2018

# Three essays on regression discontinuity design and partial identification

Yang He
*Iowa State University*

**Three essays on regression discontinuity design and partial identification**

by

**Yang He**

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:
Otávio Bartalotti, Co-major Professor
Brent Kreider, Co-major Professor
Gray Calhoun
Cindy Yu
Oleksandr Zhylyevskyy

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

# DEDICATION

Dedicated to my parents, Hongzhuan and Min, who brought me to the world and gave me the strength to chase my dream. Dedicated to my fiancée, Rui, who came into and lighted up my life. I love you with all my heart.

# TABLE OF CONTENTS

**Page**

# LIST OF TABLES

**Page**

# LIST OF FIGURES

**Page**

# ACKNOWLEDGMENTS

I thank all my professors, the sources of knowledge, wisdom and guidance. Without their support I would not have been able to complete this work. I owe my deepest gratitude to Dr. Kreider, Dr. Bartalotti and Dr. Calhoun for their scholarly guidance, and most importantly, for their trust that gives me strength and courage to push the limit of my capability. I am deeply indebted to Dr. Zhylyevskyy and Dr. Yu who provide constructive comments and innovative ideas during the completion of this work. I would additionally like to thank Dr. Bunzel and Dr. Huffman for enlightening discussions and all the staff in my department for their assistance.

# ABSTRACT

This dissertation consists of three chapters on regression discontinuity (RD) design and partial identification, which are widely used techniques in program evaluation.

The first and the second chapters discuss statistic inference for the treatment effect estimator in fuzzy RD designs. Fuzzy RD design and instrumental variables (IV) regression share similar identification strategies and numerically yield the same results under certain conditions. While the weak identification problem is widely recognized in IV regressions, it has drawn much less attention in fuzzy RD designs, where the standard t-test can also suffer from asymptotic size distortions and the confidence interval obtained by inverting such a test becomes invalid. I explicitly model fuzzy RD designs in parallel with IV regressions, and based on the extensive literature of the latter, develop tests which are robust to weak identification in fuzzy RD designs, including the Anderson-Rubin (AR) test, the Lagrange multiplier (LM) test, and the conditional likelihood ratio (CLR) test. These tests have correct size regardless of the strength of identification and their power properties are similar to those in IV regressions. Due to the similarities between a fuzzy RD design and an IV regression, one can choose either method for estimation and inference. However, it is shown that adopting a fuzzy RD design with newly proposed tests has the potential to achieve more power without introducing size distortions in hypothesis testing and is thus recommended. An extension to testing for quantile treatment effects in fuzzy RD designs is also discussed.

RD estimators are usually estimated with nonparametric methods and have bias. A new wild bootstrap procedure is proposed to correct bias and construct valid confidence intervals in fuzzy regression discontinuity designs. This procedure uses a wild bootstrap based on second order local polynomials to estimate and remove the bias from linear models. The bias-corrected estimator is then bootstrapped itself to generate valid confidence intervals. While the conventional confidence intervals generated by adopting MSE-optimal bandwidth is asymptotically not valid, the confidence

intervals generated by this procedure have correct coverage under conditions similar to Calonico, Cattaneo and Titiunik's(2014, *Econometrica*) analytical correction. Simulation studies provide evidence that this new method is as accurate as the analytical corrections when applied to a variety of data generating processes featuring heteroskedasticity, endogeneity and clustering. As an example, its usage is demonstrated through a reanalysis of the scholastic achievement data used by Angrist and Lavy (1999).

In the third chapter, a novel numerical approach is proposed to partially identify treatment effects. Endogenous treatment and measurement error are very common in survey data and pose threats to reliable estimation of treatment effects. The new approach considers these two issues simultaneously and provides bounds for treatment effects. Conceptually, treatment effects and model assumptions are formulated as linear restrictions on a large set of probability mass. One can then check if any given treatment effect is consistent with model assumptions and observed data. Compared with previous methods, the newly proposed numerical approach is general enough to be applied to various different problems and guarantees sharp bounds. An example is provided to show that how the distribution of a treatment effect and how the averages of multiple treatment effects can be partially identified through this approach.

# CHAPTER 1. ROBUST INFERENCE IN FUZZY REGRESSION DISCONTINUITY DESIGNS

Fuzzy regression discontinuity (RD) design and instrumental variables (IV) regression share similar identification strategies and numerically yield the same results under certain conditions. While the weak identification problem is widely recognized in IV regressions, it has drawn much less attention in fuzzy RD designs, where the standard t-test can also suffer from asymptotic size distortions and the confidence interval obtained by inverting such a test becomes invalid. I explicitly model fuzzy RD designs in parallel with IV regressions, and based on the extensive literature of the latter, develop tests which are robust to weak identification in fuzzy RD designs, including the Anderson-Rubin (AR) test, the Lagrange multiplier (LM) test, and the conditional likelihood ratio (CLR) test. These tests have correct size regardless of the strength of identification and their power properties are similar to those in IV regressions. Due to the similarities between a fuzzy RD design and an IV regression, one can choose either method for estimation and inference. However, it is shown that adopting a fuzzy RD design with newly proposed tests has the potential to achieve more power without introducing size distortions in hypothesis testing and is thus recommended. An extension to testing for quantile treatment effects in fuzzy RD designs is also discussed.

## 1.1 Introduction

Regression discontinuity (RD) design is a very popular way of estimating the causal effect of an endogenous treatment on various outcomes. Since the early work by Hahn et al. (2001) and Porter (2003), studies in this field have been growing fast. For example, some recent advances include design validity (McCrary, 2008; Barreca et al., 2016), bandwidth selection (Imbens and Kalyanaraman, 2012; Arai and Ichimura, 2013; Gelman and Imbens, 2017), statistical inference (Lee and Card, 2008; Calonico et al., 2014; Card et al., 2015b; Otsu et al., 2015; Bartalotti et al.,

2017a; Bartalotti and Brummet, 2017; Chiang et al., 2017), quantile treatment effects (Frandsen et al., 2012; Qu et al., 2015; Chiang and Sasaki, 2016), etc. A comprehensive review can be found in Imbens and Lemieux (2008) and Lee and Lemieux (2010). In a canonical RD design, the treatment probability conditional on a covariate experiences a discontinuity and is thought to be exogenously induced by policy rules governing the treatment assignment based on the covariate. The fact that this discontinuity in treatment probability is mirrored in average outcome allows researchers to identify the causal treatment effect. For example, in the first application by Thistlethwaite and Campbell (1960), the Certificates of Merit was given to students largely based on a qualifying score. The probability of a student receiving this award is zero if (s)he scores below a certain threshold, but jumps to around 3.4% if marginally passes this threshold. Thus, the jump in treatment probability can be used to study the causal effect of Certificates of Merit on future outcomes such as career aspirations.

The treatment effect in a RD design is usually identified as the ratio of discontinuities in the average outcome and the treatment probability. The term "fuzzy" is used to describe those RD designs where the jump in the treatment probability is less than one and the term "sharp" is used for those where the jump in the treatment probability is exactly one. Unlike sharp RD designs, fuzzy RD designs may have a weak identification problem where the standard t-test, as well as the confidence interval obtained by inverting the t-test, becomes unreliable. One can get the intuition from the analogy between a fuzzy RD design and an IV regression model, which is widely known to have a weak identification problem; that is, the convergence of a standard t-test statistic to a normal distribution is not uniform with respect to the correlation between the IV and the endogenous variable (Mikusheva et al., 2013). The fact that the treatment effect estimator in a fuzzy RD design could be numerically the same as that in an IV regression model under certain conditions suggests that weak identification can also happen in fuzzy RD designs. In practice, this problem could be exacerbated because only a fraction of the data is actually used for estimation. Feir et al. (2016) investigated a set of influential applied papers that use fuzzy RD designs and found that "weak identification appears to be a problem in at least one of the empirical

specifications" for half of the articles where enough information is reported. Though there has been tremendous development in the weak identification literature on IV regression models (Stock et al., 2002; Dufour, 2003; Mikusheva et al., 2013), the weak identification in the context of fuzzy RD designs is not well recognized. Feir et al. (2016) seems to be the only published study on statistical inference robust to weak identification in fuzzy RD designs.

In this article, I draw on insights from the weak identification literature on IV regression models and show that many widely used tests such as the Anderson-Rubin (AR) test, the Lagrange multiplier (LM) test and the likelihood ratio (LR) test can be adapted to fuzzy RD designs. This is achieved by explicitly modeling fuzzy RD designs in parallel with IV regressions. In particular, the relevance condition is captured by discontinuities in the treatment probability, and the strength of identification in a fuzzy RD design depends on not only the magnitude of the discontinuity, but also on how precisely it can be estimated. Standard inference may become unreliable when the discontinuity is small in magnitude, or of moderate size but can only be estimated with excessive noise. The goal is to develop valid tests even in the case of weak identification.

To build a theoretical model, I start by assuming that discontinuities in the treatment probability and the average outcome are observable random variables. They follow normal distribution with known covariance. Tests are then developed for jointly testing the treatment effect and its derivative following the regression probability jump and kink (RPJK) framework (Dong, 2016). I show that AR, LM, and LR statistics in this case are equivalent and have pivotal null distribution. Tests which utilize these statistics and critical values defined by percentiles of the null distribution are similar — that is, they have the same null rejection probability regardless of the value of nuisance parameters — a crucial property of the tests in order to be robust to weak identification.

I demonstrate how to apply the proposed tests when extra information on the treatment effect derivative is available. In particular, if the treatment effect derivative is known (or reasonably assumed), then the treatment effect is over identified. It is shown that both AR and LM statistics have Chi-square distribution with different degrees of freedom under the null, while the LR statistic has a null distribution affected by nuisance parameters. Following the idea of conditioning (Moreira,

2003), I provide a simple approach to finding the critical value for the LR statistic by simulating its null distribution conditioning on sufficient statistics of nuisance parameters. Consistent with previous studies (Moreira, 2003; Andrews et al., 2006; Moreira, 2009), tests based on these statistics, though all have correct size, have different power properties. In the case where the treatment effect derivative is only known to lie in a subset of $\mathbb{R}$, the projection method (Dufour, 1997; Dufour and Taamouti, 2005, 2007) can be applied and, though conservative by construction, potentially have more power than simply ignoring the information from treatment effect derivative.

The implementation of the tests is discussed. As is mentioned earlier, these tests are built by firstly assuming observable, normally distributed random discontinuities with known covariance. Tests are exactly similar under these conditions. In practice, estimators of the discontinuities and their covariance are used, resulting in tests which proved to be asymptotically similar. A key factor to guarantee this asymptotic similarity is to make sure the leading biases in estimated discontinuities shrink fast enough that they do not affect the asymptotic distribution. I make use of the recent work by Calonico et al. (2014) and show that to use bias-corrected point estimators coupled with modified variance estimators, works well for the proposed tests.

A reduced form approach in the spirit of Chernozhukov and Hansen (2008b) is proposed for even simpler implementation. It is shown that to test a fuzzy RD design estimator is equivalent to testing the smoothness of a transformed outcome under the null. As a result, hypothesis testing in fuzzy RD designs can be reduced to that done in sharp RD designs. Following this idea, I make an extension to statistic inference for quantile treatment effects in the framework of Chernozhukov and Hansen (2005). With a different set of assumptions, most importantly the rank similarity condition, I establish the smoothness of quantiles of the transformed outcome under the null. As a result, one can again test the null by simply testing the smoothness of quantiles. This approach is in line with the robust inference method for instrumental variable quantile regression proposed by Chernozhukov and Hansen (2008a).

It is well known that fuzzy RD design and IV regression share similar identification strategies and numerically yield the same results under certain conditions. As a result, even with a fuzzy RD

design one can still turn to IV regressions for estimation and, most importantly for robust inference because the literature on the latter is well developed. However, I show that there is a benefit to staying in the framework of fuzzy RD designs and using the proposed tests. Specifically, tests could be more powerful without introducing size distortions. Intuitively, this benefit comes from the fact that one can be very flexible in choosing models which best fit the data to estimate the discontinuities. While local linear estimators are advocated in the RD design literature, Card et al. (2014) argued that they are not always the best option and proposed choosing different polynomial orders depending on the data. The proposed tests perfectly accommodate the flexibility of choosing different models for the treatment/outcome variable and for the left/right side of the threshold, which is an important feature not shared by robust tests in IV regressions.

To summarize, this chapter contributes to the RD design literature in several dimensions. First, the link between fuzzy RD design and well known IV regression is explicitly examined and explored. Common test statistics such as AR, LM, and LR statistics are developed for the fuzzy RD design in both just-identified and over-identified cases. One of them is equivalent to the square of modified t-statistic proposed by Feir et al. (2016). Second, detailed estimation procedures are provided. Unlike Feir et al. (2016) who imposed the less practical assumption of under-smoothing, I allow for the use of mean squared error (MSE) optimal bandwidths that are readily available in common statistical packages. Third, despite the similarity and sometimes even equivalence between fuzzy RD design and IV regression, it is shown that adopting a fuzzy RD design with the proposed tests potentially leads to more power and is thus recommended. Lastly, a reduced form approach for hypothesis testing in fuzzy RD design is discussed. This approach is simple in computation and works equally well in both testing average treatment effects and quantile treatment effects.

The chapter is organized as follows: Section 1.2 contains the main results, including the construction of test statistics and their theoretical properties. Section 1.3 discusses the alternative implementation procedure and extension to quantile treatment effects. Section 1.4 presents simulation results to demonstrate the performance of the proposed tests and Section 1.5 examines an empirical application. Section 1.6 provides the conclusion.

## 1.2 Robust Inference in Fuzzy Regression Discontinuity Designs

### 1.2.1 The model

I consider a fuzzy RD design with the following random sample

$$\{\big(Y_i(1), Y_i(0), T_i(1), T_i(0), X_i\big)_{i=1,2,...,n}\},$$

where $X_i$ is a continuous running variable (also known as score or forcing variable), $Y_i(\cdot)$ and $T_i(\cdot)$ are the potential outcome and treatment respectively following the framework of Rubin causal model (Rubin, 1974). Given a known threshold $\bar{x}$, which is set to zero without loss of generality, the running variable $X_i$ determines whether unit $i$ is assigned treatment (when $X_i \geq 0$) or not (when $X_i < 0$). Due to incomplete compliance, the actual treatment status may be different from the assigned treatment. For subject $i$, we use $T_i(1)$ to denote the actual treatment if assigned to treatment group ($X_i \geq 0$), and $T_i(0)$ if assigned to the control group ($X_i < 0$).[1] Analogously, we use $Y_i(1)$ to denote the outcome if $i$ is actually in the treatment group (when $T_i = 1$), and $Y_i(0)$ if not (when $T_i = 0$).

In practice, the observed random sample is $\{\big(Y_i, T_i, X_i\big)_{i=1,2,...,n}\}$, where $T_i = \mathbb{1}(X_i \geq 0)T_i(1) + \mathbb{1}(X_i < 0)T_i(0)$ and $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$, with $\mathbb{1}(\cdot)$ being the indicator function. The parameter of interest is

$$\tau = \frac{\lim\limits_{x\to 0^+} \mathbb{E}(Y_i|X_i = x) - \lim\limits_{x\to 0^-} \mathbb{E}(Y_i|X_i = x)}{\lim\limits_{x\to 0^+} \mathbb{E}(T_i|X_i = x) - \lim\limits_{x\to 0^-} \mathbb{E}(T_i|X_i = x)}. \tag{1.1}$$

Under mild monotonicity and continuity conditions, Hahn et al. (2001) showed that this parameter is the average treatment effect for a subgroup of units at the threshold whose treatment decisions are affected by the running variable passing the threshold, i.e., $\mathbb{E}\left(Y_i(1) - Y_i(0)|X_i = 0, T_i(0) = 0, T_i(1) = 1\right)$.

Let $f_{\cdot}(\cdot)$ be a density function and $f_{\cdot|\cdot}(\cdot|\cdot)$ be a conditional density function. Define random vector $S_i \equiv \big(Y_i(1), Y_i(0), T_i(1), T_i(0)\big)$. I employ the continuity based framework (Sekhon and Titiunik, 2017) and adopt the following assumption:

[1] Sharp RD design is a special case of fuzzy RD design where $T_i(0) = 0$ and $T_i(1) = 1$.

**Assumption 1.1.** *For some $\epsilon > 0$, the following hold in the neighborhood $(-\epsilon, \epsilon)$ around the threshold $\bar{x} = 0$:*

*(a)* $f_X(x) > 0$.

*(b)* $T_i$ *is binary and* $T_i(1) \geq T_i(0)$.

*(c) For all* $S_i$, $f_{S|X=x}(S_i)$ *is continuous in* $x$*; its derivative* $\frac{df_{S|X=x}(S_i)}{dx}$ *exists and is continuous in* $x$.

Assumption 1.1(a) rules out discrete running variables and guarantees the existence of observations around the threshold as the sample size increases. Though minor discreteness is unavoidable in practice and not likely to affect the results, too few mass points near the threshold may cause specification error (Lee and Card, 2008) or imply measurement error (Dong, 2015; Barreca et al., 2016; Bartalotti et al., 2017a). Assumption 1.1(b) is very standard in IV models with binary instrument and binary treatment (Angrist et al., 1996). It basically rules out the possibility of defiers, who always choose the opposite of assigned treatment. Assumption 1.1(c) guarantees a certain degree of smoothness for the potential treatment and outcome at the threshold. Smoothness condition is generally required in RD designs and may take different forms depending on the specific identification strategy employed. For example, the continuity of $f_{S|X=x}(S_i)$ in $x$ is sufficient for the fuzzy RD estimator proposed by Hahn et al. (2001), but insufficient for the RPJK estimator proposed by Dong (2016) or the fuzzy quantile RD estimator proposed by Frandsen et al. (2012). In this chapter, I utilize discontinuities in both level and slope and, as a result, smoothness of $f_{S|X=x}(S_i)$ in $x$ up to its first order derivative is assumed. It is worth noting that continuity of $f_X(x)$ at the threshold is not required for the purpose of identification or inference, though its discontinuity is a signal of potential failing of Assumption 1.1(c).[2]

[2]See McCrary (2008) for a formal test of the continuity of the running variable density function.

With an intention to connect to the extensive literature on IV regression models, I assume there exists a random vector $(\Delta_{Y_n}, \Delta_{T_n})^T$ and rewrite the fuzzy RD design as two equations below:

$$\begin{aligned} \Delta_{Y_n} &= \tau \Delta_{T_n} + u_1, \\ \Delta_{T_n} &= \Pi + v_1, \end{aligned} \tag{1.2}$$

where $\Pi = \lim_{x \to 0^+} \mathbb{E}(T_i | X_i = x) - \lim_{x \to 0^-} \mathbb{E}(T_i | X_i = x)$ is the unknown discontinuity in $\mathbb{E}(T_i)$ at the threshold and $u_1, v_1$ are random errors with zero mean. The equation system (1.2) resembles a simple IV regression model, where the instrument is fixed at one and the endogenous variable is $\Delta_{T_n}$. The equation system (1.2) also differs significantly from an IV regression model because there is only one observation. This discrepancy can be well explained by the fact that fuzzy RD design shares the same identification strategies with IV regression only at the threshold, where the probability of observing any a unit is theoretically zero. As a result, $\Delta_{Y_n}$ and $\Delta_{T_n}$ can be best interpreted as unbiased estimators (though do not exist in general) of discontinuities in outcome and treatment at the threshold.

