain smoothness conditions on the joint distribution of (X_i^*, X_i). However, in a sharp RD design with continuous measurement error, their analysis could not work. To understand this, we observe the conditional mean of D_i (without the classical assumption):

$$E(D_i|X_i = x) = E(\mathbf{1}(X_i^* \geq x_0)|X_i = x) = \int_{x_0}^{\infty} f_{X^*|X}(x^*|x)dx^*.$$

Hence, the difference in the conditional means is

$$E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-) = \int_{x_0}^{\infty} \left(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-)\right) dx^*.$$

This difference vanishes under the continuity of $f_{X^*|X}(x^*|x)$ at $x = x_0$. That is, the discontinuity of the conditional probability is smoothed out because of the continuous measurement error: $E(D_i|X_i = x)$ is not discontinuous at x_0 despite the discontinuity of $E(D_i|X_i^* = x)$. As a result, we cannot identify the average treatment effect using (8) under the continuous measurement error. Indeed, our Monte Carlo simulations demonstrate this problem. Hence, we do not recommend focusing on (8) in empirical situations in which the researchers are aware of the measurement error and in which they are confident of the discontinuous rule but $E(D_i|X_i)$ is not apparently discontinuous at the threshold.

Remark 4. Yu (2012) focuses on a local polynomial estimator for τ_{XD} to study the average treatment effect at the threshold in a sharp RD design with a continuous forcing variable that contains classical measurement error. He first shows that τ_{XD} has identification bias for τ^* that is identical to equation (7). He then shows that the local polynomial estimator for τ_{XD} is consistent for τ^* if the measurement error tends to zero depending on the sample size under some rate conditions. We stress that the approaches in his paper and the present paper differ, as we state in the Introduction.

There may be at least three possible approaches to examine τ^* using the mismeasured forcing variable following the literature of measurement error (see Schennach, 2013). First, we could use parameteric specifications, as in Hullegie and Klein (2010). By correcting the identification bias based on the parametric specifications, we can identify τ^*. However, this approach is sensitive to the validity of the parametric specifications: if the parametric specifications are invalid, the identification analysis can be broken.

Second, we may be able to identify τ^* using instrumental variables or repeated measurements. It is well-known in the literature of measurement error that such additional variables are powerful tools for establishing identification with the problems of measurement error. However, valid instrumental variables or repeated measurements are not commonly available in empirical situations based on the RD designs.

Third, we can learn the average treatment effect through approximation methods. The present study uses this approach because the advantages of the approximation approach correspond to those of the RD design: they do not require parametric specifications or additional variables such as instrumental variables. While the approximation approach may not provide the exact identification for the average treatment effect, it provides meaningful information without restrictive requirements.

4 The small error variance approximation in the RD design

This section develops an approximation analysis for the average treatment effect at the threshold based on the small error variance approximation (SEVA) originally proposed by Chesher (1991). We show that the average treatment effect is approximated using an identified parameter when the standard deviation of the measurement error σ is small.

We focus on approximating $\tau^* = E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0-, D_i = 0)$ to learn the average treatment effect. In principle, we can also consider approximating $E(Y_i|X_i^* = x_0+) - E(Y_i|X_i^* = x_0-)$ based on the SEVA, although the precision of this approximation is worse. This notion comes from the same reason discussed in the previous section, that is, that $E(Y_i|X_i = x_0+) - E(Y_i|X_i = x_0-)$ vanishes because of continuous measurement error.

Before we state our formal result, we outline our approach for approximating τ^*. Extending the result in Chesher (1991), we show that the conditional density of Y_i given X_i^* and D_i is approximated as follows:

$$f_{Y|X^*D}(y|x,d) = f_{Y|XD}(y|x,d) - \sigma^2 \left(\log^{(1)} f_{X|D}(x|d)\right) f^{(1)}_{Y|XD}(y|x,d) - \frac{\sigma^2}{2} f^{(2)}_{Y|XD}(y|x,d) + o(\sigma^2).$$

for x near x_0, $y \in supp(Y)$, and $d \in \{0,1\}$.[3] The equation shows that the left-hand side is approximated up to the order $O(\sigma^2)$ by the terms on the right-hand side. Because the terms on the right-hand side relate the joint distribution of (Y_i, D_i, X_i) and σ, they are identified by

[3] As noted in Chesher (1991), while we can allow Y_i to be discrete, X_i^* and X_i must be continuous to establish the approximation. Nonetheless, to avoid complexity, we implicitly assume that Y_i is also continuous in this section.

the observable data if we extrapolate or forecast σ. Furthermore, this equation leads to

$$E(Y_i|X_i^* = x, D_i = d) = E(Y_i|X_i = x, D_i = d) - \sigma^2 \left(\log^{(1)} f_{X|D}(x|d)\right) E^{(1)}(Y_i|X_i = x, D_i = d) - \frac{\sigma^2}{2} E^{(2)}(Y_i|X_i = x, D_i = d) + o(\sigma^2).$$

The terms on the right-hand side are identified using the data. This indicates that the conditional mean of Y_i given X_i^* and D_i is approximated up to the order $O(\sigma^2)$ by the identified parameter. Accordingly, using this approximation, τ^* and the average treatment effect can be approximated up to the order $O(\sigma^2)$.

To show the approximation result in a rigorous manner, we require additional assumptions.

Assumption 3. *$E(Y_i|X_i^* = x_0+, D_i = 1)$, $E(Y_i|X_i^* = x_0-, D_i = 0)$, $E^{(s)}(Y_i|X_i = x_0+, D_i = 1)$, and $E^{(s)}(Y_i|X_i = x_0-, D_i = 0)$ exist for $s = 0, 1, 2$.*

Assumption 4. *(i) $f^{(s)}_{Y|X^*D}(y|x, d)$ is bounded in $y \in supp(Y)$ for x near x_0, $d \in \{0, 1\}$, and $s = 0, 1, \ldots, 5$. (ii) $f^{(s)}_{X^*|D}(x|d)$ is bounded in x near x_0 for $d \in \{0, 1\}$ and $s = 0, 1, \ldots, 5$. (iii) $f^{(s)}_{X|D}(x|d)$ is continuous at $x = x_0$ and bounded near x_0 for $d \in \{0, 1\}$ and $s = 0, 1$.*

Assumption 5. *$\int_{-\infty}^{\infty} y f^{(s)}_{Y|XD}(y|x, d)dy < \infty$ and $\int_{-\infty}^{\infty} y f^{(t)}_{Y|X^*D}(y|x, d)dy < \infty$ for x near x_0, $d \in \{0, 1\}$, $s = 0, 1, 2$, and $t = 0, 1, \ldots, 5$.*

The assumptions are regularity conditions for establishing the approximation for τ^*. Assumptions 3–5 ensure the existence of the limits, the boundedness of the densities, and the switching of the orders of integration and differentiation. The assumptions guarantee that the order of the approximation becomes $O(\sigma^2)$. We stress that we do not require partial differentiability of the conditional density of Y_i or X_i^* at x_0. Furthermore, the continuity of $f_{X|D}(\cdot|d)$ is implied by the continuity of $f_U(\cdot)$.

The following theorem presents an approximation for τ^* based on the SEVA. Let

$$\mu(x, d, \sigma) \coloneqq E(Y_i|X_i = x, D_i = d) - \sigma^2 g(x, d) E^{(1)}(Y_i|X_i = x, D_i = d) - \frac{\sigma^2}{2} E^{(2)}(Y_i|X_i = x, D_i = d),$$

where $g(x, d) \coloneqq \log^{(1)} f_{X|D}(x|d) = f^{(1)}_{X|D}(x|d)/f_{X|D}(x|d)$ and $d \in \{0, 1\}$.

Theorem 2. *Suppose that Assumptions 1 and 3–5 hold. When $\sigma \to 0$, it holds that*

$$\tau^* = \mu(x_0+, 1, \sigma) - \mu(x_0-, 0, \sigma) + o(\sigma^2). \tag{9}$$

Theorem 2 states that τ^* (and thus the average treatment effect) are approximated up to the order $O(\sigma^2)$ by the difference on the right-hand side when σ is small. The smaller standard deviation of the measurement error implies a more precise approximation for the average treatment effect.

The condition $\sigma \to 0$ means that σ is "sufficiently small" in the mathematical sense. While the original SEVA in Chesher (1991) does not require this condition, we need it for the SEVA

in the RD design. The reason why we require the condition (and the bounded support of U_i) is because we approximate the one-sided limits of the conditional expectations. The basic idea behind the SEVA is the convolution of the probability distributions and Taylor's theorem. In our setting, we should apply Taylor's theorem to the conditional densities at every point near the discontinuity point to approximate the one-sided limits of the conditional means. As a result, the condition $\sigma \to 0$ and the bounded support are required to ensure the establishment of the Taylor polynomials at every point near the discontinuity point. However, the condition $\sigma \to 0$ is a mathematical requirement, which does not mean that σ converges in the real world. In practice, the precision of the approximation depends on the magnitude of σ and the data-generating process. We demonstrate the approximate precision using our Monte Carlo simulations, which suggests that our approximation can work even when σ^2 accounts for about 20% of $var(X_i)$.

The terms on the right-hand side of (9) are identified by the data (Y_i, D_i, X_i) except for the standard deviation σ. In practice, we can extrapolate σ when we have additional public data or census data on X^* for the population of interest, because $\sigma^2 = var(X) - var(X^*)$ under Assumption 1. For example, suppose that we are interested in evaluating a policy program in a state based on survey data and the forcing variable is income, which may contain measurement error. In this situation, we can use public data on income in the state to extrapolate σ by estimating $var(X^*)$. Importantly, this procedure does not require additional variables on the observations $i = 1, \ldots, n$. Alternatively, we can learn the effect of the measurement error through (9) by forecasting σ, as in Battistin and Chesher (2014). Because the difference on the right-hand side of (9) is monotonic in σ, we can calculate the forecast intervals for the difference by forecasting several values of σ. Hence, Theorem 2 allows us to correct or forecast the identification bias because of the measurement error up to the order $O(\sigma^2)$.

Remark 5. Our approximate analysis could not be extended to the fuzzy RD design, because $E(D_i|X_i)$ in the fuzzy RD design is not discontinuous at the threshold owing to continuous measurement error. We consider the same setting in Section 2, except that D_i is not a deterministic function of X_i^*. Instead, in the fuzzy RD design, the conditional probability of $D_i = 1$ given X_i^* is discontinuous at the threshold: $E(D_i|X_i^* = x_0+) \neq E(D_i|X_i^* = x_0-)$. The fuzzy RD estimand is

$$\frac{E(Y_i|X_i^* = x_0+) - E(Y_i|X_i^* = x_0-)}{E(D_i|X_i^* = x_0+) - E(D_i|X_i^* = x_0-)},$$

which identifies the average treatment effect under the independence assumption between $Y_{1i} - Y_{0i}$ and D_i. Even without the independence assumption, the parameter identifies a local average treatment effect (see Hahn et al. (2001) for details). Here, $E(Y_i|X_i^*) \neq E(Y_i|X_i^*, D_i)$, because D_i is not a deterministic function of X_i^* in the fuzzy RD design.