The modeling of fuzzy RD design as in (1.2) also sheds light on the strength of identification. From the weak IV literature, it is trivial to find that the concentration parameter, which measures the strength of identification or quality of instrument, is given by $\Upsilon = \Pi^2 / \mathbb{V}(v_1)$, with $\mathbb{V}(\cdot)$ denoting variance. This formula is consistent with the one derived by Feir et al. (2016). Their formula for the concentration parameter is a function of sample size, bandwidth as well as kernel choice because they replaced $\mathbb{V}(v_1)$ with its estimator from a local linear model.[3]

Recent theoretical studies on RD designs and their applications have extended to regression kink (RK) designs (Card et al., 2015a,b), where slope changes in the average treatment and the outcome are utilized to help identify treatment effects. In particular, Dong (2016) showed that the

[3]When local linear models are employed in estimating a fuzzy RD design, Feir et al. (2016) derived the concentration parameter as follows:

$$\Upsilon(h_n) = \frac{n h_n f_X(0) \Pi^2}{k(\sigma^2_{T-} + \sigma^2_{T+})} \quad \text{with} \quad k = \frac{\int_0^\infty \left( \int_0^\infty K(s) s^2 ds - u \int_0^\infty K(s) s ds \right)^2 K(u)^2 du}{\left( \int_0^\infty K(u) du \int_0^\infty K(u) u^2 du - (\int_0^\infty K(u) u du)^2 \right)^2}$$

where $h_n = O_p(n^{-r})$ is the bandwidth satisfying $\frac{1}{5} < r < \frac{1}{3}$, $K(\cdot)$ is a kernel function. Conditional variances of the treatment variable are defined by $\sigma^2_{T-} = \lim_{x \to 0^-} \mathbb{V}[T|X = x]$ and $\sigma^2_{T+} = \lim_{x \to 0^+} \mathbb{V}[T|X = x]$.

following equality holds under Assumption 1.1:

$$\lim_{x\to 0^+}\frac{\partial\,\mathbb{E}(Y_i|X_i=x)}{\partial x}-\lim_{x\to 0^-}\frac{\partial\,\mathbb{E}(Y_i|X_i=x)}{\partial x}=\tau\Big[\lim_{x\to 0^+}\frac{\partial\,\mathbb{E}(T_i|X_i=x)}{\partial x}-\lim_{x\to 0^-}\frac{\partial\,\mathbb{E}(T_i|X_i=x)}{\partial x}\Big]+\tau'\Pi. \tag{1.3}$$

The left side of equation (1.3) is the kink in the average outcome, the difference in parenthesis on the right side is the kink in the treatment, and $\tau'$ is the first order derivative of the treatment effect with respect to the running variable evaluated at the threshold.[4] It is worth noting that $\tau'$ measures the changing rate of the treatment effect at the threshold. Thus, it serves as an indicator of external validity of the locally identified treatment effect.[5] Equation (1.3) shows that fuzzy RK estimator is valid only when $\tau'\Pi=0$. Without information on $\tau'$, equation (1.3) allows us to jointly test parameters $\tau,\tau'$; with $\tau'$ being a specific known value, equation (1.3) makes $\tau$ over identified.

I propose the following model based on equations (1.2) and (1.3):

$$\begin{aligned}\Delta_{Y_n}&=\tau\Delta_{T_n}+u_1,\\ \Delta_{Y'_n}&=\tau\Delta_{T'_n}+\tau'\Delta_{T_n}+u_2,\\ \Delta_{T_n}&=\Pi+v_1,\\ \Delta_{T'_n}&=\Pi'+v_2,\end{aligned} \tag{1.4}$$

where $\Pi'=\lim_{x\to 0^+}\partial\,\mathbb{E}(T_i|X_i=x)/\partial x-\lim_{x\to 0^-}\partial\,\mathbb{E}(T_i|X_i=x)/\partial x$ is the unknown kink in $\mathbb{E}(T_i)$ at the threshold, $u_2$ and $v_2$ are random errors with zero mean. Analogous to $\Delta_{Y_n}$ and $\Delta_{T_n}$, $\Delta_{Y'_n}$ and $\Delta_{T'_n}$ are random variables with means $\tau\Pi'+\tau'\Pi$ and $\Pi'$ respectively. For example, $\Delta_{Y'_n}$ could be an unbiased estimator of the kink in $\mathbb{E}(Y_i)$ at the threshold, and $\Delta_{T'_n}$ could be an unbiased estimator of the kink in $\mathbb{E}(T_i)$ at the threshold.

[4] Though Dong (2016) provided rigorous proof of equation (1.3), one can gain some intuition by thinking of a slightly different outcome model featured by additive and linear treatment effect $Y=Y_0+T\tau$, where $Y_0$ is a smooth outcome function without treatment and $T$ is a continuous treatment. By taking derivative with respect to the running variable and then taking difference of the limits on both sides of the threshold, one can obtain $\lim_{x\to 0^+}Y'-\lim_{x\to 0^-}Y'=\tau\left(\lim_{x\to 0^+}T'-\lim_{x\to 0^-}T'\right)+\tau'\left(\lim_{x\to 0^+}T-\lim_{x\to 0^-}T\right)$. Through out this chapter, I use superscript " $'$ " as part of variable names for those defined as first order derivatives. I will later use superscript "$T$" to denote transpose.

[5] With local policy invariance assumption, this derivative is also equal to the marginal threshold treatment effect. See details from Dong and Lewbel (2015).

Suppose a random vector $(\Delta_{Y_n}, \Delta_{Y'_n}, \Delta_{T_n}, \Delta_{T'_n})^T$ is available, the objective for a researcher is to estimate the parameter of interest, $(\tau, \tau')$, and perform inferential statistic analysis based on a realization of this random vector. The unknown constants, $\Pi$ and $\Pi'$, are nuisance parameters and of no direct interest. In the case of strong identification, standard tests work well because their asymptotic distributions approximate their finite sample distributions closely. In the case of weak identification, on the contrary, the actual distributions of standard test statistics are affected by the nuisance parameters and could be significantly different from their asymptotic distributions. For example, the parameter $\Upsilon = \Pi^2 / \mathbb{V}(v_1)$ is to a fuzzy RD design as the concentration parameter is to an IV regression model. Consequently, to apply a standard t-test in the case of very small $\Upsilon$ may fail to control its size and result in invalid confidence intervals. Statistic inference for a fuzzy RK estimator is not exempted from this threat if the parameter $\Upsilon' = \Pi'^2 / \mathbb{V}(v_2)$, defined similarly to $\Upsilon$, is very small.[6] Though many tests robust to weak identification have been proposed in the weak IV literature, they are not directly applicable to fuzzy RD designs represented by equations in (1.4).

### 1.2.2 Sufficient statistics and robust tests with known variance

In matrix notation, the equation system (1.4) can be rewritten as

$$\begin{pmatrix} \Delta_{Y_n} \\ \Delta_{Y'_n} \\ \Delta_{T_n} \\ \Delta_{T'_n} \end{pmatrix} = \begin{pmatrix} \tau & 0 \\ \tau' & \tau \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \Pi \\ \Pi' \end{pmatrix} + \begin{pmatrix} v_3 \\ v_4 \\ v_1 \\ v_2 \end{pmatrix}, \tag{1.5}$$

with

$$v_3 = \tau v_1 + u_1, \quad v_4 = \tau v_2 + \tau' v_2,$$

[6] Similar to $\Upsilon$, when local linear models are employed in estimation, it can be show that

$$\Upsilon'(h_n) = \frac{n h_n^3 f_X(0) \Pi'^2}{k'(\sigma^2_{T-} + \sigma^2_{T+})} \text{ with } k' = \frac{\int_0^\infty \left( \int_0^\infty K(s) s ds - u \int_0^\infty K(s) ds \right)^2 K(u)^2 du}{\left( \int_0^\infty K(u) du \int_0^\infty K(u) u^2 du - \left( \int_0^\infty K(u) u du \right)^2 \right)^2}.$$

or more compactly, $W_n \sim N(\mu, \Omega_n)$ where

$$\begin{aligned} W_n &= (\Delta_{Y_n}, \Delta_{Y_n'}, \Delta_{T_n}, \Delta_{T_n'})^T, \\ \mu &= (\tau\Pi, \tau'\Pi + \tau\Pi', \Pi, \Pi')^T, \\ \Omega_n &= \mathbb{V}\left[(v_3, v_4, v_1, v_2)^T\right]. \end{aligned}$$

It is worth noting that $W_n$ naturally serves as a sufficient statistic for model (1.5) because it is the only sample.[7] Following the standard practice in weak IV literature, it is assumed that elements in $W_n$ are jointly normal with known covariance matrix $\Omega_n$. Moreira (2003) proposed a novel approach to partition the sufficient statistic in an IV regression model into two independent parts, which are then used to construct most of the commonly used test statistics. I show that this approach can be adapted to model (1.5) as well. To be specific, under the null hypothesis $H_0 : (\tau, \tau')^T = (\tau_0, \tau_0')^T$, two random vectors $S_n$ and $T_n$ are defined as

$$\begin{aligned} S_n^T &= W_n^T B_0 (B_0^T \Omega_n B_0)^{-\frac{1}{2}}, \\ T_n^T &= W_n^T \Omega_n^{-1} A_0 (A_0^T \Omega_n^{-1} A_0)^{-\frac{1}{2}}, \end{aligned}$$

with $B_0$ and $A_0$ defined under the null hypothesis,

$$B_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -\tau_0 & -\tau_0' \\ 0 & -\tau_0 \end{pmatrix}, \qquad A_0 = \begin{pmatrix} \tau_0 & 0 \\ \tau_0' & \tau_0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

By construction, $S_n$ and $T_n$ are normalized for later convenience as in Andrews et al. (2006). I also define matrices $B$ and $A$ in a way similar to $B_0$ and $A_0$, but with $\tau_0$ and $\tau_0'$ replaced by true parameters $\tau$ and $\tau'$. Note that the construction of $B_0$ and $A_0$ in this chapter is different from that in the weak IV literature, mainly because the parameters of interest in model (1.5) show up in both the first two equations. Each column of $B_0$ is orthogonal to each column of $A_0$, hence $[B_0(B_0^T\Omega_n B_0)^{-\frac{1}{2}} : \Omega_n^{-1}A_0(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}]$ is a nonsingular square matrix. As a result, $S_n$ and $T_n$

[7] Any other statistics calculated from it cannot provide additional information as to the value of parameters.

are also sufficient statistics equivalent to $W_n$ because there exists one-to-one mapping between them: $W_n^T = [S_n^T : T_n^T][B_0(B_0^T\Omega_n B_0)^{-\frac{1}{2}} : \Omega_n^{-1}A_0(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}]^{-1}$. Most importantly, $S_n$ and $T_n$ are jointly normally distributed with zero correlation and thus independent. These properties are summarized in Lemma 1.1.

**Lemma 1.1.** *For the model in (1.5):*

*(a) $S_n$ and $T_n$ are sufficient statistics for $\theta = (\tau, \tau', \Pi, \Pi')^T$.*

*(b) $S_n \sim N\big((B_0^T\Omega_n B_0)^{-\frac{1}{2}}(B_0 - B)^T\mu, I_2\big)$, $T_n \sim N\big((A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}A_0^T\Omega_n^{-1}\mu, I_2\big)$; $S_n$ and $T_n$ are independent.*

Though with subscript "$n$", the proof of Lemma 1.1 does not rely on asymptotics $n \to \infty$. In other words, Lemma 1.1 is valid for all $n$ regardless of the true values of parameters. Under the null hypothesis, the statistic $S_n$ follows standard multivariate normal distribution. This is because $B_0 - B = 0$ under the null hypothesis. However, the statistic $T_n$ has a distribution depending on nuisance parameters $\Pi$ and $\Pi'$ under both null and alternative hypotheses. Let $\psi(S_n, T_n, \Omega_n, \tau_0, \tau_0')$ be a continuous statistic for testing $H_0$, the most straightforward way to achieve a similar test at level $\alpha \in (0, 1)$ is to reject $H_0$ whenever $\psi$ exceeds a critical value $c_\psi$ defined by the $1-\alpha$ quantile of its null distribution. However, the null distribution of $\psi$ is generally unknown (unless $T_n$ is not involved) and the performance of asymptotic approximation crucially depends on values of $\Pi$ and $\Pi'$. Following the conditioning idea (Moreira, 2003), the exact null distribution of $\psi$ conditioning on $T_n$ is attainable because $S_n$ is standard multivariate normal and independent with $T_n$. As a result, I define the critical value

$$c_\psi(T_n, \Omega_n, \tau_0, \tau_0', \alpha) = q_\alpha\big(\psi(Q, T_n, \Omega_n, \tau_0, \tau_0')|T_n\big) \text{ with } Q \sim N(0, I_2),$$

where $q_\alpha(\cdot)$ denotes the $1-\alpha$ quantile of a random variable. Intuitively, if the critical value is fixed, it must be the case that the test statistic has a pivotal distribution. Otherwise the test statistic has a varying distribution and the critical value must be adjusted accordingly. Many widely used test statistics are based on $S_n$ and $T_n$. I focus on the Anderson-Rubin (AR) test (Anderson and Rubin,

1949), the Lagrange multiplier (LM) test (score test) (Kleibergen, 2002; Moreira, 2002) and the conditional likelihood ratio (CLR) test (Moreira, 2003) because they are widely used in empirical studies.

**The Anderson-Rubin test.** The AR statistic is the square of $S_n$,

$$AR_0 = S_n^T S_n,$$

which follows chi-squared distribution with two degrees of freedom and is consequently pivotal. With a fixed critical value $c_{AR}(T_n, \Omega_n, \tau_0, \tau_0', \alpha) = q_\alpha(\chi_2^2)$, a test which rejects $H_0$ when $AR_0 > c_{AR}(T_n, \Omega_n, \tau_0, \tau_0', \alpha)$ is similar at level $\alpha$.

Though $AR_0$ is a statistic for testing $\tau = \tau_0$ and $\tau' = \tau_0'$ jointly, with slight modification it can also be used to test $\tau = \tau_0$ only, which is of primary interest in many cases. For example, a statistic can be constructed by replacing $B_0$ with its first column only to serve this purpose, i.e.,

$$AR_0^j = \frac{(W_n^T B_0^j)^2}{(B_0^j)^T \Omega_n B_0^j} \quad \text{with} \quad B_0^j = (1, 0, -\tau_0, 0)^T,$$

resulting in a statistic equivalent to the null-restricted statistic proposed by Feir et al. (2016). Analogously, in the case of fuzzy RK design with a discontinuity in slope, one can construct a statistic by replacing $B_0$ with its second column only, with $\tau_0'$ set to zero,[8] i.e.,

$$AR_0^k = \frac{(W_n^T B_0^k)^2}{(B_0^k)^T \Omega_n B_0^k} \quad \text{with} \quad B_0^k = (0, 1, 0, -\tau_0)^T.$$

The statistic $AR_0^k$ works similarly to $AR_0^j$ but draws on information from a kink in treatment probability at the threshold. In both cases, statistics $AR_0^j$ and $AR_0^k$ are random variables following chi-squared distribution with one degree of freedom and thus the critical value is $q_\alpha(\chi_1^2)$.

**The Lagrange multiplier test**. The conventional LM statistic is a quadratic form of the score with respect to the information matrix and has a non-pivotal distribution under the null. Kleibergen (2002) proposed a new LM statistic (also known as the K-statistic) which equals a quadratic form of the score of the concentrated log-likelihood. In the context of fuzzy RD designs,

[8] The third equation in system (1.4) is dropped because it does not provide any identification power in the case $\Pi = 0$.

it can be shown that the score is

$$S_n^T(B_0^T\Omega_n B_0)^{-\frac{1}{2}}\widehat{\mathbf{\Pi}} \text{ with } \widehat{\mathbf{\Pi}} = \begin{pmatrix} \widehat{\Pi} & 0 \\ \widehat{\Pi}' & \widehat{\Pi} \end{pmatrix},$$

where $(\widehat{\Pi}, \widehat{\Pi}')^T = T_n^T(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}$ is the maximum likelihood estimator of $(\Pi, \Pi')$ under $H_0$. The LM statistic is a quadratic of the score:

$$LM_0 = S_n^T(B_0^T\Omega_n B_0)^{-\frac{1}{2}}\widehat{\mathbf{\Pi}}\big(\widehat{\mathbf{\Pi}}^T(B_0^T\Omega_n B_0)^{-1}\widehat{\mathbf{\Pi}}\big)^{-1}\widehat{\mathbf{\Pi}}^T(B_0^T\Omega_n B_0)^{-\frac{1}{2}}S_n.$$

Notice that $\widehat{\mathbf{\Pi}}$ is almost surely invertible, hence the LM statistic is reduced to

$$LM_0 = S_n^T S_n,$$

which is exactly the same as the AR statistic. As a result, the critical value of the LM test is fixed at $c_{LM}(T_n, \Omega_n, \tau_0, \tau_0', \alpha) = q_\alpha(\chi_2^2)$.

**The Likelihood ratio test.** For a given sample, a large difference in its likelihood with and without imposing the null hypothesis provides evidence against this hypothesis. For model (1.5), the likelihood ratio statistic is

$$LR_0 = S_n^T S_n - \min_{(\tau,\tau')\in\mathbb{R}^2} W_n^T B(B^T\Omega_n B)^{-1}B^T W_n,$$

where the first part corresponds to the null-restricted likelihood and the second part corresponds to the unrestricted likelihood. At first look, the calculation of $LR_0$ involves optimization over the space of $\mathbb{R}^2$ to search for $\tau$ and $\tau'$ which minimize $W_n^T B(B^T\Omega_n B)^{-1}B^T W_n$. A closer look at this optimization problem shows that a minimum of zero is always reachable because $W_n^T B = 0$ consists of two equations and two free variables and thus a solution always exists. Hence one can conclude that

$$LR_0 = S_n^T S_n, \tag{1.6}$$

and the critical value of the LR test is again $c_{LR}(T_n, \Omega_n, \tau_0, \tau_0', \alpha) = q_\alpha(\chi_2^2)$.

To summarize, when the null hypothesis is $H_0 : (\tau, \tau')^T = (\tau_0, \tau_0')^T$, the three test statistics $AR_0$, $LM_0$ and $LR_0$ are equivalent and follow chi-squared distribution with two degrees of freedom. This conclusion is consistent with previous findings on their equivalence in the just identified case (Kleibergen, 2002; Moreira, 2003).

### 1.2.3 Over identification with known treatment effect derivative

It is widely known that the AR test is inefficient in cases of over identification because the degrees of freedom of its (limiting) distribution is always equal to the number of instruments. This is a natural result from the fact that the AR statistic is obtained by projecting the disturbances of the structural equation on all instruments. However, this drawback is not shared by the LM statistic and the LR statistic.

This subsection considers a case where $\tau$ is of primary interest and additional information on $\tau'$ is available (or assumed). For example, if $\tau'$ is known, then $\tau$ can be identified from a discontinuity either in level or slope. This is empirically relevant because assumptions on $\tau'$, depending on the context, are sometimes legitimate. For example, in estimating the potential crowd out effect of the Pell Grant on the institutional grant aid, Turner (2017) assumed that one dollar of Pell Grant has constant effect on the institutional grant aid at the margin of Pell Grant eligibility, i.e., $\tau' = 0$. Besides taking specific assumed values, $\tau'$ can also be restricted in a region, which is sometimes convincing. For example, in estimating the effect of class size on test scores in Israeli public schools (Angrist and Lavy, 1999), one may be willing to assume $\tau\tau' \leq 0$ because the marginal treatment effect (the effect of one more student on average test scores) decreases in magnitude with the treatment intensity (class size).

I proceed by assuming $\tau'$ is known. Under this assumption, the parameter $\tau_0'$ in the matrices $A_0$ and $B_0$ is replaced by $\tau'$, and statistics $S_n$ and $T_n$ are updated accordingly. The null hypothesis is reduced to $H_0 : \tau = \tau_0$.

**The Anderson-Rubin test.** The AR statistic in the over identified case has exactly the same formula as in the just identified case:

$$AR_0^* = S_n^T S_n.$$

Consequently $AR_0^*$ has the same null distribution and critical value as $AR_0$.

**The Lagrange multiplier test.** The LM statistic in the over identified case is different from that in the just identified case because one only needs to take derivative of the log likelihood with

respect to $\tau$, resulting a LM statistic as follows:

$$LM_0^* = \frac{\left(S_n^T (B_0^T \Omega_n B_0)^{-\frac{1}{2}} (\widehat{\Pi}, \widehat{\Pi}')^T\right)^2}{(\widehat{\Pi}, \widehat{\Pi}')(B_0^T \Omega_n B_0)^{-1} (\widehat{\Pi}, \widehat{\Pi}')^T}.$$

Unlike the $AR_0^*$ statistic, $LM_0^*$ projects disturbances from structural equation on an IV estimate of the endogenous variable instead of all instruments (Kleibergen, 2002). Due to the one-to-one mapping between $(\widehat{\Pi}, \widehat{\Pi}')$ and $T_n$, $(\widehat{\Pi}, \widehat{\Pi}')$ is also independent with $S_n$ and $LM_0^*$ consequently has a pivotal distribution with one degree of freedom.

**The Likelihood ratio test.** The LR statistic in the over identified case is no longer equivalent to the AR statistic because its second part, $W_n^T B(B^T \Omega_n B)^{-1} B^T W_n$, can no longer always achieve a minimum of zero due to additional identifying restrictions. Specifically, the LR statistic is

$$LR_0^* = S_n^T S_n - \min_{\tau} W_n^T B(B^T \Omega_n B)^{-1} B^T W_n.$$

The distribution of $LR_0^*$ is not pivotal. As a result, the approach of conditioning can be employed to make sure a test based on $LR_0^*$ remains similar. The key to a similar test at level $\alpha$ following this approach is to obtain a critical value defined by the $1-\alpha$ quantile of the null distribution of $LR_0^*$ conditioning on the observed statistic $T_n$. I propose numerically approximating this distribution by repeatedly computing

$$LR_0^* = Q^T Q - \min_{\tau} \widetilde{W}_n^T B(B^T \Omega_n B)^{-1} B^T \widetilde{W}_n,$$

with

$$\widetilde{W}_n^T = [Q^T : T_n^T][B_0(B_0^T \Omega_n B_0)^{-\frac{1}{2}} : \Omega_n^{-1} A_0 (A_0^T \Omega_n^{-1} A_0)^{-\frac{1}{2}}]^{-1},$$

where $T_n$ is fixed at its observed value and $Q$ is drawn from the null distribution of $S_n^T$, i.e., $Q \sim N(0, I_2)$. The critical value for the CLR test $c_{CLR}(T_n, \Omega_n, \tau_0, \tau_0', \alpha)$ is then defined by the $1-\alpha$ quantile of this empirical distribution. The test that $H_0$ is rejected when $LR_0^* > c_{CLR}(T_n, \Omega_n, \tau_0, \tau_0', \alpha)$ is similar at level $\alpha$.

The discussion above shows that one can take advantage of both jump and kink to test $\tau = \tau_0$, given that $\tau'$ takes a(n) known/assumed value. This test is similar and in general more powerful than tests which make use of the jump only. In the case where it is too strong to assume $\tau'$ takes

a specific value, it might be reasonable to constrain $\tau'$ within a certain range, i.e., $\tau' \in S_{\tau'} \subset \mathbb{R}$. Then one can perform joint test for $(\tau, \tau')$ and then use projection method (Dufour, 1997) to test $\tau$ only. To be specific, a test which rejects $\tau = \tau_0$ when $(\tau, \tau') = (\tau_0, \tau_0')$ is rejected for all $\tau_0' \in S_{\tau'}$ will have correct size, though it is no longer similar. On one hand, the projection method preserves correct size at the sacrifice of power. On the other hand, the extra information that $\tau' \in S_{\tau'} \subset \mathbb{R}$ increases the the power of testing $\tau$. Combining these two facts, it is possible that using projection method, together with a reasonable constraint that $\tau' \in S_{\tau'} \subset \mathbb{R}$, will lead to a test for $\tau = \tau_0$ more powerful than $AR_0^j$.

### 1.2.4 Estimation

The sufficient statistics and robust tests introduced in the above section are based on observable $W_n$ and a known variance $\Omega_n$. In practice, however, both $W_n$ and $\Omega_n$ are not directly available in fuzzy RD designs. In this section, I show that those robust tests remain asymptotically valid when $W_n$ and $\Omega_n$ are replaced by their estimators.

Non-parametric regressions have been widely used as standard methods in RD designs since early studies by Hahn et al. (2001) and Porter (2003). One important feature of non-parametric regressions is the choice of polynomial order and bandwidth, with both having a direct effect on the quality of estimators. For example, the trade-off between bias and variance is unavoidable and, when improperly managed, may lead to invalid distributional approximations for test statistics even asymptotically. Though there has been lots of studies on choosing polynomial order and bandwidth (see a list of studies in the introduction section), I provide a brief description of the estimation procedures based on the findings from Calonico et al. (2014). In particular, for the purpose of illustration, I focus on local linear models and discuss the requirement for data generating process (DGP) around the threshold and the bandwidth selector.

Additional assumptions regarding to the DGP around the threshold and the assumption on kernel function are listed as follows:

**Assumption 1.2.** *For some $\epsilon > 0$, the following hold in the neighborhood $(-\epsilon, \epsilon)$ around the threshold $\bar{x} = 0$:*

*(a)* $\mathbb{E}(Y_i^4 | X_i = x)$ *is bounded.*

*(b)* $\mathbb{E}(Y_i | X_i = x)$ *and* $\mathbb{E}[T_i | X_i = x]$ *are three times continuously differentiable excluding* $x = 0$.

*(c) The kernel function $K(\cdot)$ is positive, bounded and continuous on the interval $(-\kappa, \kappa)$ and zero outside that interval for some $\kappa > 0$.*

It is worth noting that the smoothness condition in Assumption 1.2(b) is different from that in Assumption 1.1(b) and neither is nested in the other. While Assumption 1.1(b) is crucial for the validity of model (1.5), Assumption 1.2(b) is necessary for estimation because we are approximating Taylor expansions (up to the second order) at the threshold by local polynomials. Bounded fourth moment of the outcome and binary treatment ensure that estimands from local polynomial models are well behaved.

The estimation for each element in $W_n$ is similar: it is the difference of coefficients from local linear models on each side of the threshold. With kernel function $K(\cdot)$ and bandwidth $h$, the following shorthand notations are employed:[9]

$$
\begin{aligned}
K_{+,h}(x) &= \frac{1}{h} K\big(\frac{x}{h}\big)\, \mathbb{1}(x \geq 0), & K_{-,h}(x) &= \frac{1}{h} K\big(\frac{x}{h}\big)\, \mathbb{1}(x < 0), \\
\mu_{Z+}(x) &= \mathbb{E}(Z_i | X_i = x \geq 0), & \mu_{Z-}(x) &= \mathbb{E}(Z_i | X_i = x < 0), \\
\mu_{Z+}^{(\eta)}(x) &= \frac{d^\eta \mu_{Z+}(x)}{dx^\eta}, & \mu_{Z-}^{(\eta)}(x) &= \frac{d^\eta \mu_{Z-}(x)}{dx^\eta}, \\
\mu_{Z+}^{(\eta)} &= \lim_{x \to 0^+} \mu_{Z+}^{(\eta)}(x), & \mu_{Z-}^{(\eta)} &= \lim_{x \to 0^-} \mu_{Z-}^{(\eta)}(x),
\end{aligned}
$$

where $Z$ is a placeholder for either $Y$ or $T$. I further introduce another placeholder $\bullet$ denoting either " $+$ " or " $-$ " to simplify the notation. Let $h_{Z,0}$ and $h_{Z,1}$ be the bandwidth for estimating

[9] The DGP and estimation are independent on the left and right side of the threshold. Thus, there is no restriction of using different kernel function and bandwidth on the two sides. For expository purpose, I use the same kernel function and bandwidth in this section.

$\mu_{Z_\bullet}$ and $\mu^{(1)}_{Z_\bullet}$ respectively. Their estimators are obtained by solving the following problems:

$$\widehat{\mu}_{Z_\bullet}(h_{Z,0}) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} (Z_i - \beta_0 - X_i\beta_1)^2 K_{\bullet,h_{Z,0}}(X_i),$$
$$\widehat{\mu}^{(1)}_{Z_\bullet}(h_{Z,1}) = \arg\min_{\beta_1} \min_{\beta_0} \sum_{i=1}^{n} (Z_i - \beta_0 - X_i\beta_1)^2 K_{\bullet,h_{Z,1}}(X_i).$$

The bandwidth which minimizes the asymptotic mean squared errors (MSE) of a point estimator, such as $\widehat{\mu}_{Z_\bullet}(h_{Z,0})$ or $\widehat{\mu}^{(1)}_{Z_\bullet}(h_{Z,1})$, are widely used since they are theoretical grounded and easy to compute (Imbens and Kalyanaraman, 2012). Designed for point estimation, this MSE-optimal bandwidth may not be the best option to serve the purpose of statistic inference. Both Hahn et al. (2001) and Porter (2003) derived asymptotic distributions for RD estimators and showed that the bias is non-negligible if the MSE-optimal bandwidth is adopted. It can be expected that to target at a minimum MSE leads to variance and squared bias which are of the same order. To address this problem, one can either use a bandwidth smaller than the MSE-optimal one (under smoothing) or explicitly correct the bias. The former is straightforward in intuition because a smaller bandwidth induces less bias and more variability. However, it is also less user-friendly because there is no widely accepted theoretical guidance in choosing the bandwidth. The latter is more flexible in terms of bandwidth choices (MSE-optimal bandwidth are allowed) and is shown to have a faster shrinking speed of coverage error rate (Calonico et al., 2017) than under smoothing. The bias can be estimated from

$$\mathbb{E}[\widehat{\mu}_{Z_\bullet}(h_{Z,0})] - \mu_{Z_\bullet} = B_{\bullet,0}\mu^{(2)}_{Z_\bullet} h^2_{Z,0}\big(1 + o_p(1)\big),$$
$$\mathbb{E}[\widehat{\mu}^{(1)}_{Z_\bullet}(h_{Z,1})] - \mu^{(1)}_{Z_\bullet} = B_{\bullet,1}\mu^{(2)}_{Z_\bullet} h_{Z,0}\big(1 + o_p(1)\big),$$

where $B_{\bullet,0}$ and $B_{\bullet,1}$ are known constants depending on the running variable and kernel function. With $h_{Z,2}$ being another bandwidth, one can estimate $\mu^{(2)}_{Z_\bullet}$ through a local quadratic model in a way similar to $\mu_{Z_\bullet}$ and $\mu^{(1)}_{Z_\bullet}$,

$$\widehat{\mu}^{(2)}_{Z_\bullet}(h_{Z,2}) = \arg\min_{\beta_2} \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Z_i - \beta_0 - X_i\beta_1 - X_i^2\beta_2)^2 K_{\bullet,h_{Z,2}}(X_i),$$

and then use $\widehat{\mu}^{(2)}_{Z_\bullet}(h_{Z,2})$ to remove the biases in $\widehat{\mu}_{Z_\bullet}(h_{Z,0})$ and $\widehat{\mu}^{(1)}_{Z_\bullet}(h_{Z,1})$. I use the differences of bias-corrected estimates to construct $\widehat{W}_n = (\widehat{\Delta}_{Y_n}, \widehat{\Delta}_{Y'_n}, \widehat{\Delta}_{T_n}, \widehat{\Delta}_{T'_n})^T$, where

$$\widehat{\Delta}_Z = \widehat{\mu}_{Z_+}(h_{Z,0}) - \widehat{\mu}_{Z_-}(h_{Z,0}) - \big(B_{+,0}\widehat{\mu}^{(2)}_{Z_+}(h_{Z,2}) - B_{-,0}\widehat{\mu}^{(2)}_{Z_-}(h_{Z,2})\big)h^2_{Z,0},$$
$$\widehat{\Delta}_{Z'} = \widehat{\mu}^{(1)}_{Z_+}(h_{Z,1}) - \widehat{\mu}^{(1)}_{Z_-}(h_{Z,1}) - \big(B_{+,1}\widehat{\mu}^{(2)}_{Z_+}(h_{Z,2}) - B_{-,1}\widehat{\mu}^{(2)}_{Z_-}(h_{Z,2})\big)h_{Z,1}.$$

Due to bias correction, the variance of $\widehat{W}_n$ differs from that of the original biased estimator. Let $\widehat{\Omega}_n$ be the estimator for $\mathbb{V}(\widehat{W}_n)$.[10] The properties of these estimators are summarized below:

**Lemma 1.2.** *Let Assumption 1.2 hold. If* $n\min\{h^5_{Z,0}, h^5_{Z,2}\}\max\{h^2_{Z,0}, h^2_{Z,2}\} \to 0$, $n\min\{h^5_{Z,1}, h^5_{Z,2}\}\max\{h^2_{Z,1}, h^2_{Z,2}\} \to 0$ *and* $n\min\{h_{Z,0}, h_{Z,1}, h_{Z,2}\} \to \infty$, *then*

$$(\widehat{W}_n - \mu)\widehat{\Omega}_n^{-\frac{1}{2}}(\widehat{W}_n - \mu)^T \to^d N(0, I_4),$$

*provided that* $\kappa\max\{h_{Z,0}, h_{Z,1}, h_{Z,2}\} < \epsilon$.

Lemma 1.2 is a natural extension of Theorem 1 in Calonico et al. (2014), who proved the asymptotic normality of bias-corrected sharp RD estimators. Since $\widehat{W}_n$ is a vector of four bias-corrected sharp RD estimators, its joint normality can be established through Cramér-Wold theorem. As is emphasized by Calonico et al. (2014), Lemma 1.2 accommodates a wide range of bandwidths, including the MSE-optimal bandwidths. With estimators $\widehat{W}_n$ and $\widehat{\Omega}_n$, the feasible sufficient statistics are defined as $\widehat{S}^T_n = \widehat{W}^T_n B_0(B_0^T\widehat{\Omega}_n B_0)^{-\frac{1}{2}}$ and $\widehat{T}^T_n = \widehat{W}^T_n\widehat{\Omega}^{-1}_n A_0(A_0^T\widehat{\Omega}^{-1}_n A_0)^{-\frac{1}{2}}$, which are then used to construct test statistics as well as critical values.

**Theorem 1.1.** *Let Assumptions 1.1 and 1.2 hold. Choose a sequence* $\{\Omega_n\}$ *such that* $\mathbb{V}(\widehat{W}_n) - \Omega_n \to^p 0$, *then*

*(a)* $(\widehat{S}_n, \widehat{T}_n) \to^d (S_n, T_n)$,

*(b)* $\big(\psi(\widehat{S}_n, \widehat{T}_n, \widehat{\Omega}_n, \tau_0, \tau'_0), c_\psi(\widehat{T}_n, \widehat{\Omega}_n, \tau_0, \tau'_0, \alpha)\big) \to^d \big(\psi(S_n, T_n, \Omega_n, \tau_0, \tau'_0), c_\psi(T_n, \Omega_n, \tau_0, \tau'_0, \alpha)\big)$,

*(c) Under the null hypothesis,* $Pr\big(\psi(\widehat{S}_n, \widehat{T}_n, \widehat{\Omega}_n, \tau_0, \tau'_0) > c_\psi(\widehat{T}_n, \widehat{\Omega}_n, \tau_0, \tau'_0, \alpha)\big) \to^p \alpha$.

[10] Its formula is straightforward but lengthy and thus left to the appendix.

Part (a) of Theorem 1.1 states that the joint distribution of feasible sufficient statistics converges to that of infeasible sufficient statistics. It is the key conclusion because, together with the continuous mapping theorem, it is sufficient for part (b) and part (c). This establishes that the AR test, the LM test and the CLR test are exactly similar for model (1.5) with infeasible $S_n$ and $T_n$. Theorem 1.1 guarantees that tests based on proper estimators of $S_n$ and $T_n$ are still asymptotically similar.

## 1.3 Discussion and Extension

### 1.3.1 Alternative implementation of AR test

The idea behind the AR test in a conventional setting is to check whether the residuals from structural equations under the null hypothesis are orthogonal to instrumental variables. In the context of fuzzy RD designs, the instrument is valid only at the threshold and the orthogonality condition reduces to continuity condition. As a result, the following lemma holds:

**Lemma 1.3.** *Let Assumption 1.1 hold. Define* $Y_i^* = Y_i - (\tau_0 + \tau_0' X_i)T_i$. *Then* $\mathbb{E}(Y_i^*)$ *and* $\frac{\partial \mathbb{E}(Y_i^*|X_i=x)}{\partial x}$ *are continuous at* $x = 0$ *under the null hypothesis.*

It is straightforward to see that $AR_0^j$ is designed to test the continuity of $\mathbb{E}(Y_i^*)$, $AR_0^k$ is designed to test the continuity of $\frac{\partial \mathbb{E}(Y_i^*|X_i=x)}{\partial x}$ and $AR_0$ is designed for a joint test. In other words, Lemma 1.3 is an alternative presentation of the model (1.5). However, Lemma 1.3 implies a much simpler approach to perform the AR test: one just need to calculate $Y_i^*$ first and then test its smoothness. Any evidence for the existence of a jump or kink in $Y^*$ signals the violation of the null hypothesis. In other words, for inferential purpose, a fuzzy RD design is transformed into a sharp RD design once $Y_i^*$, rather than the original $Y_i$, is used as the outcome.

### 1.3.2 Test quantile treatment effect

I next show that the test discussed above for average treatment effects can be adapted to one type of quantile treatment effects with slightly different assumptions.[11] These two cases share

[11] See Frandsen et al. (2012); Chiang and Sasaki (2016) for previous studies.

similar ideas, which is to firstly remove the treatment effects and then check smoothness of the outcome at the threshold. To start with, I define the quantile treatment effect as

$$\tau(p) = y(1, x, p) - y(0, x, p)|_{x=0}, \quad p \in (0, 1), \tag{1.7}$$

where $y(T_i, X_i, p) = q_p\big(Y_i(T_i)|X_i\big)$ is the conditional $p$th quantile of potential outcome and $\tau(p)$ is the difference between two $p$th quantiles at the threshold with and without treatment. Specifically, $\tau(p)$ is the parameter of interest under the context of regression discontinuity. It is worth noting that $\tau(p)$ is in general not informative about the distribution of heterogeneous treatment effects $Y_i(1) - Y_i(0)$. However, with certain assumptions such as rank similarity, it measures the treatment effect for subjects at the $p$th quantile of the outcome. Formally, I adopt assumptions similar to Chernozhukov and Hansen (2005) and study their implications on identifying and testing quantile treatment effects in fuzzy RD designs.

**Assumption 1.3.** *Let $U_i(\cdot)$ be the percentile of subject $i$ in the distribution of the outcome if every unit has treatment status indicated by $\cdot$. Let $U_i = T_i U_i(1) + (1 - T_i)U_i(0)$. For some $\epsilon > 0$, the following hold in the neighborhood $(-\epsilon, \epsilon)$ around the threshold $\bar{x} = 0$:*

*(a) No discrete response. Given $(T_i, X_i)$, the outcome $Y_i \equiv y(T_i, X_i, U_i)$ is strictly increasing in $U_i$ and $U_i \sim U(0, 1)$.*

*(b) Rank similarity. Given $X_i$ and unobservable $W_i$, $T_i \equiv t(X_i, W_i)$ and $U_i(1) \sim U_i(0)$.*

Assumption 1.3(a) requires a one-to-one mapping from the quantile $U_i$ to the outcome $Y_i$ given any $(T_i, X_i)$. This condition does not necessarily imply a continuous outcome variable but there should be no non-zero probability mass on the support of $Y_i$. The fact that $U_i \sim U(0, 1)$ is not restrictive due to the Skorohod representation of random variables. Assumption 1.3(b) imposes rank similarity conditional on factors determining treatment status. This condition is somewhat weaker than the rank invariance condition, which states that ranks do not change under different treatments. Assumption 1.3(b) allows unsystematic variation in ranks under different treatments conditional on $(X_i, W_i)$.

Similar to $\tau'$, I define $\tau'(p)$ as the first order derivative of the quantile treatment effect with respect to the running variable.[12] With null hypothesis $H_0 : \big(\tau(p), \tau'(p)\big) = \big(\tau_0(p), \tau_0'(p)\big)$, Lemma 1.4 summarizes findings on smoothness at the threshold, which can be used to construct similar tests.

**Lemma 1.4.** *Let Assumptions 1.1 and 1.3 hold.*

*(a) The quantile function $y(\cdot, x, p)$ and its derivative $\frac{\partial y(\cdot,x,p)}{\partial x}$ are continuous at $x = 0$.*

*(b) Define $Y_i^* = Y_i - \big(\tau_0(p) + \tau_0'(p)X_i\big)T_i$, then $q_p(Y_i^*|X_i = x)$ and its derivative $\frac{dq_p(Y_i^*|X_i=x)}{dx}$ are continuous at $x = 0$ under the null hypothesis.*

Lemma 1.4(a) establishes the smoothness of the quantile function at the threshold, which is stronger than the smoothness of expectation used in estimating mean treatment effects. Lemma 1.4(b) is analogous to Lemma 1.3 and can be used to construct tests robust to weak identification. For example, one can employ local quantile regression to obtain the quantiles and their derivatives for $Y_i^*$ at the threshold. To test the null is equivalent to test whether the differences in quantiles or derivatives on two sides of the threshold are significantly different zero.