It might seem that the fuzzy RD estimand is approximated by approximating each term in the estimand based on the SEVA, similar to the sharp RD design. That is, it might seem that

by extending the result in Chesher (1991), the average treatment effect in the fuzzy RD design is approximated by

$$\frac{\mu_Y(x_0+,\sigma)-\mu_Y(x_0-,\sigma)}{\mu_D(x_0+,\sigma)-\mu_D(x_0-,\sigma)},$$

where

$$\mu_Y(x,\sigma) \coloneqq E(Y_i|X_i=x) - \sigma^2\left(\log^{(1)} f_X(x)\right)E^{(1)}(Y_i|X_i=x) - \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i=x),$$
$$\mu_D(x,\sigma) \coloneqq E(D_i|X_i=x) - \sigma^2\left(\log^{(1)} f_X(x)\right)E^{(1)}(D_i|X_i=x) - \frac{\sigma^2}{2}E^{(2)}(D_i|X_i=x).$$

However, $\mu_D(x_0+,\sigma)-\mu_D(x_0-,\sigma)$ could vanish under continuous measurement error. Under the classical measurement error assumption, we observe that

$$E(D_i|X_i=x) = E(E(D_i|X_i^*,X_i=x)|X_i=x) = \int_{-\infty}^{\infty} E(D_i|X_i^*=x^*)f_{X^*|X}(x^*|x)dx^*.$$

Hence, the difference in the conditional probabilities is

$$E(D_i|X_i=x_0+) - E(D_i|X_i=x_0-) = \int_{-\infty}^{\infty} E(D_i|X_i^*=x^*)(f_{X^*|X}(x^*|x_0+) - f_{X^*|X}(x^*|x_0-))dx^*.$$

This difference vanishes under the continuity of $f_{X^*|X}(x^*|x)$ at $x=x_0$, which is implied by the continuity of $f_U(\cdot)$. Hence, $\mu_D(x_0+,\sigma)-\mu_D(x_0-,\sigma)$ could also vanish such that we cannot approximate the fuzzy RD estimand based on the SEVA.

Accordingly, we require other approaches to evaluate the effect of measurement error in the fuzzy RD design. This is beyond the scope of this paper.

5 Estimation

This section presents a nonparametric estimation procedure for the parameter that approximates the average treatment effect, that is, the difference on the right-hand side of (9). We develop the asymptotic properties of the nonparametric estimator.

Using the approximation analysis developed in the previous section, if we can consistently estimate the difference on the right-hand side of (9), the average treatment effect is approximately estimated up to the order $O(\sigma^2)$. We thus consider estimating

$$\begin{aligned}
\mu(x_0+,1,\sigma) &= E(Y_i|X_i=x_0+,D_i=1) \\
&\quad - \sigma^2 g(x_0,1)E^{(1)}(Y_i|X_i=x_0+,D_i=1) - \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i=x_0+,D_i=1), \\
\mu(x_0-,0,\sigma) &= E(Y_i|X_i=x_0-,D_i=0) \\
&\quad - \sigma^2 g(x_0,0)E^{(1)}(Y_i|X_i=x_0-,D_i=0) - \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i=x_0-,D_i=0),
\end{aligned}$$

where $g(x,d) \coloneqq f^{(1)}_{X|D}(x|d)/f_{X|D}(x|d)$. In the following, we assume that σ^2 is known, because σ^2 can be extrapolated through $\sigma^2 = var(X) - var(X^*)$ or forecast.

We can consistently estimate $g(x_0, d)$ by $\hat{g}(x_0, d)$ based on the kernel density and density derivative estimators:

$$\hat{g}(x_0, d) := \frac{\hat{f}^{(1)}_{X|D}(x_0|d)}{\hat{f}_{X|D}(x_0|d)},$$

where $\hat{f}_{X|D}(x_0|d) := (n_d h)^{-1} \sum_{i=1}^n \mathbf{1}(D_i = d) K_h(X_i)$, $\hat{f}^{(1)}_{X|D}(x_0|d) := -(n_d h^2)^{-1} \sum_{i=1}^n \mathbf{1}(D_i = d) K^{(1)}_h(X_i)$, n_d is the number of observations with $D_i = d$, $K_h(z) := K((z - x_0)/h)$, $K(\cdot)$ is some kernel function, and h is a bandwidth tending to zero as $n \to \infty$. $\hat{f}_{X|D}(x_0|d)$ and $\hat{f}^{(1)}_{X|D}(x_0|d)$ are consistent for $f_{X|D}(x_0|d)$ and $f^{(1)}_{X|D}(x_0|d)$, respectively, under the regularity conditions, similar to those in Silverman (1978) or Li and Racine (2007, Chapter 3) (see also Fan and Gijbels, 1996, Section 2.7). Accordingly, $\hat{g}(x_0, d)$ is consistent for $g(x_0, d)$ under the conditions. We thus assume the conditions implicitly and omit the details for the asymptotic properties of this estimator.

We next focus on estimating $E^{(s)}(Y_i|X_i = x_0+, D_i = 1)$ and $E^{(s)}(Y_i|X_i = x_0-, D_i = 0)$ for $s = 0, 1, 2$, which are estimated using local polynomial regressions (Fan and Gijbels, 1996). The estimators for $E^{(s)}(Y_i|X_i = x_0+, D_i = 1)$ are given by the following p-th order local polynomial regression:

$$(\hat{\alpha}^+, \hat{\beta}^{+\prime})' := \underset{(a,b')' \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^n I_i D_i \left(Y_i - a - b_1(X_i - x_0) - \cdots - b_p(X_i - x_0)^p\right)^2 K_h(X_i), \quad (10)$$

where $p \geq 2$ is some positive integer and $I_i := \mathbf{1}(X_i \geq x_0)$.[4] Then, $\hat{\alpha}^+$ is the estimator for $E(Y_i|X_i = x_0+, D_i = 1)$, and $\hat{\beta}^+_k$ is that for $(k!)^{-1} E^{(k)}(Y_i|X_i = x_0+, D_i = 1)$ for $k = 1, \ldots, p$. Similarly, the estimators for $E^{(s)}(Y_i|X_i = x_0-, D_i = 0)$ for $s = 0, 1, 2$ are given by the following p-th order local polynomial regression:

$$(\hat{\alpha}^-, \hat{\beta}^{-\prime})' := \underset{(a,b')' \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^n (1 - I_i)(1 - D_i) \left(Y_i - a - b_1(X_i - x_0) - \cdots - b_p(X_i - x_0)^p\right)^2 K_h(X_i),$$

Then, $\hat{\alpha}^-$ is the estimator for $E(Y_i|X_i = x_0-, D_i = 0)$, and $\hat{\beta}^-_k$ is that for $(k!)^{-1} E^{(k)}(Y_i|X_i = x_0-, D_i = 0)$ for $k = 1, \ldots, p$.

The parameter approximating the average treatment effect is estimated by

$$\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma),$$

where

$$\hat{\mu}(x_0+, 1, \sigma) := \hat{\alpha}^+ - \sigma^2 \hat{g}(x_0, 1) \hat{\beta}^+_1 - \sigma^2 \hat{\beta}^+_2,$$
$$\hat{\mu}(x_0-, 0, \sigma) := \hat{\alpha}^- - \sigma^2 \hat{g}(x_0, 0) \hat{\beta}^-_1 - \sigma^2 \hat{\beta}^-_2,$$

which are estimators for $\mu(x_0+, 1, \sigma)$ and $\mu(x_0-, 0, \sigma)$, respectively.

[4] In practice, the kernel function and bandwidth used for the local polynomial regressions can differ from those used to estimate $f_{X|D}(x_0|d)$ and $f^{(1)}_{X|D}(x_0|d)$.

To develop the asymptotic properties of this estimator, we introduce additional assumptions. The assumptions are standard regularity conditions for developing asymptotic properties for the local polynomial estimators, which are analogous to the conditions in Hahn et al. (2001) and Porter (2003). Let $V_i \coloneqq Y_i - E(Y_i|X_i, D_i)$.

Assumption 6. *$K(\cdot)$ is continuous, symmetric, and non-negative with compact support. For simplicity, the support is assumed to be $[-M, M]$ for some finite $M > 0$.*

Assumption 7. *$f_X(\cdot)$ is bounded, continuous, and bounded away from zero near x_0.*

Assumption 8. *$E(D_i|X_i = x_0+)$ and $E(1 - D_i|X_i = x_0-)$ exist and are non-zero.*

Assumption 9. *(i) $E(V_i^2|X_i = x, D_i = 1)$ and $E(V_i^2|X_i = x, D_i = 0)$ are bounded near x_0 and $E(V_i^2|X_i = x_0+, D_i = 1)$ and $E(V_i^2|X_i = x_0-, D_i = 0)$ exist. (ii) $E(|V_i|^{2+\zeta}|X_i = x)$ is bounded near x_0 for some $\zeta > 0$.*

Assumption 10. *(i) $E(Y_i|X_i = x, D_i = d)$ is $p+1$-times continuously differentiable for x near x_0 and $d \in \{0, 1\}$. (ii) $E^{(k)}(Y_i|X_i = x_0+, D_i = 1)$ and $E^{(k)}(Y_i|X_i = x_0-, D_i = 0)$ exist for $k = 1, \ldots, p+1$. (iii) There exists some $\tilde{M} > 0$ such that $E^{(p+1)}(Y_i|X_i = x, D_i = 1)$ is bounded for $x \in [x_0, x_0 + \tilde{M}]$ and $E^{(p+1)}(Y_i|X_i = x, D_i = 0)$ is bounded for $x \in [x_0 - \tilde{M}, x_0)$.*

To develop asymptotic properties of $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$, we first study asymptotic properties of the local polynomial estimators.