## 1.4 Monte Carlo Simulations

### 1.4.1 Comparison of size and power

I compare the size and power of a series of tests, including the standard t test and the newly proposed robust tests, through simulations from the following DGP:

$$\begin{aligned} X_i &\sim U(-1, 1), \\ T_i &= \mathbb{1}[X_i \geq 0](d_0 + d_1 X_i) + v, \\ Y_i &= \tau T_i + u, \end{aligned} \tag{1.8}$$

[12] Formally, this derivative is defined as

$$\tau'(p) = \left.\frac{\partial\big(y(1, x, p) - y(0, x, p)\big)}{\partial x}\right|_{x=0}.$$

where $v$ and $u$ are jointly standard normal with correlation $\rho$, $d_0$ and $d_i$ are the jump and kink at the threshold, and $\tau$ is a constant treatment effect.[13] The primary reason to adopt a very simple DGP like (1.8) is to isolate the confounding effects from choices of local polynomial orders and bandwidths.[14] In particular, I adopt local linear regressions with a fixed bandwidth of one to estimate $W_n$ and $\Omega_n$, i.e., the whole sample will be used and the estimators are unbiased. With this setup, it is straightforward to theoretically derive the distribution of $W_n = (\Delta_{Y_n}, \Delta_{Y'_n}, \Delta_{T_n}, \Delta_{T'_n})$ when the sample size is $n$:

$$W_n \sim N\left((d_0\tau, d_1\tau, d_0, d_1), \frac{8}{n}\begin{pmatrix} \tau^2+2\tau\rho+1 & \tau+\rho \\ \tau+\rho & 1 \end{pmatrix} \otimes \begin{pmatrix} 2 & -3 \\ -3 & 6 \end{pmatrix}\right). \tag{1.9}$$

I pick $n = 100$, $\tau = 1$ and choose different sets of $(d_0, d_1)$ to control for concentration parameters $\Upsilon$ and $\Upsilon'$. Since the distribution of $W_n$ given by (1.9) is always exact and does not rely on large "$n$", it is expected that robust tests based on (1.9) are also exactly similar.

Table 1.1 reports the probabilities of rejecting the null hypothesis $H_0 : \tau = 1$ for various tests at nominal level of 5% based on 2000 replications. The results are divided into three panels. From the top to the bottom, the identification strength ranges from very weak identification ($\Upsilon_j = \Upsilon_k = 1$) to very strong identification ($\Upsilon_j = \Upsilon_k = 100$). In each panel, results for cases of zero correlation ($\rho = 0$), negative correlation ($\rho = -0.9$) and positive correlation ($\rho = 0.9$) are reported. A total of seven tests are considered. The tests $t_j$ and $t_k$ are standard t tests for fuzzy RD design and fuzzy RK design respectively. The test $AR_j$ is the robust version of $t_j$ (also used in Feir et al. (2016)) and the test $AR_k$ is the robust version of $t_k$. The remaining three tests, AR, LM and CLR, are conducted by assuming a known $\tau'$, which is zero in the DGP of (1.8). Except for the standard t tests, the other five tests have been theoretically shown to be robust to weak identification. Results from numerical simulations confirm this conclusion. In Table 1.1, a valid test should reject the

[13] Assumption 1.1 requires a binary treatment variable. Besides the fact that binary treatments are popular, the main purpose of this condition is to make sure the outcome is additive and linear in the treatment without imposing further restrictions on functional forms. Since the outcome is already assumed to be additive and linear in the treatment in DGP (1.8), it is not necessary for the treatment to be binary.

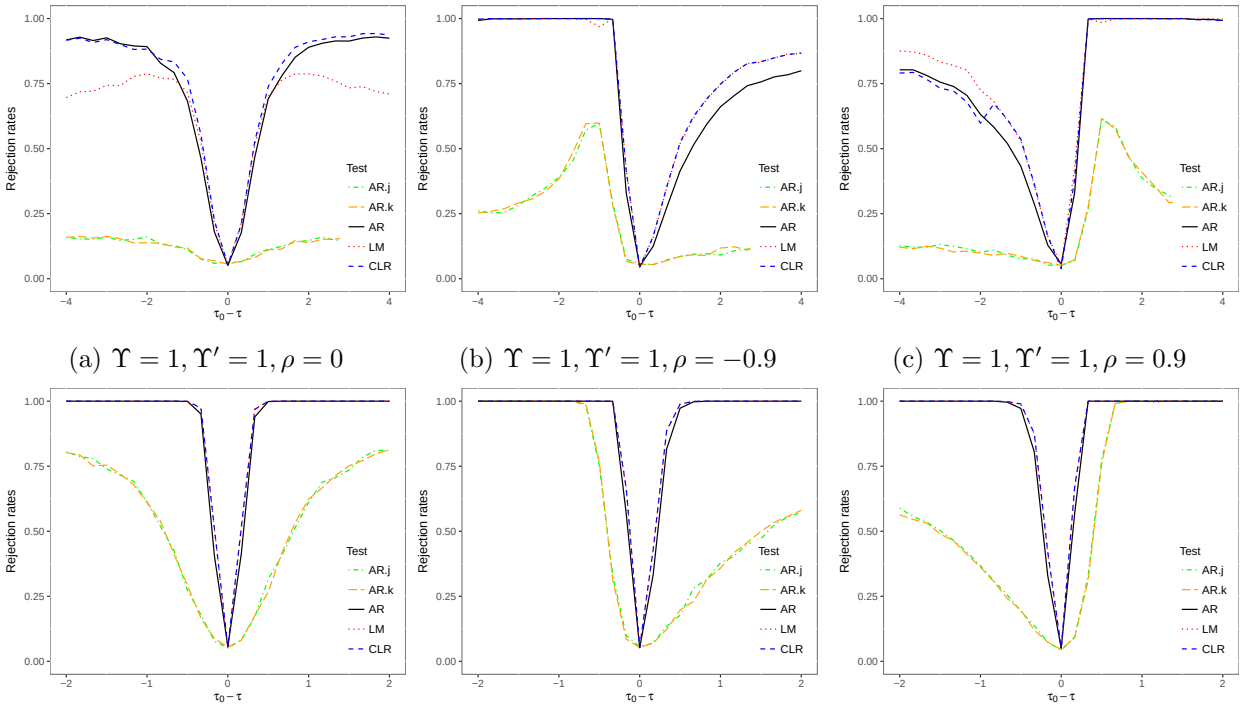
[14] The bandwidth choice partially determines the concentration parameter and, together with the curvature of $\mathbb{E}(T_i)$ and $\mathbb{E}(Y_i)$, has an effect on the magnitude of bias. Differences in bandwidth and bias are confounding factors in evaluating the performance of different tests.

null with a probability of 5% regardless of the identification strength, which is exactly the case for robust tests such as $AR_j$, $AR_k$, $AR$, $LM$ and $CLR$. On the contrary, the rejection probabilities for $t_j$ and $t_k$ are very different from 5% unless the identification is very strong (the bottom panel). In the case of weak identification (the top and middle panels), tests $t_j$ and $t_k$ underreject the null when $\rho = 0$ and overreject the null when $\rho = \pm 0.9$. Additional simulation results are available in the appendix where the nominal level is set to 5% (See Table A.1). Results from these two tables also show that the size distortion of standard t tests is more obvious when the identification is weaker. Take $t_j$ as an example, its actual size can be as large as 16.8% when the nominal size is 5%, and can be as large as 21.4% when the nominal size is 10%.

Table 1.1: Percent Rejected under $H_0 : \tau = 1$ at Nominal Level of 5%

| $\rho$ | $t_j$ | $t_k$ | $AR_j$ | $AR_k$ | $AR$ | $LM$ | $CLR$ |
|---|---|---|---|---|---|---|---|
| Panel A: $\Upsilon_j = \Upsilon_k = 1$ | | | | | | | |
| 0.0 | 0.0 | 0.0 | 5.2 | 5.0 | 5.3 | 5.4 | 5.0 |
| -0.9 | 16.8 | 15.6 | 5.1 | 5.2 | 4.8 | 5.0 | 5.1 |
| 0.9 | 15.3 | 15.4 | 4.8 | 5.4 | 4.5 | 4.7 | 3.4 |
| Panel B: $\Upsilon_j = \Upsilon_k = 10$ | | | | | | | |
| 0.0 | 1.3 | 1.6 | 4.9 | 5.1 | 5.2 | 5.4 | 5.3 |
| -0.9 | 8.2 | 8.3 | 5.2 | 4.8 | 4.8 | 4.8 | 4.9 |
| 0.9 | 8.2 | 7.5 | 4.8 | 5.1 | 4.8 | 5.0 | 4.9 |
| Panel C: $\Upsilon_j = \Upsilon_k = 100$ | | | | | | | |
| 0.0 | 3.5 | 4.4 | 3.7 | 4.8 | 4.8 | 5.1 | 5.2 |
| -0.9 | 4.5 | 5.9 | 4.8 | 5.1 | 5.0 | 5.0 | 5.0 |
| 0.9 | 5.2 | 5.3 | 5.1 | 5.1 | 6.2 | 5.4 | 5.5 |

Though all the robust tests demonstrate correct size in Table 1.1 and A.1. They differ in efficiency. In the weak IV literature, it is known that the AR test usually lose some power in the case of over identification because its degree of freedom is larger than the number of endogenous variables. On the other hand, CLR is shown to be the most powerful one among a class of invariant similar test (Andrews et al., 2006). These findings are expected to continue to hold in the context of fuzzy RD designs. Figure 1.2 plots the rejection probabilities from testing a sequence of values




(a) $\Upsilon = 1, \Upsilon' = 1, \rho = 0$ (b) $\Upsilon = 1, \Upsilon' = 1, \rho = -0.9$ (c) $\Upsilon = 1, \Upsilon' = 1, \rho = 0.9$

(d) $\Upsilon = 10, \Upsilon' = 10, \rho = 0$ (e) $\Upsilon = 10, \Upsilon' = 10, \rho = -0.9$ (f) $\Upsilon = 10, \Upsilon' = 10, \rho = 0.9$

Figure 1.2: Power of Tests at Nominal Level of 5%

for $\tau$. Not surprisingly, $AR_j$ and $AR_k$ have low power compared with other three tests because they use information from jump or kink only. Among the other three test, the $CLR$ is shown to have more power than $AR$ and $LM$.

### 1.4.2 Polynomial order, bandwidth and bias correction

The simulations considered in section 1.4.1 are based on known polynomial orders so that I have the correct specification. The main motivation is to provide a "clean" assessment on tests' performance, otherwise the comparison would become less obvious and convincing because the choices of polynomial order and bandwidth may have different effects on different tests. Since section 1.4.1 provides strong evidence for the validity of robust tests proposed in this chapter, I move to more practical cases where two widely used DGPs from Lee and Card (2008) and Ludwig and Miller (2007) are adopted. I want to show that (i) similar to standard tests, the robust

tests rely on a match between the point estimator and its variance as well, and (ii) the choices of polynomial order and bandwidth can be made flexible for $Y_i$ and $T_i$/left and right side to improve the power of tests. While the former is well recognized by many researchers, it is still new to robust tests (Feir et al. (2016) adopted the undersmoothing approach without proposing a specific bandwidth selector). The latter is related to studies by Card et al. (2014) and Gelman and Imbens (2017). Unlike them, the main purpose is to show that a proper choice of combination of different polynomial orders and bandwidths has a potential to improve the power of robust tests, which is an advantage not shared by the practice of estimating a fuzzy RD design through an IV regression model.

Two DGPs are chosen for the outcomes. The first one comes from Lee (2008) (hereafter Lee2008) and the second one comes from Ludwig and Miller (2007) (hereafter LM2007). Both these two DGPs are intensively adopted in RD literature. However, they are for sharp RD designs. Hence, I couple them with two additional DGPs for the treatment variables. In summary, I have a reduced form DGP for $\big(Y_i, T_i, X_i\big)$:

$$
\begin{aligned}
Y_i &= \mu_j^Y(X_i) + \varepsilon_i, \\
T_i &\sim B\big(1, \mu_l^T(X_i)\big), \\
X_i &\sim 2 \times Beta(2,4) - 1, \\
\varepsilon_i &\sim N(0, 0.1295^2),
\end{aligned}
$$

where the running variable $X_i$ follows Beta distribution, the treatment variable $T_i$ follows Bernoulli distribution with mean $\mu_l^T(X_i)$, and the outcome variable $Y_i$ follows normal distribution with mean $\mu_j^Y(X_i)$. The subscripts $j = 1, 2$ and $l = 1, 2$ represent two different functions for the mean outcome

and the mean treatment. The mean functions for the outcome are

$$\text{Lee2008: } \mu_1^Y(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0, \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise.} \end{cases}$$

$$\text{LM2007: } \mu_2^Y(x) = \begin{cases} 3.71 + 2.30x + 3.28x^2 + 1.45x^3 + 0.23x^4 + 0.03x^5 & \text{if } x < 0, \\ 0.26 + 18.49x - 54.81x^2 + 74.30x^3 - 45.02x^4 + 9.83x^5 & \text{otherwise.} \end{cases}$$

The mean functions for the treatment are

$$\text{Quintic: } \mu_1^T(x) = \begin{cases} (x - x^3 + x^5)/4 + 0.3 & \text{if } x < 0, \\ (x - x^3 + x^5)/4 + 0.7 & \text{otherwise.} \end{cases}$$

$$\text{Linear: } \mu_2^T(x) = \begin{cases} 0.3x + 0.3 & \text{if } x < 0, \\ 0.3x + 0.7 & \text{otherwise.} \end{cases}$$

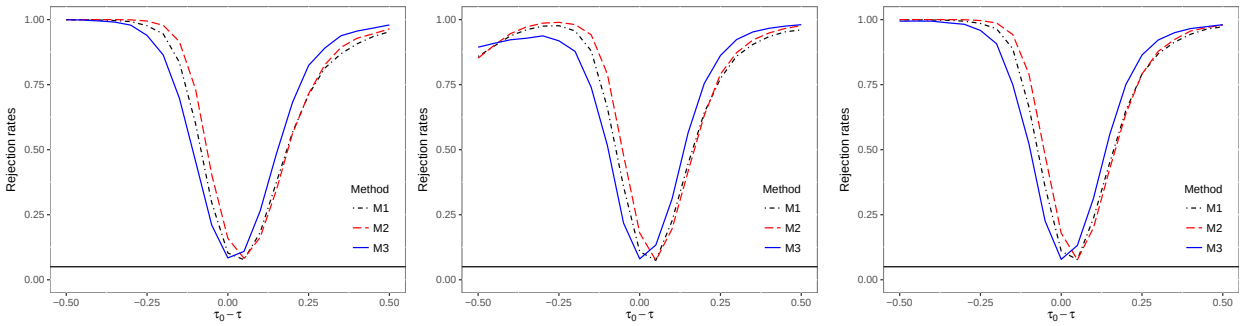
For the choices of polynomial order and bandwidth in estimation, I adopt three methods (M1, M2, and M3 for short). Among them, M1 and M2 were proposed by Calonico et al. (2014). To be specific, with M1, I choose local linear regression to estimate intercepts and local quadratic regression to estimate slopes, with a single bandwidth which minimizes the asymptotic MSE of sharp RD estimator for the outcome variable.[15] The second method, M2, is similar to M1 except that a single bandwidth is chosen to minimized the MSE of the fuzzy RD estimator. The third method, M3, is inspired by findings from Card et al. (2014), which suggest that the order of polynomial should depend on the data rather than being fixed. Following their practice, I choose polynomial order (from 1, 2 and 3) and bandwidth jointly to minimize the asymptotic MSE of the intercept/slope estimator, and this selection is done separately for the outcome/treatment variable and left/right side. Since M3 is more flexible than M1 and M2, it is expected that tests based on M3 would have more power.

[15] Imbens and Kalyanaraman (2012) proposed bandwidth selectors for sharp RD design and fuzzy RD design. They argued that these two are usually similar in practice and suggested using the one for sharp RD design for simplicity. Lee and Lemieux (2010) provided additional argument that the treatment function is usually expected to be flatter than the outcome function around the threshold, thus the MSE optimal bandwidth for estimating the treatment function is in general wider than the one for estimating the outcome function. Alternatively, one may want to choose the smaller one among these two, as is suggested by Imbens and Lemieux (2008).

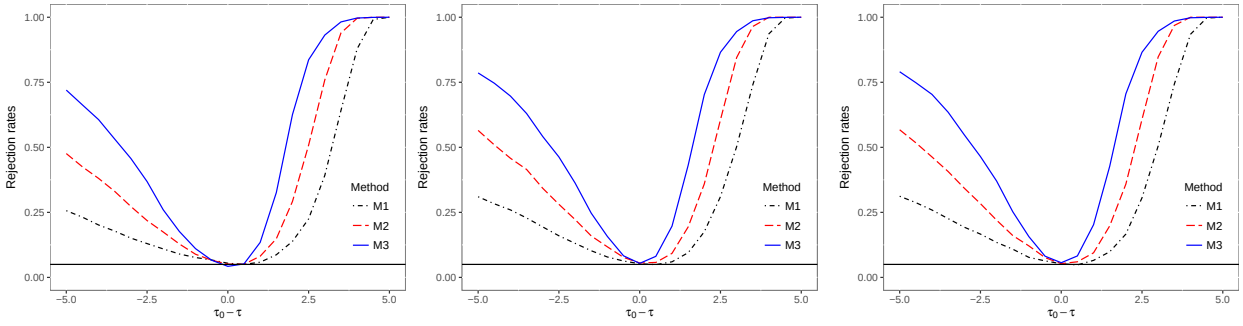
Table 1.2 reports the probabilities of rejecting the null by $AR$, $LM$ and $CLR$ tests in the over identified case.[16] Since the bandwidths are chosen to minimize the asymptotic MSE of point estimators, they may be not the best choices for statistic inference. In the first three columns of Table 1.2, where the leading bias is not corrected, the rejection probabilities are very high for Panel A and C (as high as 37% for a test at nominal level of 5%). After bias correction, the rejection probabilities decrease substantially but are still well above the nominal level. This over rejection in Panel A and C, even after bias correction, is a result of choosing large bandwidths. A closer inspection reveals that bandwidths used in Panel A and C are much wider than those used in Panel B and D. Since the outcome function in Panel B and D has more curvature, the bandwidth selector responds by selecting smaller bandwidths. In addition, the bandwidths used in M2 are larger than those in M1, which explains the higher distortion in M2. This is because the fuzzy RD estimator has more variability than the sharp RD estimator. Thus, larger biases and consequently larger bandwidths are allowed. Overall, both M1 and M2 do not perform well in controlling the size of tests in Panel A and C, while M1 does a much better job and the rejection probabilities are substantially closer to the nominal level.

Besides the actual size of tests, their power is also of interest. Figure 1.4 shows the power curves of $AR$, $LM$ and $CLR$ tests with different methods and different DGPs.[17] From the top to the bottom, the four rows of plots in Figure 1.4 correspond to the four panels A, B, C and D in Table 1.2. For the DGPs in Panel A and C, the power of all tests under M1, M2 and M3 are similar. In particular, with M3, all tests have less power left to the true parameter but more power right to the true parameter. For the DGPs in Panel B and D, the power differs substantially under M1, M2 and M3. All the three tests under consideration have the most power under M3 and the least power under M1. The better power under M2 over that of M1 is due to larger optimal bandwidths generated by M2. The drawback comes along with larger bandwidths is that M2 does not consistently perform well in controlling size of tests, as is shown in Table 1.2.

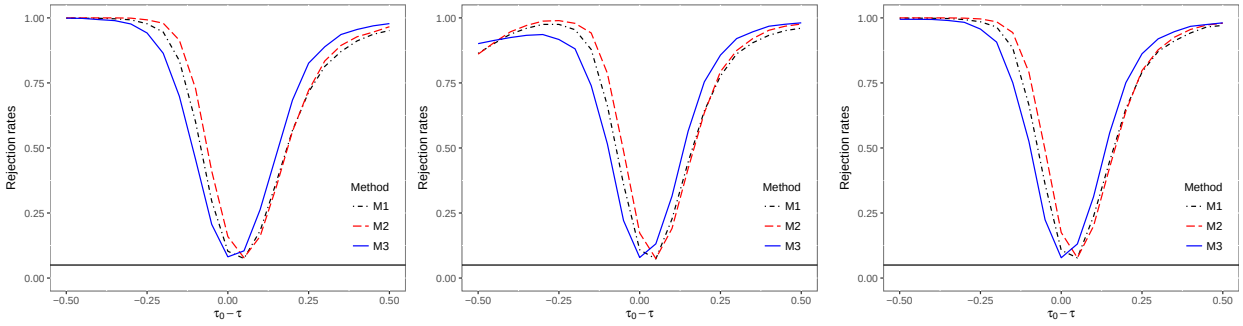
[16] Simulation results for sample sizes of 500 and 10000 are in the Appendix.
[17] Similar plots for sample sizes of 500 and 10000 are in the Appendix.



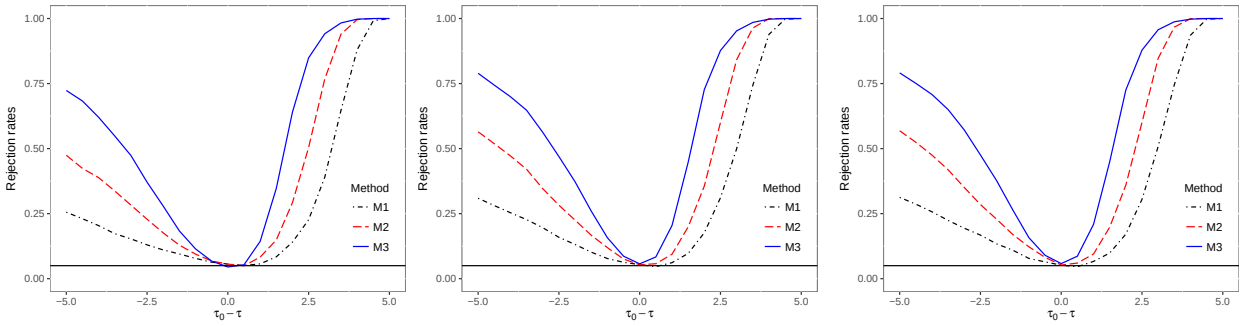

(a) AR test with $\mu_1^Y$ and $\mu_1^T$ (b) LM test with $\mu_1^Y$ and $\mu_1^T$ (c) CLR test with $\mu_1^Y$ and $\mu_1^T$




(d) AR test with $\mu_2^Y$ and $\mu_1^T$ (e) LM test with $\mu_2^Y$ and $\mu_1^T$ (f) CLR test with $\mu_2^Y$ and $\mu_1^T$




(g) AR test with $\mu_1^Y$ and $\mu_2^T$ (h) LM test with $\mu_1^Y$ and $\mu_2^T$ (i) CLR test with $\mu_1^Y$ and $\mu_2^T$




(j) AR test with $\mu_2^Y$ and $\mu_2^T$ (k) LM test with $\mu_2^Y$ and $\mu_2^T$ (l) CLR test with $\mu_2^Y$ and $\mu_2^T$

Figure 1.4: Power of Bias-corrected Tests at Nominal Level of 5% with $N = 5000$

Table 1.2: Percent Rejected at Nominal Level of 5% with $N = 5000$

| Method | $M1$ | $M2$ | $M3$ | $M1$ | $M2$ | $M3$ |
|---|---|---|---|---|---|---|
| Bias correction | No | No | No | Yes | Yes | Yes |
| Test | Panel A: $\mu_1^Y$ and $\mu_1^T$ | | | | | |
| $AR$ | 21.4 | 33.8 | 15.3 | 10.3 | 16.0 | 8.4 |
| $LM$ | 21.3 | 36.5 | 13.2 | 10.8 | 18.0 | 8.1 |
| $CLR$ | 21.4 | 36.6 | 12.4 | 11.2 | 18.1 | 7.8 |
| | Panel B: $\mu_2^Y$ and $\mu_1^T$ | | | | | |
| $AR$ | 5.8 | 6.6 | 5.9 | 5.4 | 5.1 | 4.2 |
| $LM$ | 5.4 | 7.4 | 5.9 | 5.2 | 5.5 | 5.4 |
| $CLR$ | 5.3 | 7.2 | 6.2 | 5.2 | 5.5 | 5.8 |
| | Panel C: $\mu_1^Y$ and $\mu_2^T$ | | | | | |
| $AR$ | 21.1 | 34.4 | 15.8 | 10.3 | 16.0 | 8.2 |
| $LM$ | 21.5 | 37.0 | 13.4 | 10.8 | 17.3 | 7.9 |
| $CLR$ | 21.5 | 37.4 | 12.9 | 11.1 | 17.8 | 7.4 |
| | Panel D: $\mu_2^Y$ and $\mu_2^T$ | | | | | |
| $AR$ | 5.5 | 6.3 | 6.3 | 5.6 | 5.6 | 4.5 |
| $LM$ | 5.3 | 6.8 | 5.8 | 5.4 | 5.3 | 5.7 |
| $CLR$ | 5.3 | 6.6 | 6.3 | 5.2 | 5.4 | 5.8 |

In summary, I have considered and compared three methods for estimation and statistic inference in fuzzy RD designs. Among them, M1 and M2 use local liner regressions with a single bandwidth. In this case, the estimation and inference can also be done by conventional IV regression models. The third method, M3, which has more flexibility in choosing the order of polynomial and bandwidths, are shown to leads to tests with some desirable properties when compared with M1 and M2. In particular, tests under M3 consistently perform well in controlling size and have power at least on par with, if not better than, those under M1 and M2. It is worth noting that M3 is not compatible with IV regression models. The proposed tests perfectly accommodate a flexible choice of polynomial orders and bandwidths, thus, they have advantages over the robust tests developed in the framework of IV regression.

## 1.5 Empirical Application

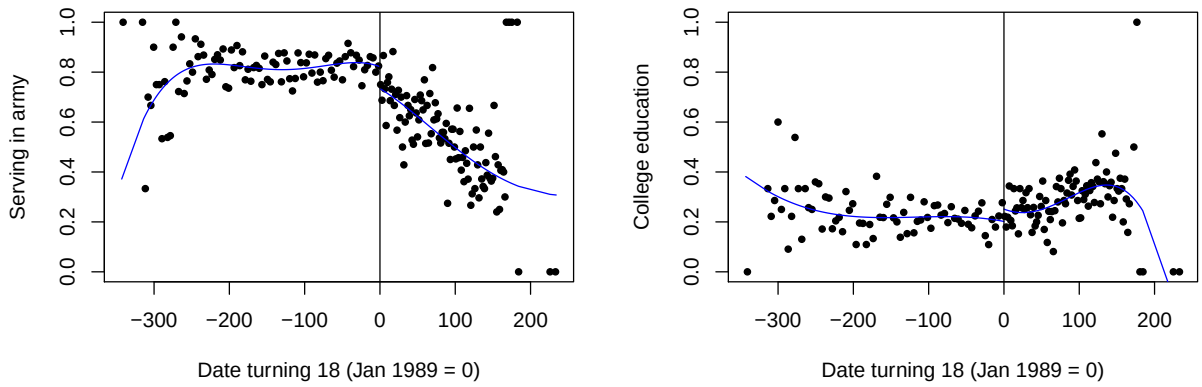
In this section, I reexamine the effect of military service on education using data from Russia Longitudinal Monitoring Survey.[18] Results of both standard and weak identification robust inference are reported and compared. I show that the standard method and the proposed method yield different results when the identification is weak (small bandwidth and few observations), but similar results when the identification is strong (large bandwidth and many observations). In particular, the confidence set derived from the standard method is too small in the case of weak identification, and thus very likely to undercover the true parameter.

In the late 1980s, the end of Cold War was followed by a significant demilitarization process in Russia. Card and Yakovlev (2014) showed that the share of Russian males who served in the army decreased sharply after 1989 and modeled this change as a RK design. Their findings suggest that military services increase risky behaviors such as alcohol consumption and smoking, resulting associated chronic illness. Dong (2016) used the same data and applied the RPJK design to estimate the effect of military services on education and earnings. Contrary to the prevailing evidence from the US and other OECD countries, Dong (2016) found that the conscription in Russia has a negative effect on college education.

Following Card and Yakovlev (2014) and Dong (2016), the running variable is the date when a male turned 18, which is the official conscription age in Russia, and the threshold is January 1989. A male who turned 18 after this threshold would have smaller probability of being drawn to the army than a male who turned 18 before this threshold. I focus the attention to males aged 30-60 in the data and their probabilities of serving the army are shown in Figure 1.6 (a).[19] Figure 1.6 (a) suggests both a jump and a kink in the probability of serving in army at the threshold. These discontinuities seem to be mirrored at the same threshold in college education, as is plotted in Figure 1.6 (b).

[18] The survey data is available at http://www.cpc.unc.edu/projects/rlms-hse.

[19] The probability of serving in army is fitted with fourth order polynomials separately on each side of the threshold.




(a) Discontinuity in military service

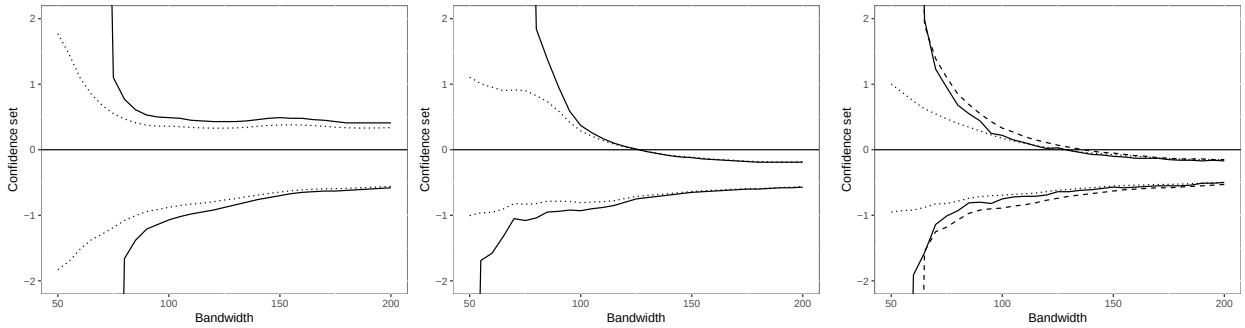
(b) Discontinuity in college education

Figure 1.6: Discontinuities in military service and college education

I estimate the effect of military services on education with three different designs. One is the standard fuzzy RD design, the second is the fuzzy RK design following Card and Yakovlev (2014), and the third is the RPJK design following Dong (2016).[20] Confidence sets derived from standard and robust tests are shown in Figure 1.8 for a wide range of bandwidths. In all the three plots, dotted lines represent the lower and upper bounds of confidence sets derived from standard t-test. In plot (a) and (b), where either a jump or kink is used for identification, the treatment effect is just identified and solid lines represent the lower and upper bounds of confidence sets from inverting a robust test. In plot (c), the treatment effect is over identified since both jump and kink are used. The confidence set derived from the CLR test is denoted by solid lines and the confidence set derived from the AR test is denoted by dashed lines.

Figure 1.8 shows that the standard and robust inference yield very similar results when the identification is strong: the confidence sets from inverting a standard test and a robust test are almost identical if the bandwidth is larger than 150, which means more data used in estimation and stronger identification. However, when the bandwidth is small and weak identification becomes a problem, the standard and robust inference yield significantly different results. The confidence

[20] Card and Yakovlev (2014) used only the kink for identification, while Dong (2016) used both jump and kink for identification. They are all valid if the true jump is zero or the treatment effect derivative is zero.




(a) Fuzzy RD (b) Fuzzy RK (c) RPJK

Figure 1.8: Confidence sets for the treatment effect in different designs

sets from inverting a standard test is a lot smaller than those from inverting a robust test. It is worth noting that smaller confidence sets from standard inference in the case of weak identification do not suggest that they are more informative. On the contrary, the information they provide is not reliable because they are derived from invalid tests. In other words, the confidence sets from standard inference are artificially small and very likely to undercover the true parameters.

## 1.6 Conclusion

Previous authors, e.g. Feir et al. (2016), have noted the weak identification problem in fuzzy RD designs and proposed a null-restricted t-test. I approach the same problem by drawing insights from the extensive literature on IV regressions. Specifically, the AR test, the LM test and the CLR test are considered and tailored to the settings of fuzzy RD designs. Different from Feir et al. (2016), I do not limit the attention to the most standard fuzzy RD design, where the identification only relies on a jump in the treatment probability. I consider a more general case where the identification relies on a jump, a kink or both in the treatment probability. Thus, the proposed tests can be applied in a wide range of research questions. In addition, I explicitly correct the bias in estimation rather than assume undersmoothing. As a result, a larger set of bandwidths are allowed, including the MSE optimal bandwidth which is available in many statistic packages. Though IV regression is an alternative option when estimating a fuzzy RD design, I still recommend the latter because

of its flexibility in estimation. The proposed tests not only have correct size, but also have good power properties when this flexibility is properly explored.

# CHAPTER 2. USING WILD BOOTSTRAP TO CONSTRUCT CONFIDENCE INTERVALS IN FUZZY REGRESSION DISCONTINUITY DESIGNS

A new wild bootstrap procedure is proposed to correct bias and construct valid confidence intervals in fuzzy regression discontinuity designs. This procedure uses a wild bootstrap based on second order local polynomials to estimate and remove the bias from linear models. The bias-corrected estimator is then bootstrapped itself to generate valid confidence intervals. While the conventional confidence intervals generated by adopting MSE-optimal bandwidth is asymptotically not valid, the confidence intervals generated by this procedure have correct coverage under conditions similar to Calonico, Cattaneo and Titiunik's(2014, *Econometrica*) analytical correction. Simulation studies provide evidence that this new method is as accurate as the analytical corrections when applied to a variety of data generating processes featuring heteroskedasticity, endogeneity and clustering. As an example, its usage is demonstrated through a reanalysis of the scholastic achievement data used by Angrist and Lavy (1999).

## 2.1 Introduction

The idea of regression discontinuity (RD) design was first used by Thistlethwaite and Campbell (1960) to estimate the causal effect of merit awards on future academic outcomes. In their application, the discontinuity in receiving merit awards as a function of test scores (refereed to in literature as "forcing variable" or "running variable", which determines the treatment assignment) creates a local randomized experiment, which allows researchers to identify the causal effect at the point of discontinuity. The idea of RD designs did not get much attention from economists in its early years, but the past decade has seen its increasing popularity in analyzing the causal impact

of policies and interventions in social science. Imbens and Lemieux (2008) and Lee and Lemieux (2010) provide recent reviews of this literature with many examples.

The identification in RD designs relies on the assumption that units arbitrarily close to the cutoff are credibly similar in predetermined characteristics. Under this "smoothness" condition, one can essentially compare units slightly above the cutoff and units slightly below the cutoff, and the difference in outcomes can be thought of as being induced by exogenous changes in treatment, giving it an interpretation of treatment effect. When the running variable completely determines the treatment, the probability of being treated jumps from zero to one at the cutoff (sharp RD designs). On the contrary, when the running variable does not entirely determine the treatment, there are both treated and untreated units on each side of the cutoff. This treatment misassignment was studied in a series of work by Trochim and Spiegelman (1980) and Trochim (1984) and was called "fuzzy" RD design thereafter. Directly comparing the outcomes on both sides of the cutoff results in an "intent-to-treat" effect but not the actual treatment effect because this difference is contributed only by part of the units. As in a Wald formulation of the treatment effect in an instrumental variable setting, the true treatment effect can be recovered by taking the ratio of difference in outcomes and difference in treatment probabilities at the cutoff. Even when units are self-selected to treatment based on anticipated gains, Hahn et al. (2001) show that this ratio can be interpreted as the local average treatment effect (LATE) under proper assumptions.

The identification of RD designs occurs exactly at the cutoff, which unavoidably requires extrapolation. Established by Fan (1992) and advocated by Hahn et al. (2001), the desirable boundary property of local linear models makes them almost standard practice in estimating RD designs. An important tunning variable in these nonparametric models is the bandwidth $h$, which controls the trade-off between bias and variance. One very popular choice of this tunning variable under the setting of RD designs is the bandwidth selector proposed by Imbens and Kalyanaraman (2012), which minimizes the asymptotic mean squared error (AMSE) of the treatment effect estimator. This bandwidth selector has the form $h = O_p(n^{-1/5})$, where $n$ is the number of observations. However, as is shown by Hahn et al. (2001), a bandwidth choice of $h = O_p(n^{-1/5})$ leads to an asymptotic

normal distribution of the treatment effect estimator centered at the true treatment effect plus a non-negligible bias term. Ignoring this bias term invalidates confidence intervals based on Wald test. Simulation studies on sharp RD designs in Calonico et al. (2014), henceforth "CCT," also confirm that conventional confidence intervals have empirical coverage well below their nominal levels. As a result, it is common practice to use ad-hoc bandwidths which shrink at a rate more than $n^{-1/5}$ so that the bias term vanishes faster in a hope that the bias will not affect asymptotic approximation.

CCT solve this problem by firstly re-centering the conventional point estimator with estimated bias term and then rescaling it by a unconventional standard error which takes into consideration the additional variability of the estimated bias. This approach results in a bias-corrected point estimator which is asymptotically normal under weaker assumptions on the bandwidth choice. Confidence intervals based on this method are accurate even when the AMSE optimal bandwidths are used.

In this chapter, a wild bootstrap procedure is proposed as an alternative to CCT's robust inference method for fuzzy RD designs. It is theoretically proved that the new bootstrap procedure is asymptotically equivalent to CCT's and supported by simulations that it performs well with finite sample. Compared with CCT's analytical method, the bootstrap procedure is very straightforward and does not require intensive analytical derivations. In addition, since the bootstrap is motivated by mimicking the true data generating process, it has the flexibility to accommodate dependent data by adjusting the resampling algorithm accordingly. In particular, this chapter demonstrates how the proposed bootstrap procedure can be applied to clustered data and perform at least as good as the analytical robust method.

The wild bootstrap procedure exploits CCT's theoretical insight by resampling from higher order local polynomials. In particular, the local linear models are estimated as usual for both outcome and treatment, resulting in a conventional biased estimator. To estimate the bias, additional local quadratic models are estimated. These second order polynomials together with the potentially correlated residuals represent the true data generating process (DGP) for bootstrap. The bias of

the conventional estimator from local linear models is therefore known under this bootstrap DGP and can be calculated by averaging the error of the linear model's estimates across many bootstrap replications. Though the local quadratic models are also not bias free, it can be shown that its bias converges to zero at a faster rate, fast enough that the bias of the local linear model can be estimated and removed using the second order polynomial. This approach is described in detail by Algorithm 2.1 and the resulting bias-corrected estimator is shown to be asymptotically normal with mean zero in Theorem 2.1.

This bias correction procedure introduces additional variability because the bias is calculated by assuming that local quadratic models represent the true DGP. However, these local quadratic models also come with uncertainty because of sampling error. So an iterated bootstrap procedure (Hall and Martin, 1988) is adopted to accommodate this additional variability: generate many bootstrap datasets from local quadratic models and calculate bias-corrected estimate for each of them. The resulting empirical distribution of bias-corrected estimate is then used to construct confidence intervals. This procedure is in line with CCT's approach, where the variance of estimated bias term and the covariance between estimated bias and original point estimator are derived analytically. This complex adjustment to the original variance is automatically embedded in the iterated bootstrap. The detailed implementation steps are described in Algorithm 2.2, and the resulting confidence intervals are shown to be asymptotically valid in Theorem 2.2.

This chapter is closely related to the work by Bartalotti et al. (2017b), who look at the robust inference in sharp RD designs, which are special cases of RD designs. The current chapter provides important generalization in several dimensions. First, it borrows the idea of bootstrapping IV models and adapts that to a more general fuzzy RD design. Second, its validity is extended and theoretically proved to any order of local polynomials and any order of derivatives of interests. Lastly, its flexibility and capability to accommodate clustered data is discussed and confirmed by simulation studies.