Lemma 1. *Suppose that Assumptions 6–10 hold. When $n \to \infty$, $h \to 0$, $nh \to \infty$, and $\sqrt{nh}h^{p+1} \to \tilde{C}$ for some $\tilde{C} \in [0, \infty)$, it holds that*

$$\left(\hat{\alpha}^+, \hat{\beta}^{+\prime}\right)' - \left(\alpha^+, \beta^{+\prime}\right)' \xrightarrow{p} 0,$$
$$\left(\hat{\alpha}^-, \hat{\beta}^{-\prime}\right)' - \left(\alpha^-, \beta^{-\prime}\right)' \xrightarrow{p} 0,$$

and

$$\sqrt{nh}H^{-1}\left(\left(\hat{\alpha}^+, \hat{\beta}^{+\prime}\right)' - \left(\alpha^+, \beta^{+\prime}\right)'\right) \rightsquigarrow N\left(B^+, \Omega^+\right), \tag{11}$$
$$\sqrt{nh}H^{-1}\left(\left(\hat{\alpha}^-, \hat{\beta}^{-\prime}\right)' - \left(\alpha^-, \beta^{-\prime}\right)'\right) \rightsquigarrow N\left(B^-, \Omega^-\right), \tag{12}$$

where

$$\begin{aligned}
H &\coloneqq diag(1, h^{-1}, \ldots, h^{-p}),\\
B^+ &\coloneqq \tilde{C}E^{(p+1)}(Y_i|X_i = x_0+, D_i = 1)(\Gamma^+)^{-1}(\gamma_{p+1}, \ldots, \gamma_{2p+1})',\\
B^- &\coloneqq \tilde{C}E^{(p+1)}(Y_i|X_i = x_0-, D_i = 0)(\Gamma^-)^{-1}\left((-1)^{p+1}\gamma_{p+1}, \ldots, (-1)^{2p+1}\gamma_{2p+1}\right)',\\
\Omega^+ &\coloneqq \frac{E(V_i^2|X_i = x_0+, D_i = 1)}{E(D_i|X_i = x_0+)f_X(x_0)}(\Gamma^+)^{-1}\Delta^+(\Gamma^+)^{-1},\\
\Omega^- &\coloneqq \frac{E(V_i^2|X_i = x_0-, D_i = 0)}{E(1 - D_i|X_i = x_0-)f_X(x_0)}(\Gamma^-)^{-1}\Delta^-(\Gamma^-)^{-1},
\end{aligned}$$

$$\Gamma^+ \coloneqq (\gamma_{k+l-2})_{(k,l)}, \qquad \Gamma^- \coloneqq \left((-1)^{k+l+1}\gamma_{k+l-2}\right)_{(k,l)}, \qquad \textit{for} \quad k,l = 1,\dots,p+1,$$

$$\Delta^+ \coloneqq (\delta_{k+l-2})_{(k,l)}, \qquad \Delta^- \coloneqq \left((-1)^{k+l+1}\delta_{k+l-2}\right)_{(k,l)}, \qquad \textit{for} \quad k,l = 1,\dots,p+1,$$

$$\gamma_q \coloneqq \int_0^M u^q K(u)du, \qquad \delta_q \coloneqq \int_0^M u^q K^2(u)du, \qquad \textit{for} \quad q = 0,\dots,2p+1.$$

Lemma 1 shows that the vectors of local polynomial estimators are consistent and asymptotically normal. The asymptotic distributions are not centered at zero because of the presence of the asymptotic biases. The asymptotic biases of the vectors of the local polynomial estimators are of order $O(Hh^{p+1})$ and depend on the one-sided derivatives of $E(Y_i|X_i, D_i)$.

Lemma 1 states that the convergence rates of the vectors of the local polynomial estimators are $1/(\sqrt{nh}H^{-1})$. Specifically, the convergence rates of $\hat{\alpha}^+$ and $\hat{\alpha}^-$ are of order $1/\sqrt{nh}$, those of $\hat{\beta}_1^+$ and $\hat{\beta}_1^-$ are of order $1/(\sqrt{nh}h)$, and those of $\hat{\beta}_2^+$ and $\hat{\beta}_2^-$ are of order $1/(\sqrt{nh}h^2)$. These results are consistent with the results in the literature of local polynomial regressions, such as those in Fan and Gijbels (1992), Ruppert and Wand (1994), and Masry (1996a,b). From these results, we expect the convergence rate of $\hat{\mu}(x_0+,1,\sigma) - \hat{\mu}(x_0-,0,\sigma)$ to be of order $1/(\sqrt{nh}h^2)$.

The asymptotic properties of $\hat{\mu}(x_0+,1,\sigma) - \hat{\mu}(x_0-,0,\sigma)$ are developed in the following theorem.

Theorem 3. *Suppose that Assumptions 6–10 hold and* $\hat{g}(x_0,d) \xrightarrow{p} g(x_0,d)$ *for* $d \in \{0,1\}$. *When* $n \to \infty$, $h \to 0$, $nh \to \infty$, *and* $\sqrt{nh}h^{p+1} \to \tilde{C}$ *for some* $\tilde{C} \in [0,\infty)$, *it holds that*

$$\hat{\mu}(x_0+,1,\sigma) - \hat{\mu}(x_0-,0,\sigma) - (\mu(x_0+,1,\sigma) - \mu(x_0-,0,\sigma)) \xrightarrow{p} 0,$$

and

$$\sqrt{nh}h^2\left(\hat{\mu}(x_0+,1,\sigma) - \hat{\mu}(x_0-,0,\sigma) - (\mu(x_0+,1,\sigma) - \mu(x_0-,0,\sigma))\right) \rightsquigarrow N(B,\Omega),$$

where $B \coloneqq \sigma^2 e_3'(B^- - B^+)$, $\Omega \coloneqq \sigma^4 e_3'(\Omega^+ + \Omega^-)e_3$, *and* $e_3 \coloneqq (0,0,1,0,\dots,0)'$ *is the* $p+1$ *vector and* B^+, B^-, Ω^+, *and* Ω^- *are defined in Lemma 1.*

Theorem 3 shows that the estimator for the parameter approximating the average treatment effect is consistent for the parameter and asymptotically normal. The asymptotic distribution is not centered at zero because of the presence of the asymptotic bias. The convergence speed of the estimator is of order $1/(\sqrt{nh}h^2)$, which is expected by the discussion above: the convergence rate of the estimator is determined by those of the estimators for $E^{(2)}(Y_i|X_i = x_0+, D_i = 1)$ and $E^{(2)}(Y_i|X_i = x_0-, D_i = 0)$. This convergence rate is slower than $1/\sqrt{nh}$, although this result is standard. This convergence speed is a limitation of the local polynomial estimators, which cannot be overcome if we use the local polynomial regressions. We can overcome the slow convergence rate using parametric methods such as parametric polynomial regressions, although we do not pursue this issue here because this paper focuses on nonparametric methods.

Remark 6. The selection of the kernel functions and bandwidths are practically concerned. In particular, the precision of the nonparametric estimator $\hat{\mu}(x_0+,1,\sigma) - \hat{\mu}(x_0-,0,\sigma)$ largely

depends on selecting the bandwidths, as other nonparametric estimators do. We explain the details of bandwidth selection in our Monte Carlo simulations (Section 6). We note that the bandwidth selection for RD designs developed in Imbens and Kalyanaraman (2012) and Arai and Ichimura (2014) cannot directly apply our setting. This is because our estimand, the difference on the right-hand side of (9), is not typical in the sharp RD design.

6 Monte Carlo simulations

This section presents the results of the Monte Carlo simulations. We first describe the simulation designs and the implementation of our approximate analysis, and then we report the results.

The simulations are conducted with R 3.1.1 for Windows 7. 1000 replications are used for the simulation. We set the sample size to $n = 2500$, which may look somewhat large, although this is required to execute higher-order local polynomial regressions.

6.1 Designs

We consider two designs for the potential outcomes:

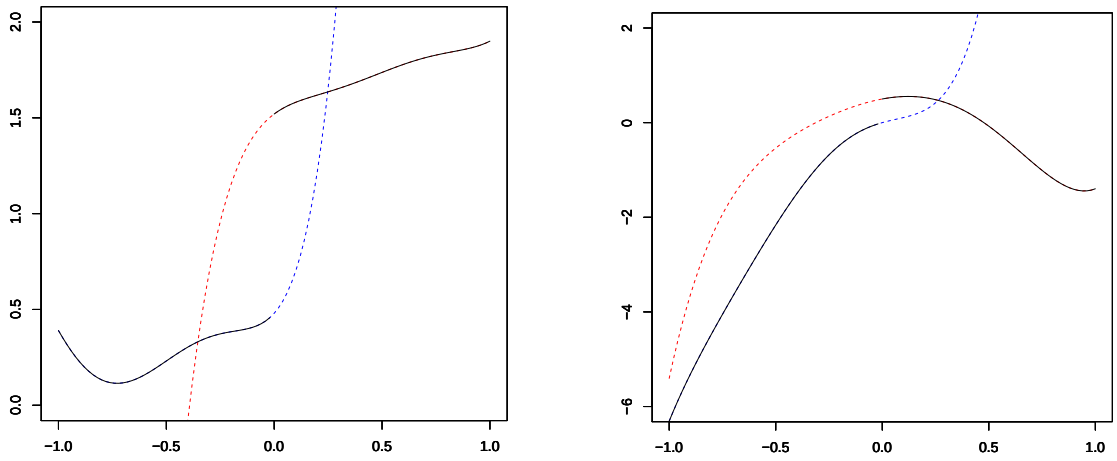
$$\text{Design A:} \quad \begin{aligned} Y_{1i} &= 1.52 + 0.84X_i^* - 3.0(X_i^*)^2 + 7.99(X_i^*)^3 - 9.01(X_i^*)^4 + 3.56(X_i^*)^5 + e_i, \\ Y_{0i} &= 0.48 + 1.27X_i^* + 7.18(X_i^*)^2 + 20.21(X_i^*)^3 + 21.54(X_i^*)^4 + 7.33(X_i^*)^5 + e_i, \end{aligned}$$

$$\text{Design B:} \quad \begin{aligned} Y_{1i} &= 0.5 + 0.84X_i^* - 0.3(X_i^*)^2 - 2.397(X_i^*)^3 - 0.901(X_i^*)^4 + 3.56(X_i^*)^5 + e_i, \\ Y_{0i} &= 0 + 1.27X_i^* - 3.59(X_i^*)^2 + 14.147(X_i^*)^3 + 23.694(X_i^*)^4 + 10.995(X_i^*)^5 + e_i, \end{aligned}$$

where $X_i^* \sim i.i.d.\ 2Beta(2,4) - 0.7$ and $e_i \sim i.i.d.N(0, 0.1295^2)$ in both designs, which implies $E(X_i^*) = -1/30$, $var(X_i^*) = 32/252$. The treatment is $D_i = \mathbf{1}(X_i^* \geq 0)$, that is, $x_0 = 0$, in each design. Design A is similar to Imbens and Kalyanaraman (2012, Lee design), Arai and Ichimura (2014, Design 1), and Calonico et al. (2014, Model 1), which is motivated by Lee (2008)'s data. Design B is analogous to Calonico et al. (2014, Model 3). However, the average treatment effects and $E(X_i^*)$ are bigger here, which reveal the riskiness of the mismeasured forcing variable. For illustration, the conditional means are plotted in Figure 3.