The chapter is organized as follows. Section 2.2 describes the basic fuzzy RD approach, its usual implementation, and the CCT's robust inference method. Section 2.3 presents the proposed boot-

strap procedures to estimate bias and construct confidence interval. Their asymptotic properties are discussed and summarized in two theorems. Section 2.4 provides simulation evidence that the bootstrap procedure effectively reduces bias and generates valid confidence intervals. An extended application to clustered data is discussed in Section 2.5. Section 2.6 demonstrates the usage of this bootstrap procedure by applying it to the scholastic achievement data used by Angrist and Lavy (1999). Finally, Section 2.7 concludes.

## 2.2 Background

This section provides additional details of identification assumptions and traditional estimation methods in fuzzy RD designs. It also briefly introduces the robust confidence interval proposed by CCT. Notations defined in this section and following sections are consistent with CCT where possible to aid readers familiar with that paper.

In a typical fuzzy RD setting, researchers are interested in the local causal effect of treatment at a given cutoff. For any unit $i$, a triple $(X_i, T_i, Y_i)$ is observed, where $X_i$ is a continuous running variable which determines treatment assignment, $T_i$ is a binary variable which indicates actual treatment status and $Y_i$ is the outcome. In sharp RD designs, the treatment actually received is the same as the assigned treatment, i.e., $T_i = \mathbb{1}(X_i \geq c)$, with $c$ being the cutoff. In fuzzy RD designs, however, the received treatment is not a deterministic function of running variable $X_i$. Instead, the probability $\Pr(T_i = 1 \mid X_i)$ is between zero and one in both sides but experiences a sudden change at the cutoff. For subject $i$, we use $T_i(1)$ to denote the actual treatment if assigned to treatment group ($X_i \geq 0$), and $T_i(0)$ if assigned to the control group ($X_i < 0$). Analogously, we use $Y_i(1)$ to denote the outcome if $i$ is actually in the treatment group (when $T_i = 1$), and $Y_i(0)$ if not (when $T_i = 0$).

In practice, the observed random sample is $\{(Y_i, T_i, X_i)_{i=1,2,\ldots,n}\}$, where $T_i = \mathbb{1}(X_i \geq 0)T_i(1) + \mathbb{1}(X_i < 0)T_i(0)$ and $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$, with $\mathbb{1}(\cdot)$ being the indicator function. The

parameter of interest is

$$\zeta = \frac{\tau_Y}{\tau_T} = \frac{\lim_{x \to 0^+} \mathbb{E}(Y_i | X_i = x) - \lim_{x \to 0^-} \mathbb{E}(Y_i | X_i = x)}{\lim_{x \to 0^+} \mathbb{E}(T_i | X_i = x) - \lim_{x \to 0^-} \mathbb{E}(T_i | X_i = x)}, \tag{2.1}$$

where the symbol $\mathbb{E}$ represents the expectation and $\tau_Y$ and $\tau_T$ represent the sharp RD estimators, i.e., difference in expectations at the cutoff. Intuitively, this is a Wald estimator in the limit where the assigned treatment serves as an instrument. The reduced-form difference in expected outcome, $\tau_Y$, reveals the "intent-to-treat" (ITT) effect. The treatment effect is recovered by dividing ITT effect by the first stage difference in treatment probabilities. When the treatment effect is not constant across units, $\zeta$ should be interpreted with caution. If treatment status is independent of treatment effects at the cutoff, $\zeta$ is the average treatment effect (ATE) at the cutoff. This assumption rules out self-selection based on anticipated gain. Hahn et al. (2001) show that under a less restrictive assumption that the running variable is independent of the joint distribution of treatment effect and treatment status at the cutoff, the local average treatment effect (LATE) is identified.

The formula for $\zeta$ shows that it is a ratio of two sharp RD estimators. Due to this symmetry, I use "$Z$" as a placeholder for either outcome variable $Y$ or treatment variable $T$ to ease the notation. In addition, I introduce conditional expectations $\mu_{Z+}(x)$ and $\mu_{Z-}(x)$, conditional variances $\sigma^2_{Z+}(x)$ and $\sigma^2_{Z-}(x)$, the $\eta$th order derivative of conditional expectations $\mu^{(\eta)}_{Z+}(x)$ and $\mu^{(\eta)}_{Z-}(x)$ and their limits. Formally, they are defined as

$$\begin{aligned}
\mu_{Z+}(x) &= \mathbb{E}(Z_i(1) \mid X_i = x) & \mu_{Z-}(x) &= \mathbb{E}(Z_i(0) \mid X_i = x) \\
\sigma^2_{Z+}(x) &= \mathbb{V}(Z_i(1) \mid X_i = x) & \sigma^2_{Z-}(x) &= \mathbb{V}(Z_i(0) \mid X_i = x) \\
\mu^{(\eta)}_{Z+}(x) &= \frac{d^\eta \mu_{Z+}(x)}{dx^\eta} & \mu^{(\eta)}_{Z-}(x) &= \frac{d^\eta \mu_{Z-}(x)}{dx^\eta} \\
\mu^{(\eta)}_{Z+} &= \lim_{x \to 0^+} \mu^{(\eta)}_{Z+}(x) & \mu^{(\eta)}_{Z-} &= \lim_{x \to 0^-} \mu^{(\eta)}_{Z-}(x)
\end{aligned}$$

where the symbol $\mathbb{V}(\cdot)$ represents variance. The treatment effect $\zeta$ is nonparametrically estimable because $\mu_{Z-}$ and $\mu_{Z+}$ can be estimated consistently under Assumption 2.1, which lists standard conditions in the fuzzy RD literature. (See, in particular, (Hahn et al., 2001), (Porter, 2003) and CCT.)

**Assumption 2.1** (Behavior of the DGP near the cutoff)**.** *The random variables* $\{X_i, T_i, Y_i\}_{i=1}^n$ *form a random sample of size* $n$*. There exists a positive number* $\kappa_0$ *such that the following conditions hold for all* $x$ *in the neighborhood* $(-\kappa_0,\ \kappa_0)$ *around zero:*

1. *The density of* $X_i$ *is continuous and bounded away from zero at* $x$*.*
2. $\mathbb{E}[Z_i^4 \mid X_i = x]$ *is bounded.*
3. $\mu_{Z-}(x)$ *and* $\mu_{Z+}(x)$ *are three times continuously differentiable.*
4. $\sigma^2_{Z-}(x)$ *and* $\sigma^2_{Z+}(x)$ *are continuous and bounded away from zero.*
5. $\mu_{T-}(0) \neq \mu_{T+}(0)$*.*

Part 1 ensures that the number of data points arbitrarily close to the cutoff increases as the sample size grows. Part 3 imposes necessary smoothness condition to allow an approximation by second order polynomials. Part 2 and 4 put standard restrictions on moments to ensure that the estimated local polynomials are well behaved. Part 5 requires that the treatment assignment as an instrument is valid, in the sense that it induces a first stage difference in treatment probability. In practice, local polynomial regression is widely used to estimate RD designs because of nice boundary properties.[1] As an illustration, I focus here on local linear regression using kernel function $K(\cdot)$. For simplicity, suppose a common bandwidth, $h$, is chosen for both the outcome and the treatment, the estimated treatment effect is

$$\hat{\zeta}(h) = \frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)} = \frac{\hat{\mu}_{Y+}(h) - \hat{\mu}_{Y-}(h)}{\hat{\mu}_{T+}(h) - \hat{\mu}_{T-}(h)}, \tag{2.2}$$

with

$$\hat{\mu}_{Z+}(h) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} \mathbb{1}\{X_i \geq 0\}(Z_i - \beta_0 - X_i\beta_1)^2 \frac{1}{h} K(\frac{X_i}{h})$$

and

$$\hat{\mu}_{Z-}(h) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} \mathbb{1}\{X_i < 0\}(Z_i - \beta_0 - X_i\beta_1)^2 \frac{1}{h} K(\frac{X_i}{h}).$$

[1] See Fan and Gijbels (1996) for discussions on the boundary properties of local polynomial regression. See Gelman and Imbens (2017) for discussions on the choices of global and local polynomial regression and its order.

The conditional expectations $\mu_{Z+}$ and $\mu_{Z-}$ are consistently estimated by $\hat{\mu}_{Z+}(h)$ and $\hat{\mu}_{Z+}(h)$ when $h \to 0$.[2] The asymptotic distribution of the quotient estimator $\frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)}$ can be derived by applying the delta method. Let $V_Z$ be the asymptotic variance of $\hat{\tau}_Z(h)$ and $C_{YT}$ be the asymptotic covariance between $\hat{\tau}_Y(h)$ and $\hat{\tau}_T(h)$, i.e.,

$$\begin{pmatrix} \sqrt{nh}(\hat{\tau}_Y(h) - \tau_Y) \\ \sqrt{nh}(\hat{\tau}_T(h) - \tau_T) \end{pmatrix} \to^d N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_Y & C_{YT} \\ C_{YT} & V_T \end{pmatrix} \right),$$

then it follows that

$$\sqrt{nh}(\hat{\zeta}(h) - \zeta) \to^d N(0, \frac{1}{\tau_T^2} V_Y - \frac{2\tau_Y}{\tau_T^3} C_{YT} + \frac{\tau_Y^2}{\tau_T^4} V_T).$$

Let $V(h) = \mathbb{V}(\hat{\zeta}(h) \mid X_1, ..., X_n)$, then $\frac{\hat{\zeta}(h)-\zeta}{\sqrt{V(h)}} \to^d N(0,1)$ and the confidence intervals can be constructed as

$$\hat{\zeta}(h) \pm q_{1-\alpha/2} V(h)^{1/2} \tag{2.3}$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The above asymptotic distribution is valid only when bandwidth $h$ shrinks fast enough such that the bias of $\hat{\zeta}_Z(h)$ is negligible relative to $\sqrt{V(h)}$. Formally, $h = o_p(n^{-1/5})$ is required. With a bandwidth of order $O_p(n^{-1/5})$, Hahn et al. (2001) show that the asymptotic distribution is normal but not centered at zero. Using (2.3) to construct confidence intervals without considering this first-order bias in distributional approximation leads to a coverage rate lower than the nominal level. Imbens and Kalyanaraman (2012) develop plug-in bandwidth selector for RD estimators, which is optimal in the sense that squared error loss of the point estimator is minimized. Ludwig and Miller (2005) propose using cross validation to select bandwidth which minimizes squared prediction errors but find that the loss function is very "flat" and leads to relatively large bandwidth.

Two different approaches are adopted in empirical studies. One is undersmoothing. In this case, instead of using the MSE-optimal bandwidth, researchers may want to choose a smaller bandwidth in order to meet the requirement of $h = o_p(n^{-1/5})$. However, this often leads to a series of ad-hoc bandwidths without theoretical basis. Another approach is bias correction. In this case, the leading

[2] Unless otherwise stated, all limits in this chapter are assumed to hold as $n \to \infty$.

bias is consistently estimated in an attempt to remove distortion of the asymptotic approximation. However, this approach does not perform well because the estimated bias introduces additional variability. The CCT's approach is based on bias correction, but redefines the variance component for normalization so that the additional variability is accounted for.

For any bandwidth $h \to 0$, the first-order bias of fuzzy RD estimator from local linear regression is

$$\mathbb{E}(\hat{\zeta}(h) \mid X_1, ..., X_n) - \zeta = h^2(\frac{1}{\tau_T}\boldsymbol{B}_Y(h) - \frac{\tau_Y}{\tau_T^2}\boldsymbol{B}_T(h))(1 + o_p(1)), \tag{2.4}$$

with

$$\boldsymbol{B}_Z(h) = \frac{\mu_{Z+}^{(2)}}{2}\mathfrak{B}_+(h) - \frac{\mu_{Z-}^{(2)}}{2}\mathfrak{B}_-(h).$$

The terms $\mathfrak{B}_+(h)$ and $\mathfrak{B}_-(h)$, explicitly defined in appendix, are observed quantities that depend on the kernel, bandwidth and running variable. To explicitly calculate the first-order bias, one needs to estimate $\tau_Z$, $\mu_{Z+}^{(2)}$ and $\mu_{Z-}^{(2)}$. Among them $\tau_Z$ is consistently estimated by the local linear regression with bandwidth $h$. CCT propose a local second-order regression with a (potentially) different bandwidth $b$ to estimate the second order derivatives $\mu_{Z+}^{(2)}$ and $\mu_{Z-}^{(2)}$. This procedure gives the bias-corrected estimator

$$\hat{\zeta}^{bc}(h, b) = \hat{\zeta}(h) - \Delta(h, b),$$

with

$$\Delta(h, b) = h^2(\frac{1}{\hat{\tau}_T(h)}\hat{\boldsymbol{B}}_Y(h, b) - \frac{\hat{\tau}_Y(h)}{\hat{\tau}_T^2(h)}\hat{\boldsymbol{B}}_T(h, b)),$$
$$\hat{\boldsymbol{B}}_Z(h, b) = \frac{\hat{\mu}_{Z+}^{(2)}(b)}{2}\mathfrak{B}_+(h) - \frac{\hat{\mu}_{Z-}^{(2)}(b)}{2}\mathfrak{B}_-(h).$$

Notice that the bias $\Delta(h, b)$ is estimated with uncertainty. As a result, the variance of bias-corrected estimator $\hat{\zeta}^{bc}(h, b)$ is different from $V(h)$. CCT propose a new formula for the variance of bias-corrected estimator and use it for normalization:

$$\frac{\hat{\zeta}^{bc}(h, b) - \zeta}{V^{bc}(h, b)^{1/2}} \to^d N(0, 1), \tag{2.5}$$

where $V^{bc}(h, b) = V(h) + C(h, b)$ and $C(h, b)$ captures the adjustment to variance introduced by the bias-correction term. This distributional approximation is valid even when $h = O_p(n^{-1/5})$, as

long as certain conditions on $h$ and $b$ are satisfied. Assumption 2.2 specifies the bandwidth and kernel conditions assumed by CCT, which I will also use in this chapter.

**Assumption 2.2** (Bandwidth and kernel)**.** *Let $h$ be the bandwidth used to estimate the local linear model and let $b$ be the bandwidth used to estimate the second local quadratic model. Then $nh \to \infty$, $nb \to \infty$, and $n \times \min(h,b)^5 \times \max(h,b)^2 \to 0$ as $n \to \infty$. The kernel function $K(\cdot)$ is positive, bounded, and continuous on the interval $[-\kappa,\ \kappa]$ and zero outside that interval for some $\kappa > 0$.*

Assumption 2.2 does not require $nh^{1/5} \to 0$. Instead, it only requires that $nh^{1/5}b^{1/2} \to 0$ when $h < b$ or $nb^{1/5}h^{1/2} \to 0$ when $h > b$. This assumption also allows for the vast majority kernel functions commonly used in practice.

To simplify notation, let $m = \min(h,b)$ and define the scaled and truncated kernel functions

$$K_{+,h}(x) = \tfrac{1}{h}K(x/h)\,\mathbb{1}\{x \geq 0\} \qquad K_{-,h}(x) = \tfrac{1}{h}K(x/h)\,\mathbb{1}\{x < 0\}$$

and

$$K_{+,b}(x) = \tfrac{1}{b}K(x/b)\,\mathbb{1}\{x \geq 0\} \qquad K_{-,b}(x) = \tfrac{1}{b}K(x/b)\,\mathbb{1}\{x < 0\}.$$

In the next section, a simple bootstrap procedure is proposed to construct robust confidence intervals based on the insight provided by CCT's bias-corrected estimator. This bootstrap procedure is straightforward in the sense that no derivation of analytical formulas for the bias, variance and covariance terms is required. The bias-corrected estimator and its confidence interval are numerically different from CCT's but asymptotically equivalent.

## 2.3 Bootstrap Algorithm

In this section, two bootstrap algorithms are presented to obtain bias-corrected point estimator and its confidence intervals in the fuzzy RD designs. Their correctness is justified in two theorems and proved in the appendix. The idea behind both algorithms is to use local second-order polynomials to approximate the distribution of $(X_i, T_i, Y_i)$ around the cutoff. These second order polynomials, together with the variance structure, have known properties and act as the "true"

DGP as sample size increases. Assumption 2.2 guarantees that the estimated "true" DGP is close to the unknown DGP in the sense that distributional approximation derived from the "true" DGP is asymptotically valid. This can be best illustrated from the special case where $h = b$, which translates to $nb^7 \to 0$ under Assumption 2.2. By the same argument that $h = o_p(n^{-1/5})$ is required for valid inference in a RD design estimated by local linear regression, $b = o_p(n^{-1/7})$ is required in a RD design estimated by local quadratic regression.

The first algorithm consistently estimates the bias term. In particular, after fitting local second-order regressions of outcome $Y_i$ and $T_i$ on running variable $X_i$ at each side of the cutoff, one can create many datasets through residual bootstrap. Each dataset generates a traditional fuzzy RD estimate, which are used to calculate the bias. Below is the detailed procedures in Algorithm 2.1.

**Algorithm 2.1** (Bias estimation)**.** *Assume $h$ and $b$ are bandwidths as defined by Assumption 2.2.*

1. *Estimate local second order polynomials $\hat{g}_{Z-}$ and $\hat{g}_{Z+}$ with least squares using $K_{-,b}$ and $K_{+,b}$ for weights:*

$$\hat{g}_{Z-}(x) = \hat{\beta}_{Z-,0} + \hat{\beta}_{Z-,1}x + \hat{\beta}_{Z-,2}x^2, \qquad \hat{g}_{Z+}(x) = \hat{\beta}_{Z+,0} + \hat{\beta}_{Z+,1}x + \hat{\beta}_{Z+,2}x^2$$

*with*

$$(\hat{\beta}_{Z-,0}, \hat{\beta}_{Z-,1}, \hat{\beta}_{Z-,2})' = \underset{\beta_0,\beta_1,\beta_2}{\arg\min} \sum_{i=1}^{n} (Z_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2 K_{-,b}(X_i)$$
$$(\hat{\beta}_{Z+,0}, \hat{\beta}_{Z+,1}, \hat{\beta}_{Z+,2})' = \underset{\beta_0,\beta_1,\beta_2}{\arg\min} \sum_{i=1}^{n} (Z_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2 K_{+,b}(X_i).$$

*Let*

$$\hat{g}_Z(x) = \begin{cases} \hat{g}_{Z-}(x) & \text{if } x < 0 \\ \hat{g}_{Z+}(x) & \text{otherwise} \end{cases}$$

*and calculate the residuals $\hat{\varepsilon}_{Zi} = Z_i - \hat{g}_Z(X_i)$ for all $i$.*

2. *Repeat the following steps $B_1$ times to produce the bootstrap estimates $\hat{\eta}_1^*(h), \ldots, \hat{\eta}_{B_1}^*(h)$. For the $k$th replication:*

(a) *Draw i.i.d. random variables* $e_i^*$ *with mean zero, variance one, and bounded fourth moments independent of the original data and construct*

$$\varepsilon_{Zi}^* = \hat{\varepsilon}_{Zi} e_i^*,$$

*and*

$$Z_i^* = \hat{g}_Z(X_i) + \varepsilon_{Zi}^*$$

*for all* $i$.

(b) *Calculate* $\hat{\mu}_{Z+}^*(h)$ *and* $\hat{\mu}_{Z-}^*(h)$ *by estimating the local linear model on the bootstrap data set using* $K_{+,h}$ *and* $K_{-,h}$ *for weights:*

$$\hat{\mu}_{Z-}^*(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{-,h}(X_i)$$
$$\hat{\mu}_{Z+}^*(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{+,h}(X_i).$$

(c) *Save* $\hat{\zeta}_k^*(h) = \frac{\hat{\mu}_{Y+}^*(h) - \hat{\mu}_{Y-}^*(h)}{\hat{\mu}_{T+}^*(h) - \hat{\mu}_{T-}^*(h)}$.

3. *Estimate the bias as*

$$\Delta^*(h, b) = \frac{1}{B_1} \sum_{k=1}^{B_1} \hat{\zeta}_k^*(h) - \frac{\hat{g}_{Y+}(0) - \hat{g}_{Y-}(0)}{\hat{g}_{T+}(0) - \hat{g}_{T-}(0)}. \tag{2.6}$$

Algorithm 2.1 consists of three steps. The first step estimates the bootstrap DGP, which is captured by second order local polynomials. The second step creates a series of new samples through wild bootstrap and finds the traditional fuzzy RD estimate for each sample. Notice that pairs of residuals are multiplied by the same realization of random number $e^*$ to preserve the correlation between $Y_i$ and $T_i$.[3] In addition, the fact that $T_i^*$ in bootstrap sample is no longer binary does not impact the validity of the algorithm because mean function and heteroskedasticity are preserved. The last step calculates the bias from local linear estimator by definition. Under Assumption 2.1, 2.2 and assume that $B_1$ is large enough, the procedure described by Algorithm 2.1

[3]If two independent random variables are used to generate $Y_i^*$ and $T_i^*$ respectively, $Y_i^*$ and $T_i^*$ will also be independent from each other.

gives a consistent estimator of the bias component that converges fast enough in probability that it can be used as a correction, resulting in a bias-corrected estimator that has the same asymptotic distribution as in (2.5). This conclusion is formally given in Theorem 2.1.

**Theorem 2.1.** *Under Assumptions 2.1 and 2.2,*

$$\frac{\hat{\zeta}(h) - \Delta^*(h,b) - \zeta}{V^{bc}(h,b)^{1/2}} \to^d N(0,1), \tag{2.7}$$

*where* $\Delta^*(h,b)$ *is defined by equation* (2.6).

Theorem 2.1 enables one to construct valid confidence interval in the form of $\hat{\zeta}(h) - \Delta^*(h,b) \pm V^{bc}(h,b)^{1/2}$. However, the term $V^{bc}(h,b)$ still needs to be calculated. The second algorithm circumvents the analytical derivation of $V^{bc}(h,b)$ through an iterated bootstrap. In particular, the first layer bootstrap is designed to mimic the randomness due to sampling error and the second layer bootstrap, as described in Algorithm 2.1, is designed to estimate bias due to model misspecification. The additional variability introduced by the bias correction term will be automatically accounted for by this iterated bootstrap. The detailed procedure is given in Algorithm 2.2.

**Algorithm 2.2** (Distribution)**.** *Assume* $h$ *and* $b$ *are bandwidths as defined by Assumption 2.2 and Algorithm 2.1.*

1. *Estimate* $\hat{g}_{Z+}$ *and* $\hat{g}_{Z-}$ *and generate* $\hat{g}_Z(\cdot)$ *and the residuals* $\hat{\varepsilon}_{Zi}$ *just as in Algorithm 2.1.*

2. *Repeat the following steps* $B_2$ *times to produce bootstrap estimates of the bias-corrected estimate. For the* $k$*th replication:*

   (a) *Draw i.i.d. random variables* $e_i^*$ *with mean zero, variance one, and bounded fourth moments independent of the original data and construct*

   $$\varepsilon_{Zi}^* = \hat{\varepsilon}_{Zi} e_i^*,$$

   *and*

   $$Z_i^* = \hat{g}_Z(X_i) + \varepsilon_{Zi}^*.$$

   *for all* $i = 1, \ldots, n$.

(b) *Calculate* $\hat{\mu}^*_{Z+}(h)$ *and* $\hat{\mu}^*_{Z-}(h)$ *by estimating the local linear model on the bootstrap data set using* $K_{+,h}$ *and* $K_{-,h}$ *for weights:*

$$\hat{\mu}^*_{Z-}(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{-,h}(X_i),$$
$$\hat{\mu}^*_{Z+}(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{+,h}(X_i).$$

(c) *Apply Algorithm 2.1 to the bootstrapped data set* $(X_1, T_1^*, Y_1^*), \ldots, (X_n, T_n^*, Y_n^*)$ *using the same bandwidths* $h$ *and* $b$ *that are used in the rest of this algorithm but reestimating all of the local polynomials on the bootstrap data. Generate* $B_1$ *new bootstrap samples and let* $\Delta^{**}(h, b)$ *represent the bias estimator returned by Algorithm 2.1.*

(d) *Save the estimator* $\hat{\zeta}^*_k(h) = \frac{\hat{\mu}^*_{Y+}(h) - \hat{\mu}^*_{Y-}(h)}{\hat{\mu}^*_{T+}(h) - \hat{\mu}^*_{T-}(h)}$, *and its bias* $\Delta^{**}_k(h, b)$.

3. *Define* $\zeta^* = \frac{\hat{g}_{Y+}(0) - \hat{g}_{Y-}(0)}{\hat{g}_{T+}(0) - \hat{g}_{T-}(0)}$ *and use the empirical CDF of* $\hat{\zeta}^*_1(h) - \Delta^{**}_1(h, b) - \zeta^*, \ldots, \hat{\zeta}^*_{B_2}(h) - \Delta^{**}_{B_2}(h, b) - \zeta^*$ *as the sampling distribution of* $\hat{\zeta}(h) - \Delta^*(h, b) - \zeta$.

Algorithm 2.2 also consists of three steps. The first step estimates the bootstrap DGP, which is captured by second order local polynomials. The second step creates a series of new samples, to which the Algorithm 2.1 is applied. The last step uses the empirical distribution of bias-corrected estimator to construct confidence intervals. As before, $B_2$ is assumed large enough so that simulation error can be ignored. The validity of Algorithm 2.2 is established in the following theorem.

**Theorem 2.2.** *Under Assumptions 2.1 and 2.2,*

$$\mathbb{V}^*(\hat{\zeta}^*(h) - \Delta^{**}(h, b))/V^{bc}(h, b) \to^p 1$$

*and*

$$\sup_x \left| \Pr^* \left[ \frac{\hat{\zeta}^*(h) - \Delta^{**}(h, b) - \zeta^*}{\mathbb{V}^*(\hat{\zeta}^*(h) - \Delta^{**}(h, b))^{1/2}} \leq x \right] - \Pr \left[ \frac{\hat{\zeta}(h) - \Delta^*(h, b) - \zeta}{V^{bc}(h, b)^{1/2}} \leq x \right] \right| \to^p 0.$$

Theorem 2.2 enables one to construct confidence intervals in the following form:

$$\big(\hat{\zeta}(h) - \Delta^*(h, b) + \zeta^* - (\hat{\zeta}^*(h) - \Delta^{**}(h, b))_{1-\alpha/2}, \hat{\zeta}(h) - \Delta^*(h, b) + \zeta^* + (\hat{\zeta}^*(h) - \Delta^{**}(h, b))_{\alpha/2}\big),$$

where all the terms with superscript $*$ are defined in Algorithm 2.2. Different from the analytical one, this confidence interval is not centered at the bias-corrected point estimator. Several remarks on implementing these algorithms are listed below.

**Remark 2.1.** *The proposed bias correction differs from CCT's analytical formula in finite sample. While the analytical bias is obtained by firstly linearizing* $\mathbb{E}\left(\frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)} - \frac{\tau_Y}{\tau_T}\right)$ *and then only evaluating its first order terms, Algorithm 2.1 directly estimates* $\mathbb{E}\left(\frac{\hat{\tau}^*_Y(h)}{\hat{\tau}^*_T(h)} - \frac{\tau^*_Y}{\tau^*_T}\right)$ *through bootstrap. Both methods consistently estimate the bias.*

**Remark 2.2.** *When the original treatment is binary, the bootstrap sample will no longer have binary treatment. Though it creates some difficulty for interpretation, it does not invalidate the estimation and inference because its conditional expectation and covariance with outcome variable remain unchanged.*

**Remark 2.3.** *The iterated bootstrap is less computationally intensive than it appears to be because of two reasons. First, the wild bootstrap creates new residuals but leaves the regressors unchanged, which means the design matrices only need to be computed once even when they are repeatedly used in fitting local polynomials.*[4] *Second, the number of data points actually used in estimation is a lot smaller than the full sample.*

The bootstrap procedure used in these two algorithms is in line with conventional bootstrap procedure for IV regression. When generating new samples from an IV model using residual bootstrap, one usually first estimates both reduced equation and structural equation and then randomly draws residual pairs from these two equations. Here in the fuzzy RD designs, two reduced equations are estimated. Instead of randomly drawing residual pairs, wild bootstrap is adopted to accommodate potential heteroskedasticity.

Both the CCT's approach and the bootstrap approach presented above are robust to bandwidth choice, in the sense that traditional MSE-optimal bandwidth is allowed for valid inference,

[4] *To fit local polynomials is equivalent to estimate weighted least square, i.e., the estimated parameter is* $(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{Y}$*, where* $\mathbf{X}$ *is matrix of regressors and* $\mathbf{K}$ *is weighting matrix determined by kernel function. Both* $\mathbf{X}$ *and* $\mathbf{K}$ *are not affected by the bootstrap so one just need to compute* $(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}$ *once and then reused it in the bootstrap calculations. Then each bootstrap replication just requires a single matrix-vector multiplication.*

but they are not robust to weak identification.[5] The wild bootstrap requires an initial estimation of the model, based on which resampling is conducted. Weak instrument makes it difficult to precisely estimate the model and thus the approximation to the true data generating process is poor. Alternatives which can improve performance include imposing the null hypothesis or bootstrapping (asymptotically) pivotal statistics (Davidson and Flachaire, 2008; Cameron et al., 2008). Statistical inference from the two analytical methods relies on the assumption that the estimate is asymptotically normal, which is likely to be very skewed when identification is weak.

Evidence of the usefulness of the new procedure proposed above and its relative performance to the analytical bias correction proposed in CCT are presented in a series of Monte Carlo simulations in Section 2.4.

## 2.4 Simulation

The proposed bootstrap algorithms are applied to a variety of data generating processes (DGP). The baseline DGP is similar to CCT but re-designed to fit the fuzzy RD designs:

$$
\begin{aligned}
X_i &\sim 2 \times beta(2,4) - 1 \\
T_i &= \mathbb{1}\{u_{ti} \leq \Phi^{-1}(0.5 - \frac{c}{2})\}\, \mathbb{1}\{X_i < 0\} + \mathbb{1}\{u_{ti} \leq \Phi^{-1}(0.5 + \frac{c}{2})\}\, \mathbb{1}\{X_i \geq 0\} \\
Y_i &= \mu_j(X_i) + T_i\zeta_j + u_{yi},
\end{aligned}
$$

where $u_{ti} \sim N(0,1)$ and $c = 0.9$. The equation for $T_i$ indicates that $\mu_{T-} = 0.5 - c/2$ and $\mu_{T+} = 0.5 + c/2$. As a result, the expected treatment conditional on running variable is constant on both sides but the discontinuity at the cutoff is exactly $c$. In the equation for $Y_i$, the first part on the right, $\mu_j(X_i)$ with $j = 1, 2, 3$, is the conditional expected outcome without treatment, which is continuous at the cutoff. The second part on the right, $T_i\zeta_j$, captures the additive treatment effect. In particular, the conditional expected outcome takes the following forms:

[5] As a measurement of the strength of instrumental variable, the concentration parameter in the setting of fuzzy RD designs is determined by the effective sample size ($nh$), density of the running variable at the cutoff ($f(0)$), variance of the treatment variable at the cutoff ($\sigma^2_{T-}(0), \sigma^2_{T+}(0)$) and discontinuity in treatment probability ($\mu_{T+}(0) - \mu_{T-}(0)$) (Feir et al., 2016).

$$\mu_1(x) = \begin{cases} 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0 \\ 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise,} \end{cases}$$

$$\mu_2(x) = \begin{cases} 2.30x + 3.28x^2 + 1.45x^3 + 0.23x^4 + 0.03x^5 & \text{if } x < 0, \\ 18.49x - 54.81x^2 + 74.30x^3 - 45.02x^4 + 9.83x^5 & \text{otherwise,} \end{cases}$$

$$\mu_3(x) = \begin{cases} 1.27x + 3.59x^2 + 14.147x^3 + 23.694x^4 + 10.995x^5 & \text{if } x < 0 \\ 0.84x - 0.30x^2 + 2.397x^3 - 0.901x^4 + 3.56x^5 & \text{otherwise.} \end{cases}$$

These conditional mean functions are adapted from DGPs for sharp RD designs by preserving the curvature but removing the discontinuity at the cutoff. The first mean function is designed to match features of U.S. congressional election data (Lee, 2008; Imbens and Kalyanaraman, 2012). The second mean function is designed to match the relation between children mortality rate and county poverty rate from analysis of Head Start data (Ludwig and Miller, 2005). The last mean function is similar to the first one except for some coefficients. CCT motivates this in an attempt to generate plausible model with sizable distortion when conventional t-test is performed. The true treatment effects for these three models are $\zeta_1 = 0.04, \zeta_2 = -3.45, \zeta_3 = 0.04$.

To accommodate a variety of different error structure in empirical data, the following three cases are considered.

1. Baseline case. The simplest case where errors are independently and identically distributed:

$$\begin{pmatrix} u_{ti} \\ u^*_{yi} \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad \rho = 0, \quad u_{yi} = 0.1295 u^*_{yi}.$$

2. Heteroskedasticity. The disturbance term in the outcome equation has a standard error changing with the running variable, i.e., everything being the same as in the baseline case except $u_{yi} = (0.1295 + 9x_i^2)u^*_{yi}$.[6]

[6]The motivation is to keep the standard error unchanged from the homoskedastic case at the cutoff, so that estimators from these two cases are more comparable in the sense that they are equivalent in the limit.

3. Endogeneity. The treatment status is correlated with unobserved characteristics which affect the outcome. This is modeled by the correlation between $u_{ti}$ and $u_{yi}$, i.e., everything being the same as in the baseline case except $\rho \in \{-0.9, 0.9\}$.

In the implementation of Algorithm 2.1 and 2.2, the two-point distribution proposed in Mammen (1993) is adopted for creating bootstrap samples. This auxiliary distribution is

$$e_i^* = \begin{cases} \frac{1+\sqrt{5}}{2} & \text{with probability } \frac{\sqrt{5}-1}{2\sqrt{5}}, \\ \frac{1-\sqrt{5}}{2} & \text{otherwise,} \end{cases}$$

with zero mean and unit second and third moments. Its property ensures that the skewness of the bootstrap error terms is the same as the skewness of the residuals, which is a desirable condition not imposed in Algorithm 2.1 and 2.2.[7] In addition, the residuals are transformed before applying bootstrap because they are on average underestimated by least squares. Specifically, instead of directly using $\hat{\varepsilon}_{Zi}$, the "HC3" type transformation $\hat{\varepsilon}_{Zi}/(1 - H_{ii})$ is applied, with $H_{ii}$ being the diagonal element of projection matrix.[8] This is based on jackknife covariance estimator and is shown to outperform the original heteroskedasticity-robust covariance estimator (MacKinnon and White, 1985). Simulation studies by Davidson and Flachaire (2008) and MacKinnon (2013) also provide some evidence in favor of "HC3" transformation.

The bootstrap approach uses $B_1 = 500$ replications to compute bias and $B_2 = 999$ replications to obtain empirical distribution of bias-corrected estimator. Besides the bootstrap approach, two additional approaches are estimated for comparison: the CCT's robust estimator and the conventional estimator.[9] The two bandwidths for the bootstrap approach and the CCT's approach are the same and are obtained by utilizing bandwidth selector from CCT. The bandwidth used in the conventional approach is chosen by MSE-optimal bandwidth selector proposed by Imbens and

[7] Some later studies also show good properties of the simpler Rademacher distribution (Flachaire, 2005; Davidson and Flachaire, 2008).

[8] Local regressions project $\mathbf{K}^{1/2}\mathbf{Y}$ onto space of $\mathbf{K}^{1/2}\mathbf{X}$, with $\mathbf{K}$ being the weighting matrix determined by kernel function. So the projection matrix will be $\mathbf{K}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}^{1/2}$.

[9] All simulations are conducted with R software. Packages *rdrobust (V0.94)* and *RDD (V0.57)* are used to estimate the CCT's robust estimator and conventional RD estimator respectively. By default, the former one uses the nearest neighbor variance estimator and the latter one uses "HC1" type heteroskedasticity-robust variance estimator.

Kalyanaraman (2012).[10] These three approaches are applied to a total number of 5000 simulated samples with a sample size of 1000. Triangular kernel is used throughout all the simulations in this chapter.[11]

Simulation results are shown in Table 2.1 and 2.2. For the estimated treatment effect, its bias, standard error and root of mean squared error are reported in the first three columns. For the confidence interval, its empirical coverage and average length are reported in the fourth and fifth columns. The last three columns list the bandwidths used in the two robust methods ($h_{CCT}, b_{CCT}$) and the conventional method ($h_{IK}$). For each sample, the wild bootstrap approach uses the same bandwidths as the CCT's robust approach.

Table 2.1 presents these results for data with and without heteroskedasticity. The baseline case is listed in Panel A. The two robust methods, wild bootstrap and CCT's approach, generate point estimates with very similar bias and standard error (identical for DGP 1 and 3 and slightly different for DGP 2). In contrast, the conventional approach reports 3-5 times larger bias. This is not surprising since the two robust methods explicitly correct the bias. The conventional method also fails to deliver a valid interval (coverage rates are 68.1%, 2.6% and 87.2% for the three DGPs respectively). Improvement is achieved by the robust methods by reducing bias and increasing interval length. Except for DGP 2, they both generate intervals with empirical coverage close to the nominal level and the wild bootstrap is lightly better (93.1% VS 91.5% for DGP 1 and 95.3% VS 94.1% for DGP 3). However, for DGP 2, even the robust methods report great size distortion. This is because DGP 2 has great curvature around the cutoff and makes precise fitting very difficult.[12] Still, the two robust methods improve significantly from the conventional method in coverage (from 2.6% to around 87%) at the sacrifice of slightly longer intervals (from 0.186 to around 0.21).

[10] As is suggested by Imbens and Kalyanaraman (2012), the optimal bandwidth choices in fuzzy RD designs are often similar to those based on the optimal bandwidth for the numerator only. For simplicity, all bandwidths are calculated ignoring the fact that the RD design is fuzzy.

[11] Results with other kernel functions are similar and available in a separate document.

[12] In particular, the DGP 2 shows great curvature just right to the cutoff. On the right side, its second derivative at the cutoff is -109.62, so local linear regression is likely to create large bias. Its third derivative at the cutoff is 445.8, so local quadratic regression is likely to create large bias.

Table 2.1: Empirical coverage and average interval length

| DGP | Method | Bias | SD | RMSE | EC(%) | IL | $h_{CCT}$ | $b_{CCT}$ | $h_{IK}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | Panel A: homoskedastic data | | | | | | | |
| 1 | Wild bootstrap | 0.015 | 0.054 | 0.056 | 93.1 | 0.197 | 0.197 | 0.323 | |
| | CCT robust | 0.015 | 0.054 | 0.056 | 91.5 | 0.191 | 0.197 | 0.323 | |
| | Conventional | 0.042 | 0.032 | 0.053 | 68.1 | 0.116 | | | 0.400 |
| 2 | Wild bootstrap | 0.037 | 0.058 | 0.069 | 86.9 | 0.210 | 0.165 | 0.299 | |
| | CCT robust | 0.039 | 0.060 | 0.071 | 86.6 | 0.212 | 0.165 | 0.299 | |
| | Conventional | 0.215 | 0.079 | 0.229 | 2.6 | 0.186 | | | 0.216 |
| 3 | Wild bootstrap | 0.005 | 0.053 | 0.053 | 95.3 | 0.205 | 0.162 | 0.317 | |
| | CCT robust | 0.005 | 0.053 | 0.054 | 94.1 | 0.200 | 0.162 | 0.317 | |
| | Conventional | -0.025 | 0.044 | 0.050 | 87.3 | 0.157 | | | 0.205 |
| | | Panel B: heteroskedastic data | | | | | | | |
| 1 | Wild bootstrap | 0.004 | 0.078 | 0.079 | 95.8 | 0.294 | 0.110 | 0.189 | |
| | CCT robust | 0.004 | 0.071 | 0.071 | 94.0 | 0.268 | 0.110 | 0.189 | |
| | Conventional | 0.029 | 0.048 | 0.057 | 90.8 | 0.185 | | | 0.237 |
| 2 | Wild bootstrap | 0.028 | 0.066 | 0.072 | 92.9 | 0.255 | 0.149 | 0.259 | |
| | CCT robust | 0.030 | 0.067 | 0.073 | 91.3 | 0.251 | 0.149 | 0.259 | |
| | Conventional | 0.232 | 0.109 | 0.256 | 5.8 | 0.213 | | | 0.226 |
| 3 | Wild bootstrap | 0.001 | 0.069 | 0.069 | 96.2 | 0.294 | 0.110 | 0.190 | |
| | CCT robust | 0.001 | 0.069 | 0.069 | 94.4 | 0.267 | 0.110 | 0.190 | |
| | Conventional | -0.039 | 0.061 | 0.072 | 83.0 | 0.187 | | | 0.230 |

Note: EC denotes empirical coverage and IL denote average interval length based on 5000 simulations; nominal coverage probabilities are 95% for each estimator. The columns $h_{CCT}$ and $b_{CCT}$ list average optimal bandwidths following CCT's method.The column $h_{IK}$ lists average optimal bandwidth minimizing MSE.