The observable mismeasured forcing variable is generated as $X_i = X_i^* + \sigma U_i$, where U_i is the i.i.d. truncated normally distributed random variable with mean 0 and standard deviation 1, whose support is $[-3, 3]$. U_i is independent of the other variables. We consider three values for the standard deviation of the measurement error: $\sigma = 0.12, 0.15, 0.18$. Under each σ, the magnitudes of σ^2 account for about 10%, 15%, and 20% of $var(X_i)$, respectively. For illustration, the densities of X_i^* and X_i for each σ are plotted in Figure 4.

We evaluate the performance of four estimators. The first is the estimator based on the SEVA developed in Section 5 (we denote it as "ESEVA"): $\hat{\mu}(x_0+, 1, \sigma) - \hat{\mu}(x_0-, 0, \sigma)$. For the kernel density estimation, we use the Epanechnikov kernel function $K_1(u) = 3/4(1-u^2)\mathbf{1}(|u| \leq 1)$ and the normal-scale rule bandwidth given by $h_d = 2.34\hat{\sigma}_{X,d} n_d^{-1/5}$, where $\hat{\sigma}_{X,d}$ is the square root of the sample variance of X_i for observations with $D_i = d$ for $d \in \{0, 1\}$ and n_d is the number of



(a) Design A (b) Design B

Figure 3: The conditional means in each design are plotted. The dotted red and blue lines are $E(Y_{1i}|X_i^* = x)$ and $E(Y_{0i}|X_i^* = x)$, respectively. The solid black line is $E(Y_i|X_i^* = x)$.

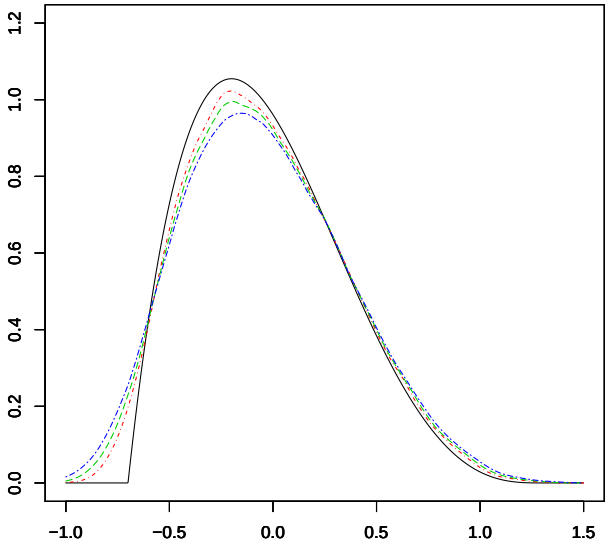


Figure 4: The densities of X_i^* and X_i for each σ are plotted. The black line is the density of X_i^*, the red one is that of X_i for $\sigma = 0.12$, the green one is that for $\sigma = 0.15$, and the blue one is that for $\sigma = 0.18$.

observations with $D_i = d$. For the kernel density derivative estimation, we employ the kernel function proposed by Jones (1994), $K_2^{(1)}(u) = u(1-u)^2\mathbf{1}(|u| \leq 1)/4$ and the normal scale rule bandwidth $h_d = \hat{\sigma}_{X,d}(112\sqrt{\pi}/n_d)^{1/7}$. We use the local linear regression to estimate the one-sided limits of the conditional expectation and the second-order local polynomial regressions to estimate the first and second derivatives. We use separate-order local polynomial regressions because of the different convergence rates of the estimators and the automatic boundary adaptive property of the local polynomial regressions (Fan and Gijbels, 1996).[5] For all local polynomial regressions, we employ the triangle kernel function $K_3(u) = (1-|u|)\mathbf{1}(|u| \leq 1)$. The bandwidth for the p-th order local polynomial regression required to estimate the ν-th right derivative is selected by the plug-in method developed in Fan and Gijbels (1996, p.67), that is,

$$h_{\nu,p} = C_{\nu,p}(K)\left(\frac{\hat{\sigma}^2(x_0+)}{(\hat{E}^{(p+1)}(Y_i|X_i = x_0+, D_i = 1))^2 \hat{f}_X(x_0)\hat{E}(D_i|X_i = x_0+)n^+}\right)^{1/(2p+3)},$$

for $(\nu, p) = (0,1), (1,2), (2,2)$. $\hat{\sigma}^2(x_0+)$ is the local linear estimator for the conditional variance of Y_i given $D_i = 1$ and $X_i = x_0+$. $\hat{f}_X(x_0)$ is the kernel density estimator for $f_X(x_0)$. $\hat{E}(D_i|X_i = x_0+)$ is the local linear estimator for $E(D_i|X_i = x_0+)$. n^+ is the number of observations with $D_i = 1$ and $X_i \geq x_0$. $C_{\nu,p}(K)$ is a constant depending on the kernel function whose definition is given in Fan and Gijbels (1996, p.67). For the triangle kernel, we set $C_{0,1}(K_3) = 2.9925$, $C_{1,2}(K_3) = 3.5218$, and $C_{2,2}(K_3) = 3.1077$. To select the bandwidth for the second order local polynomial regression, we need a pilot estimate for $\hat{E}^{(3)}(Y_i|X_i = x_0+, D_i = 1)$, which is estimated using the third-order local polynomial regression. Selecting the bandwidths for the left derivatives is analogous. The variance of the measurement error σ^2 is estimated through the difference in the sample variance of X_i and that of X^* using artificial additional data on X^* whose sample size is 5000.

The second is the estimator for τ_X, $\hat{\tau}_X \coloneqq \tilde{\alpha}_Y^+ - \tilde{\alpha}_Y^-$, where

$$(\tilde{\alpha}_Y^+, \tilde{\beta}_Y^+)' \coloneqq \underset{(a,b)'\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^n I_i \left(Y_i - a - b_1(X_i - x_0)\right)^2 K\left(\frac{X_i - x_0}{h}\right),$$
$$(\tilde{\alpha}_Y^-, \tilde{\beta}_Y^-)' \coloneqq \underset{(a,b)'\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^n (1 - I_i) \left(Y_i - a - b_1(X_i - x_0)\right)^2 K\left(\frac{X_i - x_0}{h}\right).$$

That is, $\hat{\tau}_X$ is the estimator based on the local linear regressions. For the estimator, we employ the triangle kernel function and the plug-in bandwidth discussed in Fan and Gijbels (1996).

The third is the estimator for τ_{XD}, $\hat{\tau}_{XD} \coloneqq \hat{\alpha}^+ - \hat{\alpha}^-$, based on the local linear regressions developed in Lemma 1 (i.e., $p = 1$). For this estimator, we employ the triangle kernel function and the plug-in bandwidth discussed previously.

[5] We can also use third-order polynomial regressions to estimate the second derivatives. However, we find that the performance of these estimators is worse than that of estimators based on the second-order polynomial regressions. This is because the plug-in bandwidth for the third-order polynomial regressions is of order $n^{-1/9}$, which leads to oversmoothing bandwidth under sample size 2500. We require a larger sample size to employ the plug-in bandwidth of order $n^{-1/9}$. Hence, we employ the second-order polynomial regressions, which lead to the plug-in bandwidth of order $n^{-1/7}$.

		ESEVA		$\hat{\tau}_X$		$\hat{\tau}_{XD}$		Fuzzy	
σ	true	bias	std	bias	std	bias	std	bias	std
0.12	1.04	0.0710	0.0343	-1.0083	0.1121	0.1028	0.0224	-0.0755	7.5655
0.15	1.04	0.0799	0.0324	-0.9984	0.1080	0.1196	0.0208	0.2164	4.1668
0.18	1.04	0.0920	0.0316	-1.0144	0.1044	0.1367	0.0205	0.8930	19.5753

Table 1: Monte Carlo simulation results with design A

		ESEVA		$\hat{\tau}_X$		$\hat{\tau}_{XD}$		Fuzzy	
σ	true	bias	std	bias	std	bias	std	bias	std
0.12	0.5	0.0853	0.0919	-0.4925	0.0789	0.2539	0.0497	0.4399	20.6762
0.15	0.5	0.1044	0.1031	-0.4916	0.0901	0.3336	0.0562	-1.5143	33.8111
0.18	0.5	0.1586	0.1087	-0.4915	0.0994	0.4236	0.0605	0.1016	8.9724

Table 2: Monte Carlo simulation results with design B

The fourth is the estimator in the fuzzy RD design discussed in Remark 3, that is, the estimator of (8) (we denote it as "Fuzzy"). Specifically, the estimator is $(\tilde{\alpha}_Y^+ - \tilde{\alpha}_Y^-)/(\tilde{\alpha}_D^+ - \tilde{\alpha}_D^-)$, where $\tilde{\alpha}_D^+$ and $\tilde{\alpha}_D^-$ are obtained by the minimization problems here in which Y_i is replaced with D_i. For the estimator, we employ the triangle kernel function and the plug-in bandwidth in Fan and Gijbels (1996).

6.2 Results

The results of the Monte Carlo simulations with designs A and B are reported in Tables 1 and 2, respectively. The column labeled "true" reports the true average treatment effect at the threshold. The bias for the average treatment effect and the standard deviation of each estimator are presented in the tables.

The simulation results demonstrate that the approximate analysis based on the SEVA is informative for learning the average treatment effect. In both designs, the biases of ESEVA are moderate for each σ. However, the biases of ESEVA with design B are somewhat larger than those with design A. Nonetheless, the biases of ESEVA are considerably smaller than those of the remaining estimators in all cases. The biases of ESEVA increase as σ increases, although this is expected by our analysis developed in Section 4. Accordingly, the simulation results corroborate our approximate analysis based on the SEVA.

The standard deviations of ESEVA are moderate in both designs for all σ. However, as σ becomes larger with design B, the standard deviations increase. This is expected using our approximation analysis: when σ is large, the effects of the second and third terms in $\hat{\mu}(x_0+, 1, \sigma)$ and $\hat{\mu}(x_0-, 0, \sigma)$ on the standard deviation increase. Nonetheless, the standard deviations become smaller as n increases, because the asymptotic variance of ESEVA is of order $1/(\sqrt{nh}h^2)$.

The simulation results reveal that $\hat{\tau}_X$ has severe bias for identifying the average treatment effect. The bias of $\hat{\tau}_X$ is critical even for small σ. As expected by the analysis in Section 3, the estimates of $\hat{\tau}_X$ in both designs are close to 0, which leads us to the misleading consequence

σ	mean	std
0.12	0.0276	0.0946
0.15	0.0372	0.0929
0.18	0.0231	0.0888

Table 3: Estimates of $E(D_i|X_i = x_0+) - E(D_i|X_i = x_0-)$

in which there is no treatment effect. This result indicates that $\hat{\tau}_X$ is not consistent for the average treatment effect because of the identification bias caused by the measurement error.

The biases of $\hat{\tau}_{XD}$ are relatively moderate compared with those of $\hat{\tau}_X$. However, the biases of $\hat{\tau}_{XD}$ are larger than those of ESEVA in all cases. In particular, the biases of $\hat{\tau}_{XD}$ with design B are about three times as large as those of ESEVA. Furthermore, the mean squared errors of $\hat{\tau}_{XD}$ are bigger than those of ESEVA in all cases. These results suggest that ESEVA functions better than $\hat{\tau}_{XD}$ in all cases.

The performance of Fuzzy is poor. The estimator is unstable and both the bias and the standard deviation are incoherent in each setting. This is because the measurement error causes the discontinuity of the conditional mean of D_i to vanish, as discussed in Remark 3. The mean and standard deviation of the estimates of the discontinuity of $E(D_i|X_i = x)$ at x_0 with design A are reported in Table 3,[6] which reveals that the discontinuity vanishes because of the measurement error even for small σ. According to the simulation results, we do not recommend using the fuzzy RD estimand in situations in which the discontinuity size of $E(D_i|X_i)$ is apparently small despite a confident discontinuous rule.

To summarize, the simulation results corroborate our theoretical analysis. The measurement error leads the difference in the mean outcomes just above and below the threshold and the discontinuous size of the conditional means of D_i to vanish. In addition, the approximate analysis based on the SEVA works more successfully than the remaining estimators: the biases and the mean squared errors of ESEVA are smaller than those of $\hat{\tau}_X$, $\hat{\tau}_{XD}$, and Fuzzy.

7 Conclusion

This paper presents a nonparametric analysis in the sharp RD design in which the forcing variable contains measurement error. We show that the average treatment effect given the "true" forcing variable at the discontinuity point cannot be identified based on the difference in the mean outcomes given the mismeasured forcing variable. We present the exact form of the identification bias, which leads us to the misleading consequence in which there is no treatment effect even if there exists a significant treatment effect. To examine the average treatment effect using the mismeasured forcing variable, we propose approximating it using the small error variance approximation originally developed by Chesher (1991). We develop an estimation method for the parameter that approximates the average treatment effect based on

[6] Because the data-generating processes of (D_i, X_i^*, X_i) are the same in each design, we have similar results for the estimates of the discontinuity with each design. We thus report the results only with design A.

local polynomial regressions and the kernel density estimation. Monte Carlo simulations reveal that the identification bias caused by the measurement error is critical, and they corroborate the performance of our approximation analysis.

Future work: While this paper focuses only on the sharp RD design, it is worth investigating the effect of measurement error in the fuzzy RD design in which the forcing variable may contain measurement error. The conditional probability of the treatment may vanish because of the continuous measurement error such that the small error variance approximation cannot be executed in the fuzzy RD design, as we discuss in Remark 5. Thus, we would need other approaches to examine the causal effect in the fuzzy RD design with a mismeasured forcing variable.

A Appendix

This appendix presents proofs of the theorems and lemma in the text.

A.1 Proof of Theorem 1

Proof of (6): First,

$$\begin{aligned} Y_i &= E(Y_i|X_i^*) + Y_i - E(Y_i|X_i^*) \\ &= E(D_i Y_{1i}|X_i^*) + E((1-D_i)Y_{0i}|X_i^*) + W_i \\ &= \mathbf{1}(X_i^* \geq x_0)E(Y_{1i}|X_i^*) + \mathbf{1}(X_i^* < x_0)E(Y_{0i}|X_i^*) + W_i \end{aligned}$$

where $W_i \coloneqq Y_i - E(Y_i|X_i^*)$. We thus have

$$\begin{aligned} E(Y_i|X_i = x_0+) = &\int_{x_0}^{\infty} E(Y_{1i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0+)dx^* \\ &+ \int_{-\infty}^{x_0} E(Y_{0i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0+)dx^* + E(W_i|X_i = x_0+), \qquad (13) \\ E(Y_i|X_i = x_0-) = &\int_{x_0}^{\infty} E(Y_{1i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0-)dx^* \\ &+ \int_{-\infty}^{x_0} E(Y_{0i}|X_i^* = x^*) f_{X^*|X}(x^*|x_0-)dx^* + E(W_i|X_i = x_0-), \qquad (14) \end{aligned}$$

by Assumption 2 and the dominated convergence theorem.

Under Assumption 1, $f_{X^*|X}(x^*|x)$ is continuous at $x = x_0$ for $x^* \in \mathbb{R}$ such that $f_{X^*|X}(x^*|x_0+) = f_{X^*|X}(x^*|x_0-)$. This is because

$$\begin{aligned} f_{X^*|X}(x^*|x) &= \frac{f_{X^*X}(x^*, x)}{f_X(x)} \\ &= \frac{f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})}{\int_{-\infty}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*}, \end{aligned}$$

by Assumption 1 and the convolution of the probability distributions. Hence, the continuity of $f_{X^*|X}(x^*|\cdot)$ follows from that of $f_U(\cdot)$ and the dominated convergence theorem.

To show (6), it thus suffices to show that for any x,

$$E(W_i|X_i = x) = 0. \tag{15}$$

Because $E(W_i|X_i = x) = E(W_i|X_i = x, D_i = 1)\Pr(D_i = 1|X_i = x) + E(W_i|X_i = x, D_i = 0)\Pr(D_i = 0|X_i = x)$ by the law of iterated expectations, we show $E(W_i|X_i = x, D_i = d) = 0$ for $d \in \{0, 1\}$. To this end, we first compute $f_{W|XD}(w|x, 1)$:

$$\begin{aligned}
f_{W|XD}(w|x,1) &= \frac{f_{XW|D}(x,w|1)}{f_{X|D}(x|1)} \\
&= \frac{\sigma^{-1}\int_{-\infty}^{\infty} f_{X^*W|D}(x^*,w|1) f_{U|D}(\frac{x-x^*}{\sigma}|1)dx^*}{\sigma^{-1}\int_{-\infty}^{\infty} f_{X^*|D}(x^*|1) f_{U|D}(\frac{x-x^*}{\sigma}|1)dx^*} \\
&= \frac{\int_{x_0}^{\infty} f_{X^*W}(x^*,w) f_U(\frac{x-x^*}{\sigma})dx^*/(1-F_{X^*}(x_0))}{\int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*/(1-F_{X^*}(x_0))} \\
&= \frac{\int_{x_0}^{\infty} f_{W|X^*}(w|x^*) f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*}{\int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*},
\end{aligned}$$

where the second and third equalities follow from Assumption 1 and the convolution of the probability distributions. Then, we have

$$\begin{aligned}
E(W_i|X_i = x, D_i = 1) &= \frac{\int w \int_{x_0}^{\infty} f_{W|X^*}(w|x^*) f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*dw}{\int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*} \\
&= \frac{\int_{x_0}^{\infty} E(W_i|X_i^* = x^*) f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*}{\int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma})dx^*} \\
&= 0,
\end{aligned} \tag{16}$$

where the second equality follows from Fubini's theorem and the third equality follows from $E(W_i|X_i^*) = 0$. Similarly, we have

$$E(W_i|X_i = x, D_i = 0) = 0. \tag{17}$$

Therefore, we obtain (15). Consequently, we have (6) by (13), (14), and (15).

Proof of (7): The proof is almost identical to that in Yu (2012). First,

$$\begin{aligned}
Y_i &= E(Y_i|X_i^*) + Y_i - E(Y_i|X_i^*) \\
&= E(Y_{0i}|X_i^*) + \mathbf{1}(X_i^* \geq x_0)\{E(Y_{1i} - Y_{0i}|X_i^*) - \tau^*\} + \mathbf{1}(X_i^* \geq x_0)\tau^* + W_i \\
&= m(X_i^*) + \mathbf{1}(X_i^* \geq x_0)\tau^* + W_i \\
&= m(X_i^*) + D_i\tau^* + W_i,
\end{aligned} \tag{18}$$

where $W_i \coloneqq Y_i - E(Y_i|X_i^*)$ and $m(X_i^*) \coloneqq E(Y_{0i}|X_i^*) + \mathbf{1}(X_i^* \geq x_0)\big(E(Y_{1i} - Y_{0i}|X_i^*) - \tau^*\big)$.

To show (7), we compute

$$\begin{aligned}
&E(Y_i|X_i = x_0+, D_i = 1) \\
&= E(m(X_i^*)|X_i = x_0+, D_i = 1) + \tau^* + E(W_i|X_i = x_0+, D_i = 1),
\end{aligned} \tag{19}$$

and

$$E(Y_i|X_i = x_0-, D_i = 0) = E(m(X_i^*)|X_i = x_0-, D_i = 0) + E(W_i|X_i = x_0-, D_i = 0). \quad (20)$$

By (16) and (17), we have $E(W_i|X_i = x_0+, D_i = 1) = E(W_i|X_i = x_0-, D_i = 0) = 0$. To evaluate $E(m(X_i^*)|X_i = x, D_i = 1)$, we compute $f_{X^*|XD}(x^*|x, 1)$:

$$\begin{aligned} f_{X^*|XD}(x^*|x,1) &= \frac{f_{X^*X|D}(x^*, x|1)}{f_{X|D}(x|1)} \\ &= \frac{\mathbf{1}(x^* \geq x_0) f_{X^*X}(x^*, x)/(1 - F_{X^*}(x_0))}{f_{X|D}(x|1)} \\ &= \frac{\mathbf{1}(x^* \geq x_0) f_{X^*|X}(x^*|x) f_X(x)/(1 - F_{X^*}(x_0))}{f_{X|D}(x|1)} \\ &= \frac{\mathbf{1}(x^* \geq x_0) f_{X^*|X}(x^*|x) f_X(x)/(1 - F_{X^*}(x_0))}{\sigma^{-1} \int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma}) dx^*/(1 - F_{X^*}(x_0))}, \end{aligned}$$

where the first and third equalities follow from Bayes' theorem and the fourth equality follows from Assumption 1 and the convolution of the probability distributions. Then, we have

$$\begin{aligned} E(m(X_i^*)|X_i = x, D_i = 1) &= \frac{\int m(x^*) \mathbf{1}(x^* \geq x_0) f_{X^*|X}(x^*|x) f_X(x) dx^*}{\sigma^{-1} \int_{x_0}^{\infty} f_{X^*}(x^*) f_U(\frac{x-x^*}{\sigma}) dx^*} \\ &= \sigma \frac{\int_{x_0}^{\infty} m(x^*) f_{X^*|X}(x^*|x) f_X(x) dx^*}{\int_{x_0}^{\infty} f_{X^*U}(x^*, \frac{x-x^*}{\sigma}) dx^*} \\ &= \frac{\int_{x_0}^{\infty} m(x^*) f_{X^*|X}(x^*|x) f_X(x) dx^*}{\int_{x_0}^{\infty} f_{X^*X}(x^*, x) dx^*} \\ &= \frac{\int_{x_0}^{\infty} m(x^*) f_{X^*|X}(x^*|x) dx^*}{\int_{x_0}^{\infty} f_{X^*|X}(x^*|x) dx^*}, \end{aligned} \quad (21)$$

by the convolution of the probability distributions. Similarly, we obtain

$$E(m(X_i^*)|X_i = x, D_i = 0) = \frac{\int_{-\infty}^{x_0} m(x^*) f_{X^*|X}(x^*|x) dx^*}{\int_{-\infty}^{x_0} f_{X^*|X}(x^*|x) dx^*}. \quad (22)$$

Consequently, we obtain the desired result by (16), (17), (19), (20), (21), and (22).

□

A.2 Proof of Theorem 2

In this proof, for generic A and B, we write $A = B + o(\sigma^2)$ by $A \approx B$ for notational simplicity. We first show that $E(Y_i|X_i^* = x_0+, D_i = 1) \approx \mu(x_0+, 1, \sigma)$. Fix $\varepsilon > 0$. Under Assumptions 3 and 4, there exists some $e > 0$ such that

$$\begin{aligned} &|E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0 + e, D_i = 1)| < \varepsilon, \\ &|\mu(x_0+, 1, \sigma) - \mu(x_0 + e, 1, \sigma)| < \varepsilon. \end{aligned}$$

By the triangle inequality, it thus holds that

$$
\begin{aligned}
&|E(Y_i|X_i^* = x_0+, D_i = 1) - \mu(x_0+, 1, \sigma)| \\
\leq &|E(Y_i|X_i^* = x_0+, D_i = 1) - E(Y_i|X_i^* = x_0 + e, D_i = 1)| \\
&+ |E(Y_i|X_i^* = x_0 + e, D_i = 1) - \mu(x_0 + e, 1, \sigma)| + |\mu(x_0 + e, 1, \sigma) - \mu(x_0+, 1, \sigma)| \\
< &|E(Y_i|X_i^* = x_0 + e, D_i = 1) - \mu(x_0 + e, 1, \sigma)| + 2\varepsilon.
\end{aligned}
$$

Hence, we obtain the desired result if we show

$$
E(Y_i|X_i^* = x_0 + e, D_i = 1) \approx \mu(x_0 + e, 1, \sigma), \tag{23}
$$

for any $e > 0$, because $\varepsilon > 0$ is arbitrary. The proof of (23) is similar to that in Chesher (1991). We set $x = x_0 + e$ for notational simplicity.

To prove (23), we calculate an approximation for $f_{Y|XD}(y|x, 1)$. To this end, we first compute the approximation for $f_{YX|D}(y, x|1)$. For any y and u, we have

$$
\begin{aligned}
f_{YXU|D}(y, x, u|1) &= f_{YX^*U|D}(y, x - \sigma u, u|1) \\
&= f_{Y|X^*UD}(y|x - \sigma u, u, 1) f_{X^*U|D}(x - \sigma u, u|1) \\
&= f_{Y|X^*D}(y|x - \sigma u, 1) f_{X^*|D}(x - \sigma u|1) f_U(u),
\end{aligned}
$$

where the second equality follows from Bayes' theorem, and the third equality follows from Assumption 1. Applying Taylor's theorem around $\sigma = 0$, it holds that for sufficiently small $\sigma > 0$

$$
\begin{aligned}
&f_{YXU|D}(y, x, u|1) \\
&\approx f_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) f_U(u) \\
&- \sigma u f_U(u) \left\{ f^{(1)}_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) + f_{Y|X^*D}(y|x, 1) f^{(1)}_{X^*|D}(x|1) \right\} \\
&+ \frac{1}{2}\sigma^2 u^2 f_U(u) \left\{ f^{(2)}_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) + 2 f^{(1)}_{Y|X^*D}(y|x, 1) f^{(1)}_{X^*|D}(x|1) + f_{Y|X^*D}(y|x, 1) f^{(2)}_{X^*|D}(x|1) \right\},
\end{aligned}
$$

under Assumptions 1 and 4. Integrating the both sides with respect to u, we have

$$
\begin{aligned}
f_{YX|D}(y, x|1) \approx &f_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) \\
&+ \frac{\sigma^2}{2} \left\{ f^{(2)}_{Y|X^*D}(y|x, 1) f_{X^*|D}(x|1) \right. \\
&\left. + 2 f^{(1)}_{Y|X^*D}(y|x, 1) f^{(1)}_{X^*|D}(x|1) + f_{Y|X^*D}(y|x, 1) f^{(2)}_{X^*|D}(x|1) \right\},
\end{aligned} \tag{24}
$$

by Assumptions 1 and 4.

Similar to Equation (2·5) in Chesher (1991), we have the approximation of $1/f_{X|D}(x|1)$ for sufficiently small $\sigma > 0$:

$$
\frac{1}{f_{X|D}(x|1)} \approx \frac{1}{f_{X^*|D}(x|1)} - \frac{\sigma^2}{2} \frac{f^{(2)}_{X^*|D}(x|1)}{(f_{X^*|D}(x|1))^2}, \tag{25}
$$

under Assumptions 1 and 4.

Therefore, multiplying (24) by (25), we have

$$f_{Y|XD}(y|x,1) \approx f_{Y|X^*D}(y|x,1) + \frac{\sigma^2}{2}\left\{2f^{(1)}_{Y|X^*D}(y|x,1)\left(\log^{(1)} f_{X^*|D}(x|1)\right) + f^{(2)}_{Y|X^*D}(y|x,1)\right\}.$$

Furthermore, under Assumption 4, this approximation leads to

$$f^{(s)}_{Y|XD}(y|x,1) = f^{(s)}_{Y|X^*D}(y|x,1) + O(\sigma^2),$$

for $s = 1, 2$. We thus have

$$f_{Y|XD}(y|x,1) \approx f_{Y|X^*D}(y|x,1) + \frac{\sigma^2}{2}\left\{2f^{(1)}_{Y|XD}(y|x,1)\left(\log^{(1)} f_{X^*|D}(x|1)\right) + f^{(2)}_{Y|XD}(y|x,1)\right\}.$$

This leads to

$$\begin{aligned}E(Y_i|X_i = x, D_i = 1) \approx{}& E(Y_i|X_i^* = x, D_i = 1) + \sigma^2\left(\log^{(1)} f_{X^*|D}(x|1)\right)E^{(1)}(Y_i|X_i = x, D_i = 1)\\ &+ \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i = x, D_i = 1),\end{aligned}$$

under Assumptions 1 and 3–5. Thus, we have

$$\begin{aligned}E(Y_i|X_i^* = x, D_i = 1) \approx{}& E(Y_i|X_i = x, D_i = 1) - \sigma^2\left(\log^{(1)} f_{X^*|D}(x|1)\right)E^{(1)}(Y_i|X_i = x, D_i = 1)\\ &- \frac{\sigma^2}{2}E^{(2)}(Y_i|X_i = x, D_i = 1),\end{aligned}$$

It holds that

$$\begin{aligned}f_{X^*|D}(x|1) &= f_{X|D}(x|1) + O(\sigma^2),\\ f^{(1)}_{X^*|D}(x|1) &= f^{(1)}_{X|D}(x|1) + O(\sigma^2),\end{aligned}$$

under Assumptions 1, 4, and 5 similar to Equation (2·4) in Chesher (1991). Therefore, it holds that

$$E(Y_i|X_i^* = x, D_i = 1) \approx \mu(x, 1, \sigma).$$

Accordingly, we obtain (23) and show $E(Y_i|X_i^* = x_0+, D_i = 1) \approx \mu(x_0+, 1, \sigma)$.

Similarly, we can show that

$$E(Y_i|X_i^* = x_0-, D_i = 0) \approx \mu(x_0-, 0, \sigma).$$

Consequently, we obtain the desired result.

□

A.3 Proof of Theorem 3

By Lemma 1, the consistency of $\hat{g}(x_0, d)$ for $g(x_0, d)$, and Slutsky's theorem, it holds that

$$\sqrt{nh}h^2\left(\hat{\mu}\left(x_0+, 1, \sigma\right) - \mu\left(x_0+, 1, \sigma\right)\right) \rightsquigarrow N\left(-\sigma^2 e_3' B^+, \sigma^4 e_3' \Omega^+ e_3\right), \tag{26}$$

$$\sqrt{nh}h^2\left(\hat{\mu}\left(x_0-, 0, \sigma\right) - \mu\left(x_0-, 0, \sigma\right)\right) \rightsquigarrow N\left(-\sigma^2 e_3' B^-, \sigma^4 e_3' \Omega^- e_3\right). \tag{27}$$

Because the left-hand sides of (26) and (27) are independent, we obtain the desired result using the continuous mapping theorem.

□

A.4 Proof of Lemma 1

In this proof, we denote a generic constant as C. We only provide the proof of (11), because that of (12) is analogous. The proof is an extension of those of Hahn, Todd, and van der Klaauw (1999) and Porter (2003). The minimization problem (10) is rewritten as

$$\min_{(a,b')'\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} I_i D_i (Y_i^* - (a-\alpha^+) - (b_1-\beta_1^+)(X_i-x_0) - \cdots - (b_p-\beta_p^+)(X_i-x_0)^p)^2 K\left(\frac{X_i-x_0}{h}\right),$$

where

$$\begin{aligned}
Y_i^{*+} &:= Y_i - \alpha^+ - \beta_1^+(X_i-x_0) - \cdots - \beta_p^+(X_i-x_0)^p, \\
\alpha^+ &:= E(Y_i|X_i = x_0+, D_i=1), \qquad \beta_k^+ := \frac{1}{k!}E^{(k)}(Y_i|X_i=x_0+, D_i=1) \quad \text{for } k=1,\ldots,p+1.
\end{aligned}$$

Define

$$\begin{aligned}
\tilde{Z}_i &:= (1, X_i-x_0, \ldots, (X_i-x_0)^p)', \qquad \tilde{Z} := (\tilde{Z}_1,\ldots,\tilde{Z}_n)', \\
r &:= (a, b_1, \ldots, b_p)', \qquad \gamma^+ := (\alpha^+, \beta_1^+, \ldots, \beta_p^+)', \qquad \hat{\gamma}^+ := (\hat{\alpha}^+, \hat{\beta}_1^+, \ldots, \hat{\beta}_p^+)' \\
Y^{*+} &:= (Y_1^{*+},\ldots,Y_n^{*+})', \qquad A_h^+ := diag\left(K\left(\frac{X_1-x_0}{h}\right) I_1 D_1, \ldots, K\left(\frac{X_n-x_0}{h}\right) I_n D_n\right).
\end{aligned}$$

Then, the minimization problem is

$$\begin{aligned}
&\operatorname*{argmin}_{r\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} I_i D_i \left(Y_i^{*+} - (r-\gamma^+)'\tilde{Z}_i\right)^2 K\left(\frac{X_i-x_0}{h}\right) \\
&= \operatorname*{argmin}_{r\in\mathbb{R}^{p+1}} \left(Y^{*+} - \tilde{Z}(r-\gamma^+)\right)' A_h^+ \left(Y^{*+} - \tilde{Z}(r-\gamma^+)\right).
\end{aligned}$$

By the first-order condition, we have

$$\begin{aligned}
\hat{\gamma}^+ - \gamma^+ &= (\tilde{Z}' A_h^+ \tilde{Z})^{-1} \tilde{Z}' A_h^+ Y^{*+} \\
&= H(H\tilde{Z}' A_h^+ \tilde{Z} H)^{-1} H\tilde{Z}' A_h^+ Y^{*+} \\
&= H(Z' A_h^+ Z)^{-1} Z' A_h^+ Y^{*+},
\end{aligned}$$

where $H := diag(1, h^{-1}, \ldots, h^{-p})$, $Z := (Z_1,\ldots,Z_n)'$, and $Z_i := (1, (X_i-x_0)/h, \ldots, (X_i-x_0)^p/h^p)'$. It holds that

$$\begin{aligned}
H^{-1}(\hat{\gamma}^+ - \gamma^+) &= (Z' A_h^+ Z)^{-1} Z' A_h^+ Y^{*+} \\
&= \left(\frac{1}{nh}\sum_{i=1}^{n} Z_i Z_i' K\left(\frac{X_i-x_0}{h}\right) I_i D_i\right)^{-1} \frac{1}{nh}\sum_{i=1}^{n} Z_i Y_i^{*+} K\left(\frac{X_i-x_0}{h}\right) I_i D_i.
\end{aligned}$$

Therefore, we have the following decomposition:

$$
\begin{aligned}
&\sqrt{nh}H^{-1}(\hat{\gamma}^+ - \gamma^+) \\
&= \left(\frac{1}{nh}\sum_{i=1}^{n} Z_i Z_i' K_h(X_i) I_i D_i\right)^{-1} && (28)\\
&\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right)\right)\right. && (29)\\
&+ \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right) - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\right) && (30)\\
&+ \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right) - \tilde{B}_n^+ && (31)\\
&\left. + \tilde{B}_n^+\right), && (32)
\end{aligned}
$$

where $\tilde{B}_n^+ \coloneqq \sqrt{nh}h^{p+1}\beta_{p+1} f_X(x_0) E(D_i|X_i = x_0+)(\gamma_{p+1}, \ldots, \gamma_{2p+1})'$. In the following, we study each term separately. Term (28) is shown to converge in probability to some constant. Term (29) is shown to converge in distribution to the normal distribution. Terms (30) and (31) are shown to be asymptotically negligible. The multiplication of (28) with (32) converges to B^+ as $\sqrt{nh}h^{p+1} \to \tilde{C} \in [0, \infty)$ by the following proof.

Term (28): We show that

$$
\left(\frac{1}{nh}\sum_{i=1}^{n} Z_i Z_i' K_h(X_i) I_i D_i\right)^{-1} \xrightarrow{p} \frac{1}{E(D_i|X_i = x_0+) f_X(x_0)} (\Gamma^+)^{-1}, \tag{33}
$$

where $\Gamma^+ \coloneqq (\gamma_{k+l-2})_{(k,l)}$ for $k, l = 1, \ldots, p+1$ and $\gamma_q \coloneqq \int_0^M u^q K(u) du$ for $q = 0, \ldots, 2p$. For $q = 0, \ldots, 2p$,

$$
\begin{aligned}
&E\left(\frac{1}{nh}\sum_{i=1}^{n}\left(\frac{X_i - x_0}{h}\right)^q K_h(X_i) I_i D_i\right) \\
&= h^{-1} E\left(\left(\frac{X_i - x_0}{h}\right)^q K_h(X_i) I_i D_i\right) \\
&= h^{-1} E\left(\left(\frac{X_i - x_0}{h}\right)^q K_h(X_i) I_i E(D_i|X_i)\right) \\
&= h^{-1}\int_{x_0}^{x_0+Mh}\left(\frac{x - x_0}{h}\right)^q K\left(\frac{x - x_0}{h}\right) E(D_i|X_i = x) f_X(x) dx \\
&= \int_0^M u^q K(u) E(D_i|X_i = x_0 + uh) f_X(x_0 + uh) du \\
&= E(D_i|X_i = x_0+) f_X(x_0)\int_0^M u^q K(u) du + o(1) \\
&= E(D_i|X_i = x_0+) f_X(x_0)\gamma_q + o(1),
\end{aligned}
$$

where the first equality follows from the i.i.d. assumption, the second equality follows from the law of iterated expectations, the third equality follows from the definition of I_i and Assumption

6, and the fifth equality follows from Assumptions 7 and 8 and the dominated convergence theorem. We also have

$$\begin{aligned} var\left(\frac{1}{nh}\sum_{i=1}^{n}\left(\frac{X_i-x_0}{h}\right)^q K_h(X_i)I_iD_i\right) &= \frac{1}{nh^2}var\left(\left(\frac{X_i-x_0}{h}\right)^q K_h(X_i)I_iD_i\right) \\ &\leq \frac{1}{nh^2}E\left(\left(\frac{X_i-x_0}{h}\right)^{2q} K_h^2(X_i)I_iD_i\right) \\ &\leq \frac{1}{nh^2}E\left(\left(\frac{X_i-x_0}{h}\right)^{2q} K_h^2(X_i)I_i\right) \\ &= \frac{1}{nh}\int_0^M u^{2q}K^2(u)f_X(x_0+uh)du \\ &= \frac{1}{nh}f_X(x_0)\int_0^M u^{2q}K^2(u)du + o\left(\frac{1}{nh}\right) \\ &= o(1), \end{aligned}$$

where the first equality follows from the i.i.d. assumption, the second inequality follows from $D_i \leq 1$, the second equality follows from Assumption 6, and the third equality follows from Assumptions 6 and 7 and from the dominated convergence theorem. Therefore, we have shown (33) using Markov's inequality and the continuous mapping theorem.

Term (29): We show that

$$\begin{aligned} &\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(Z_iY_i^{*+}K_h(X_i)I_iD_i - E\left(Z_iY_i^{*+}K_h(X_i)I_iD_i\middle|X_i,D_i\right)\right) \\ &\rightsquigarrow N\left(0, E(V_i^2|X_i=x_0+,D_i=1)E(D_i|X_i=x_0+)f_X(x_0)\Delta^+\right), \end{aligned} \tag{34}$$

where $\Delta^+ \coloneqq (\delta_{k+l-2})_{(k,l)}$ with $k,l=1,\ldots,p+1$ and $\delta_l \coloneqq \int_0^M u^l K^2(u)du$ for $l=0,\ldots,2p$. To this end, we use the Cramer–Wald device. We observe that

$$\begin{aligned} &\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(Z_iY_i^{*+}K_h(X_i)I_iD_i - E\left(Z_iY_i^{*+}K_h(X_i)I_iD_i\middle|X_i,D_i\right)\right) \\ &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} Z_iI_iK_h(X_i)D_i(Y_i^{*+} - E(Y_i^{*+}|X_i,D_i)) \\ &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} Z_iI_iK_h(X_i)D_i(Y_i - E(Y_i|X_i,D_i)) \\ &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} Z_iI_iK_h(X_i)D_iV_i, \end{aligned}$$

where $V_i \coloneqq Y_i - E(Y_i|D_i,X_i)$. Let λ be a nonzero finite vector and

$$U_{n,i} \coloneqq \frac{1}{\sqrt{nh}}\lambda' Z_iI_iK\left(\frac{X_i-x_0}{h}\right)D_iV_i.$$

By the law of iterated expectations, $E(U_{n,i}) = 0$. For the variance, it holds that for $l = 0, \ldots, 2p$,

$$
\begin{aligned}
&E\left(\frac{1}{nh}\left(\frac{X_i - x_0}{h}\right)^l I_i K_h^2(X_i) D_i V_i^2\right) \\
&= \frac{1}{nh} E\left(\left(\frac{X_i - x_0}{h}\right)^l I_i K_h^2(X_i) E(D_i V_i^2 | X_i)\right) \\
&= \frac{1}{nh} E\left(\left(\frac{X_i - x_0}{h}\right)^l I_i K_h^2(X_i) E(V_i^2 | X_i, D_i = 1) E(D_i | X_i)\right) \\
&= \frac{1}{nh} \int_{x_0}^{x_0 + Mh} \left(\frac{x - x_0}{h}\right)^l K^2\left(\frac{x - x_0}{h}\right) E(V_i^2 | X_i = x, D_i = 1) E(D_i | X_i = x) f_X(x) dx \\
&= \frac{1}{n} \int_0^M u^l K^2(u) E(V_i^2 | X_i = x_0 + hu, D_i = 1) E(D_i | X_i = x_0 + hu) f_X(x_0 + hu) du \\
&= \frac{1}{n} E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \int_0^M u^l K^2(u) du + o(n), \\
&= \frac{1}{n} E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \delta_l + o(n),
\end{aligned}
$$

where the first equality follows from the law of iterated expectations and the fifth equality follows from Assumptions 6–9 and the dominated convergence theorem. Thus, we have

$$
\sum_{i=1}^n var(U_{n,i}) \to E(V_i^2 | X_i = x_0+, D_i = 1) E(D_i | X_i = x_0+) f_X(x_0) \lambda' \Delta^+ \lambda.
$$

We next check Lyapunov's condition. Considering some $\zeta > 0$, it holds that

$$
\begin{aligned}
&\sum_{i=1}^n E|U_{n,i}|^{2+\zeta} \\
&= \sum_{i=1}^n \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \frac{1}{nh} E\left(|\lambda' Z_i|^{2+\zeta} \left|K\left(\frac{X_i - x_0}{h}\right)\right|^{2+\zeta} I_i D_i |V_i|^{2+\zeta}\right) \\
&\leq \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \frac{1}{h} E\left(|\lambda' Z_i|^{2+\zeta} \left|K\left(\frac{X_i - x_0}{h}\right)\right|^{2+\zeta} I_i |V_i|^{2+\zeta}\right) \\
&\leq C\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \frac{1}{h} E\left(\sum_{l=0}^p \left|\lambda_l \left(\frac{X_i - x_0}{h}\right)^l\right|^{2+\zeta} \left|K\left(\frac{X_i - x_0}{h}\right)\right|^{2+\zeta} I_i E(|V_i|^{2+\zeta} | X_i)\right) \\
&\leq C\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}} \left(\sup_{x \in [0, x_0 + Mh]} E(|V_i|^{2+\zeta} | X_i = x)\right) \int_0^M |K(u)|^{2+\zeta} f_X(x_0 + uh) \sum_{l=0}^p |\lambda_l u^l|^{2+\zeta} du \\
&= O\left(\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\right) = o(1),
\end{aligned}
$$

where the first inequality follows from $D_i \leq 1$, the second inequality follows from the law of iterated expectations and Loéve's C_r inequality, and the second equality follows from Assumptions 6, 7, and 9. Therefore, by Lyapunov CLT and the Cramer–Wald device, we have shown (34).

Term (30): We show that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\left(\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right) - E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\right) \\ &= O_p(h^{p+1}) = o_p(1).\end{aligned} \tag{35}$$

To this end, we first define

$$\mu_j^+(x) := E(Y_i | X_i = x, D_i = 1) - \left(\alpha^+ + \beta_1^+(x - x_0) + \cdots + \beta_j^+(x - x_0)^j\right),$$

for $j \in \mathbb{N}$, and we use

$$\frac{1}{h^{p+1}} \sup_{x \in (x_0, x_0 + Mh]} \left|\mu_{p+1}^+(x)\right| = o(1), \tag{36}$$

by Assumption 10 and Taylor's theorem. It is clear that the expectation of the left-hand side of (35) is zero by the law of iterated expectations. The variance is

$$\begin{aligned}var\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right)\right) &= \frac{1}{h} var\left(E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i \middle| X_i, D_i\right)\right) \\ &\leq \frac{1}{h} E\left(E\left(\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)^2 \middle| X_i, D_i\right)\right) \\ &\leq \frac{1}{h} E\left(Z_i^2 (Y_i^{*+})^2 K_h^2(X_i) I_i\right),\end{aligned}$$

where the first equality follows from the i.i.d. assumption and the second inequality follows from the law of iterated expectations. For the elements, for $l = 0, \ldots, p$, we have

$$\begin{aligned}&\frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} (Y_i^{*+})^2 K_h^2(X_i) I_i\right) \\ =&\frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} (\mu_{p+1}^+(X_i) + \beta_{p+1}(X_i - x_0)^{p+1})^2 K_h^2(X_i) I_i\right) \\ \leq& C\frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} \left(\left(\mu_{p+1}^+(X_i)\right)^2 + \beta_{p+1}^2 (X_i - x_0)^{2(p+1)}\right) K_h^2(X_i) I_i\right) \\ \leq& C\left(\sup_{x \in (x_0, x_0 + Mh]} \left(\mu_{p+1}^+(x)\right)^2\right) \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} K_h^2(X_i) I_i\right) \\ &+ C\beta_{p+1}^2 \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} (X_i - x_0)^{2(p+1)} K_h^2(X_i) I_i\right) \\ =& Ch^{2(p+1)} \left(\frac{1}{h^{p+1}} \sup_{x \in (x_0, x_0 + Mh]} \left|\mu_{p+1}^+(x)\right|\right)^2 \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2l} K_h^2(X_i) I_i\right) \\ &+ Ch^{2(p+1)} \beta_{p+1}^2 \frac{1}{h} E\left(\left(\frac{X_i - x_0}{h}\right)^{2(l+p+1)} K_h^2(X_i) I_i\right) \\ =& o\left(h^{2(p+1)}\right) + O\left(h^{2(p+1)}\right),\end{aligned}$$

where the first inequality follows from Loéve's C_r inequality and where the last equality follows from Assumptions 6 and 7 and (36). Therefore, (35) holds by Chebyshev's inequality.

Term (31): We show that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\\&-\sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)(\gamma_{p+1},\ldots,\gamma_{2p+1})' = o(1).\end{aligned} \tag{37}$$

We first observe that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(Z_i Y_i^{*+} K_h(X_i) I_i D_i\right)\\=&\frac{\sqrt{n}}{\sqrt{h}} E\left(Z_i K_h(X_i) I_i E\left(Y_i^{*+} D_i|X_i, D_i\right)\right)\\=&\frac{\sqrt{n}}{\sqrt{h}} E\left(Z_i K_h(X_i) I_i E\left(Y_i^{*+}|X_i, D_i = 1\right) E(D_i|X_i)\right)\\=&\frac{\sqrt{n}}{\sqrt{h}} E\left(Z_i K_h(X_i) I_i \left(\mu_{p+1}^{+}(X_i) + \beta_{p+1}(X_i - x_0)^{p+1}\right) E(D_i|X_i)\right),\end{aligned}$$

where the first and second equalities follow from the law of iterated expectations. Then, for $l = 0,\ldots,p$, we have

$$\begin{aligned}&\left|\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} E\left(\left(\frac{X_i - x_0}{h}\right)^l Y_i^{*+} K_h(X_i) I_i D_i\right) - \sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right|\\=&\left|\frac{\sqrt{n}}{\sqrt{h}} E\left(\left(\frac{X_i - x_0}{h}\right)^l K_h(X_i) I_i \left(\mu_{p+1}^{+}(X_i) + \beta_{p+1}^{+}(X_i - x_0)^{p+1}\right) E(D_i|X_i)\right)\right.\\&\left.- \sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right|\\\le&\left|\frac{\sqrt{n}}{\sqrt{h}} h^{p+1}\beta_{p+1}^{+} E\left(\left(\frac{X_i - x_0}{h}\right)^{l+p+1} K_h(X_i) I_i E(D_i|X_i)\right)\right.\\&\left.- \sqrt{nh}h^{p+1}\beta_{p+1}^{+} f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right|\\&+\left|\frac{\sqrt{n}}{\sqrt{h}} E\left(\left(\frac{X_i - x_0}{h}\right)^l K_h(X_i) I_i \mu_{p+1}^{+}(X_i) E(D_i|X_i)\right)\right|\\\le&\left|\sqrt{nh}h^{p+1}\beta_{p+1}^{+}\left(\int_0^M u^{l+p+1} K(u) E(D_i|X_i = x_0 + uh) f_X(x_0 + uh) du\right.\right.\\&\left.\left.- f_X(x_0) E(D_i|X_i = x_0+)\gamma_{l+p+1}\right)\right|\\&+\left|\sqrt{nh}h^{p+1}\left(\frac{1}{h^{p+1}}\sup_{x\in(x_0,x_0+Mh]}\left|\mu_{p+1}^{+}(x)\right|\right)\int_0^M u^l K(u) E(D_i|X_i = x_0 + uh) f_X(x_0 + uh) du\right|\\=&o(\sqrt{nh}h^{p+1}) = o(1),\end{aligned}$$

where the first inequality follows from the triangle inequality and the last equality follows from Assumptions 6–8 and the dominated convergence theorem. Thus, we obtain (37).

Consequently, we have the desired result by Slutsky's theorem.

□

References

J. D. Angrist and V. Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

Y. Arai and H. Ichimura. Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator. *GRIPS Discussion Paper 14-03, National Graduate Institute for Policy Studies*, 2014.

E. Battistin and A. Chesher. Treatment effect estimation with covariate measurement error. *Journal of Econometrics*, 178(2):707–715, 2014.

E. Battistin, A. Brugiavini, E. Rettore, and G. Weber. The retirement consumption puzzle: evidence from a regression discontinuity approach. *The American Economic Review*, 99(5): 2209–2226, 2009.

J. Bound, C. Brown, and N. Mathiowetz. Measurement error in survey data. *Handbook of Econometrics*, 5:3705–3843, 2001.

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Forthcoming in Econometrica*, 2014.

D. Card and L. D. Shore-Sheppard. Using discontinuous eligibility rules to identify the effects of the federal medicaid expansions on low-income children. *Review of Economics and Statistics*, 86(3):752–766, 2004.

D. Card, C. Dobkin, and N. Maestas. The impact of nearly universal insurance coverage on health care utilization: Evidence from medicare. *American Economic Review*, 98(5):2242–2258, 2008.

A. Chesher. The effect of measurement error. *Biometrika*, 78(3):451–462, 1991.

A. Chesher and C. Schluter. Welfare measurement and measurement error. *The Review of Economic Studies*, 69(2):357–378, 2002.

Y. Dong. Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics*, 2014.

J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036, 1992.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. CRC Press, 1996.

B. R. Frandsen, M. Frölich, and B. Melly. Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, 168(2):382–395, 2012.

J. Hahn, P. Todd, and W. van der Klaauw. Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. *NBER Working Paper*, 7131:1–32, 1999.

J. Hahn, P. Todd, and W. van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

P. Hullegie and T. J. Klein. The effect of private health insurance on medical care utilization and self-assessed health in germany. *Health Economics*, 19(9):1048–1062, 2010.

G. Imbens and K. Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79:933–959, 2012.

G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

M. Jones. On kernel density derivative estimation. *Communications in Statistics-Theory and Methods*, 23(8):2133–2139, 1994.

T. G. Koch. Using rd design to understand heterogeneity in health insurance crowd-out. *Journal of Health Economics*, 32(3):599–611, 2013.

D. S. Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.

D. S. Lee and D. Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

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

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

E. Masry. Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101, 1996b.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008.

Z. Pei. Regression discontinuity design with measurement error in the assignment variable. mimeo, 2011.

J. Porter. Estimation in the regression discontinuity model. mimeo, 2003.

D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, pages 1346–1370, 1994.

D. W. Schanzenbach. Do school lunches contribute to childhood obesity? *Journal of Human Resources*, 44(3):684–709, 2009.

S. M. Schennach. Measurement error in nonlinear models - a review. *Advances in Economics and Econometrics Tenth World Congress*, 3(51):296–337, 2013.

B. W. Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184, 1978.

D. L. Thistlethwaite and D. T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.

P. Yu. Identification of treatment effects in regression discontinuity designs with measurement error. mimeo, 2012.