Panel B in Table 2.1 lists the results when the data is heteroskedastic.[13] A significant difference from the homoskedastic case is the bandwidth choice. For the two robust methods, the bandwidths are reduced from the homoskedastic case while for the conventional method, this happens only to DGP 1. The increased noise in the data may reduce the perceived curvature by bandwidth selector and thus a smaller bandwidth is picked. Smaller bandwidth leads to smaller bias and larger variance. As a result, intervals in Panel B have higher coverage rate with longer interval length. The overall pattern in Panel B is similar to Panel A because all the three methods are robust to heteroskedasticity.

[13]In Panel A where homoskedastic DGP is used, there still exists heteroskedasticity from the perspective of estimation due to model specification, i.e., to use polynomials with order lower than the true one.

Table 2.2 presents results when the treatment is endogenous, which is almost always true and probably the primary reason to choose RD designs as the identification strategy. The case with positive self-selection is listed in Panel A and negative self-selection in Panel B. Again, the estimate from conventional method has significantly larger bias than the other two robust methods. As for interval estimates, the wild bootstrap and the CCT's approach work reasonably well in all cases except for DGP 2, where the empirical coverage is around 90% with positive self-selection and 85% with negative self-selection. The conventional method performs significantly worse, with empirical coverage rate as low as 1.7% (DGP 2 with negative self-selection). The sign of correlation has little effect on the bias because the bias is caused by model misspecification rather than imperfect instrumental variable.

To summarize, the wild bootstrap approach proposed in this chapter performs significantly better than the conventional method and is at least on par with the CCT's analytical methods. This wild bootstrap procedure automatically accommodate various types of covariance structure and thus is a simple alternative to obtain valid confidence intervals in RD designs.

## 2.5 Extension: Clustered Data

This section explores the application of the bootstrap procedure to clustered data in RD designs and provides evidence for its usefulness. Clustered data are very common in empirical studies. Units within the same cluster are usually dependent and ignoring this dependence is likely to invalidate statistic inference. There is enormous literature on handling clustered data.[14] In short, one can either explicitly estimate the dependence structure with some additional specifications, such as random coefficient models, or account for the dependence after estimation, such as using cluster-robust variance estimator (Liang and Zeger, 1986; Arellano, 1987).

To use cluster-robust variance estimator in statistical inference is very popular partly because it does not require assumption on the dependence structure and partly because its availability in almost all statistical software. Its validity is based on asymptotics when the number of clusters

[14]Specifically, see Wooldridge (2003); Cameron et al. (2012); Cameron and Miller (2015) for an overview on this topic.

Table 2.2: Empirical coverage and average interval length (endogenous treatment)

| DGP | Method | Bias | SD | RMSE | EC(%) | IL | $h_{CCT}$ | $b_{CCT}$ | $h_{IK}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | Panel A: $\rho = 0.9$ | | | | | | | |
| 1 | Wild bootstrap | 0.016 | 0.054 | 0.056 | 95.7 | 0.203 | 0.197 | 0.323 | |
| | CCT robust | 0.017 | 0.055 | 0.057 | 93.1 | 0.196 | 0.197 | 0.323 | |
| | Conventional | 0.043 | 0.033 | 0.054 | 70.7 | 0.121 | | | 0.398 |
| 2 | Wild bootstrap | 0.037 | 0.064 | 0.074 | 90.4 | 0.220 | 0.168 | 0.302 | |
| | CCT robust | 0.041 | 0.067 | 0.078 | 89.7 | 0.233 | 0.168 | 0.302 | |
| | Conventional | 0.226 | 0.092 | 0.244 | 3.0 | 0.207 | | | 0.222 |
| 3 | Wild bootstrap | 0.004 | 0.062 | 0.062 | 95.9 | 0.214 | 0.161 | 0.316 | |
| | CCT robust | 0.007 | 0.055 | 0.056 | 94.8 | 0.202 | 0.161 | 0.316 | |
| | Conventional | -0.024 | 0.043 | 0.049 | 86.5 | 0.156 | | | 0.204 |
| | | Panel B: $\rho = -0.9$ | | | | | | | |
| 1 | Wild bootstrap | 0.015 | 0.053 | 0.056 | 91.3 | 0.198 | 0.199 | 0.324 | |
| | CCT robust | 0.013 | 0.055 | 0.056 | 91.1 | 0.190 | 0.199 | 0.324 | |
| | Conventional | 0.042 | 0.031 | 0.052 | 65.7 | 0.113 | | | 0.402 |
| 2 | Wild bootstrap | 0.037 | 0.052 | 0.064 | 85.5 | 0.205 | 0.161 | 0.296 | |
| | CCT robust | 0.038 | 0.052 | 0.064 | 84.4 | 0.190 | 0.161 | 0.296 | |
| | Conventional | 0.201 | 0.064 | 0.211 | 1.7 | 0.165 | | | 0.208 |
| 3 | Wild bootstrap | 0.005 | 0.053 | 0.053 | 95.6 | 0.206 | 0.163 | 0.317 | |
| | CCT robust | 0.003 | 0.054 | 0.054 | 94.5 | 0.203 | 0.163 | 0.317 | |
| | Conventional | -0.027 | 0.045 | 0.052 | 89.1 | 0.160 | | | 0.207 |

Note: EC denotes empirical coverage and IL denote average interval length based on 5000 simulations; nominal coverage probabilities are 95% for each estimator. The columns $h_{CCT}$ and $b_{CCT}$ list average optimal bandwidths following CCT's method.The column $h_{IK}$ lists average optimal bandwidth minimizing MSE.

grows to infinity, which is, unfortunately, not trivial to establish in nonparametric models. The main obstacle is that shrinking tunning variable is likely to destroy the dependence structure. For local polynomial regressions, Wang (2003) and Chen et al. (2008) point out that the existence of joint density of running variable and clustering variable ensures that all clusters will eventually include only a single unit as the bandwidth shrinks to zero. As a result, the clustering structure disappears. A special case where this does not happen is that clustering occurs at the running variable level (Chen and Jin, 2005; Bartalotti and Brummet, 2017).[15]

[15]For example, in panel data where each individual are observed for multiple times and the running variable is at individual level, each individual is a cluster and will not vanish with shrinking bandwidth. Lee and Card (2008) consider another example in RD designs where clustering occurs at the running variable level and cluster-robust variance estimator is recommended in inference.

Though there is no asymptotics specifically developed for general RD designs with clustered data, currently available softwares usually provide options to take this dependence into consideration.[16] After all, the estimation of RD designs is no different from linear regression once the bandwidth is given and conventional cluster-robust variance estimator can be easily applied. Bartalotti (2018) adopted a fixed bandwidth framework to study general clustering in RD designs and proposed higher order correction based on bootstrap.

Bootstrap is also known to be applicable to clustered data. Cameron et al. (2008) provide a comprehensive survey of bootstrap method and show that proper bootstrap procedures outperform the conventional cluster-robust variance estimator when the number of clusters is small (five to thirty).

To check the flexibility and robustness of wild bootstrap procedure proposed in this chapter, I slightly revise the resampling algorithm to accommodate clustering and test its performance with pseudo clustered data. Following Brownstone and Valletta (2001) and Cameron et al. (2008), the wild bootstrap procedure for clustered data is quite straightforward: for units in the same cluster, their residuals are multiplied by the same random number drawn from the auxiliary distribution. For example,

$$Z^*_{gi} = \hat{g}_Z(X_{gi}) + \hat{\varepsilon}_{Z_{gi}} e^*_g,$$

where $e^*_g$, a random number from distribution with zero mean and unit variance, is shared by all units in the same group. For the purpose of simulation, it is assumed that errors in the outcome equation is clustered according to a random effect model, in particular,

$$u_{ygi} = u^*_{yg} + u^*_{yi}, \quad u^*_{yg}, u^*_{yi} \sim i.i.d.\ N(0, \frac{0.1295}{\sqrt{2}}),$$

with $g = 1, 2, \ldots, G$ being a group indicator. This design ensures that each individual error has a standard error of 0.1295, which is the same as the baseline case. However, half of its variability is contributed by a random effect at the group level.

[16] For example, both the *rdrobust* and *RDD* packages used in this chapter offer the option to specify a clustering variable.

Simulation results for $G = 5, 10, 25$ are reported in Table 2.3.[17] Again, two other methods besides the wild bootstrap method are estimated.[18] All the three methods fail to give a good interval estimate, which are well below the nominal level. This is not surprising because interval estimates from the two analytical methods (the CCT's robust approach and the conventional method) are based on large $G$ asymptotics. The wild bootstrap approach consistently performs better than the conventional method, but does not improve much from the CCT's robust approach. The wild bootstrap procedure proposed in this chapter is similar to the "wild bootstrap-se" method considered by Cameron et al. (2008). Their simulation results show that "wild bootstrap-se" method still suffers from size distortion with small number of clusters and is inferior to "wild bootstrap-t" method. The "wild bootstrap-t" method works well because (1) it imposes the null hypothesis so that estimation is more precise and (2) it bootstraps asymptotically pivotal t-statistics and achieves refinement.

This simple experiment shows that the wild bootstrap procedure can not only give valid confidence interval with independent data, it can also be easily applied to clustered data with slight adjustment to its resampling algorithm and performs at least as good as the analytical robust method.

## 2.6 Application

In this section, I apply the bootstrap procedure to the data used in Angrist and Lavy (1999).[19] In their paper, the effects of class size on scholastic achievement are estimated using the Maimonides' rule as instrument.

The rule that maximum class size is 40 has been adopted by Israeli public schools to determine the division of enrollment cohorts into classes since 1969. Following this rule, when the enrollment

[17] Since the RD designs is estimated separately on each side, $G$ means the number of clusters on each side. It is assumed there is no clusters crossing the cutoff.

[18] For the conventional method, I use the MSE-optimal bandwidth selector ignoring the fact that data is actually clustered. The conventional method uses cluster-robust variance estimator to construct confidence interval. For the CCT's robust approach, I used their companion R package *rdrobust*, which accommodates clustered data in both bandwidth selection and interval construction.

[19] The data is available at http://economics.mit.edu/faculty/angrist/data1/data/anglavy99.

Table 2.3: Empirical coverage and average interval length (clustered data)

| DGP | Method | Bias | SD | RMSE | EC(%) | IL | $h_{CCT}$ | $b_{CCT}$ | $h_{IK}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Panel A: $G = 5$ | | | | |
| 1 | Wild bootstrap | 0.018 | 0.081 | 0.083 | 87.0 | 0.268 | 0.251 | 0.318 | |
| | CCT robust | 0.018 | 0.081 | 0.083 | 86.8 | 0.274 | 0.251 | 0.318 | |
| | Conventional | 0.043 | 0.071 | 0.083 | 83.7 | 0.249 | | | 0.392 |
| 2 | Wild bootstrap | 0.037 | 0.085 | 0.093 | 83.4 | 0.274 | 0.165 | 0.297 | |
| | CCT robust | 0.039 | 0.086 | 0.094 | 84.0 | 0.289 | 0.165 | 0.297 | |
| | Conventional | 0.214 | 0.101 | 0.237 | 22.5 | 0.275 | | | 0.216 |
| 3 | Wild bootstrap | 0.007 | 0.080 | 0.080 | 88.6 | 0.270 | 0.200 | 0.312 | |
| | CCT robust | 0.007 | 0.080 | 0.081 | 89.0 | 0.276 | 0.200 | 0.312 | |
| | Conventional | -0.023 | 0.076 | 0.080 | 87.5 | 0.261 | | | 0.202 |
| | | | | | Panel B: $G = 10$ | | | | |
| 1 | Wild bootstrap | 0.017 | 0.068 | 0.070 | 90.2 | 0.240 | 0.230 | 0.321 | |
| | CCT robust | 0.018 | 0.068 | 0.071 | 88.9 | 0.236 | 0.230 | 0.321 | |
| | Conventional | 0.043 | 0.055 | 0.070 | 83.8 | 0.200 | | | 0.396 |
| 2 | Wild bootstrap | 0.036 | 0.071 | 0.079 | 87.1 | 0.250 | 0.166 | 0.299 | |
| | CCT robust | 0.038 | 0.071 | 0.081 | 86.2 | 0.253 | 0.166 | 0.299 | |
| | Conventional | 0.213 | 0.089 | 0.231 | 12.9 | 0.239 | | | 0.216 |
| 3 | Wild bootstrap | 0.005 | 0.067 | 0.067 | 92.7 | 0.243 | 0.186 | 0.316 | |
| | CCT robust | 0.005 | 0.068 | 0.068 | 91.5 | 0.240 | 0.186 | 0.316 | |
| | Conventional | -0.025 | 0.062 | 0.067 | 88.0 | 0.220 | | | 0.204 |
| | | | | | Panel C: $G = 25$ | | | | |
| 1 | Wild bootstrap | 0.016 | 0.061 | 0.063 | 91.7 | 0.216 | 0.213 | 0.323 | |
| | CCT robust | 0.016 | 0.061 | 0.063 | 89.6 | 0.210 | 0.213 | 0.323 | |
| | Conventional | 0.043 | 0.043 | 0.060 | 78.7 | 0.157 | | | 0.399 |
| 2 | Wild bootstrap | 0.038 | 0.065 | 0.075 | 86.8 | 0.228 | 0.165 | 0.300 | |
| | CCT robust | 0.040 | 0.066 | 0.077 | 86.6 | 0.230 | 0.165 | 0.300 | |
| | Conventional | 0.214 | 0.084 | 0.230 | 6.4 | 0.210 | | | 0.216 |
| 3 | Wild bootstrap | 0.004 | 0.060 | 0.060 | 94.1 | 0.221 | 0.174 | 0.317 | |
| | CCT robust | 0.004 | 0.060 | 0.061 | 92.6 | 0.216 | 0.174 | 0.317 | |
| | Conventional | -0.025 | 0.053 | 0.059 | 86.6 | 0.186 | | | 0.205 |

Note: EC denotes empirical coverage and IL denote average interval length based on 5000 simulations; nominal coverage probabilities are 95% for each estimator. The columns $h_{CCT}$ and $b_{CCT}$ list average optimal bandwidths following CCT's method.The column $h_{IK}$ lists average optimal bandwidth minimizing MSE.

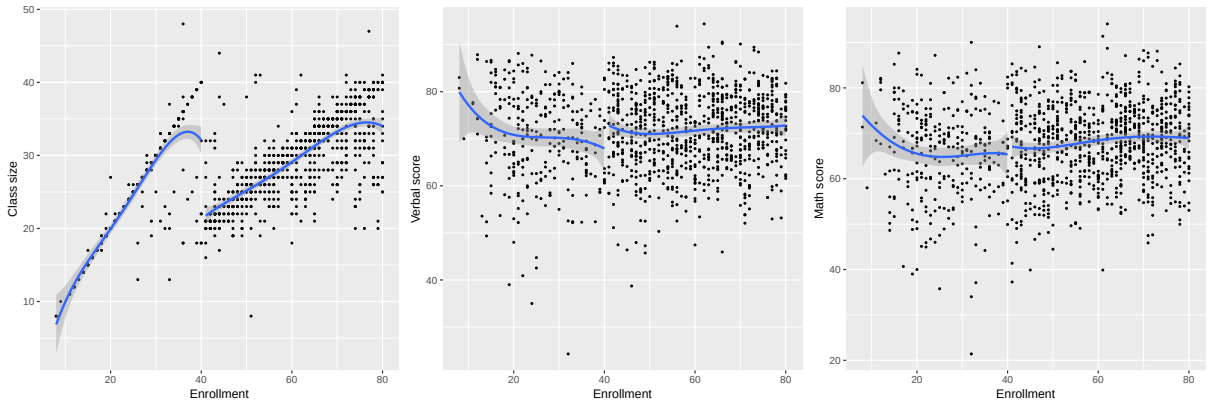



Figure 2.1: Class size, average verbal and math scores

increases and passes multiples of 40, an additional class is required. Since the total enrollment is roughly evenly divided into all classes, an additional class causes a sudden drop in class size. Ideally, when the enrollment grows from 40 to 41, class size will drop by half. Because of student turnover and imperfect enforcement of this rule, the empirical data fit into a fuzzy RD design.

The first discontinuity in class size for the 4th grade is considered. The sample used in this application includes 1164 classes from schools with enrollments no larger than 80. The outcome variables are average verbal and math test scores at class level. The discontinuities in class size and outcomes against enrollment are visualized in Figure 2.1. Each dot in these plots represents a class and the regression lines are fitted by fourth order polynomials. The shaded areas indicate confidence interval. The first plot clearly shows the discontinuity in class size. The second plot suggests a discontinuity in average verbal score, but not as significant as that in class size. The last plot does not provide much evidence for a discontinuity in average math score.

Similar to the simulations in Section 2.4, three methods are applied to estimate the effect of class size on average verbal/math scores and results are shown in Table 2.4. The first column lists the original point estimates from local linear regression, which depends only on the bandwidth choice. This explains why estimates from wild bootstrap and CCT's approach are identical and they are close to the conventional estimate. The second column lists the bias-corrected point estimates based on bootstrap bias correction and analytical bias correction. They are very close to each other but

Table 2.4: The effect of class size on average verbal score and average math score.

| | ATE | | 95% CI | | $h_{CCT}$ | $b_{CCT}$ | $h_{IK}$ |
|---|---|---|---|---|---|---|---|
| | Original | Corrected | | | | | |
| | Panel A: Average verbal score | | | | | | |
| Wild bootstrap | -0.449 | -0.575 | (-1.100 | 0.131 ) | 12.391 | 18.278 | |
| CCT robust | -0.449 | -0.575 | (-1.111 | -0.040) | 12.391 | 18.278 | |
| Conventional | -0.488 | | (-1.104 | 0.129 ) | | | 7.952 |
| | Panel B: Average math score | | | | | | |
| Wild bootstrap | -0.185 | -0.263 | (-0.924 | 0.466) | 11.612 | 17.683 | |
| CCT robust | -0.185 | -0.272 | (-0.884 | 0.340) | 11.612 | 17.683 | |
| Conventional | -0.202 | | (-0.802 | 0.398) | | | 9.200 |

differ a lot from the original estimates (the magnitude increases from 0.449 to 0.575 for average verbal score and 0.185 to 0.263∼0.272 for average math score).

Consistent with what Figure 2.1 shows, only one out of three intervals for the treatment effect on average verbal score excludes zero and all three intervals for the treatment on average match score include zero. The interval from wild bootstrap is wider than that from robust analytical approach, suggesting that it is more conservative, which also can be found from previous simulation studies.

## 2.7 Conclusion

A new wild bootstrap procedure is proposed to correct bias and construct valid confidence interval in fuzzy RD designs. This new method builds upon the developments and intuition advanced by CCT but is implemented through a novel iterated bootstrap. In particular, the local second order models are estimated for generating bootstrap samples. The first layer of bootstrap is performed in order to obtain the empirical distribution of bias-corrected treatment effect, which is made possible by utilizing a second layer of bootstrap to estimate the bias from linear models. This new procedure is proved to be theoretically valid and empirically supported by simulation studies.

# CHAPTER 3. BOUNDING TREATMENT EFFECTS WITH MISCLASSIFIED DISCRETE DATA

A novel numerical approach is proposed to partially identify treatment effects. Endogenous treatment and measurement error are very common in survey data and pose threats to reliable estimation of treatment effects. The new approach considers these two issues simultaneously and provides bounds for treatment effects. Conceptually, treatment effects and model assumptions are formulated as linear restrictions on a large set of probability mass. One can then check if any given treatment effect is consistent with model assumptions and observed data. Compared with previous methods, the newly proposed numerical approach is general enough to be applied to various different problems and guarantees sharp bounds. An example is provided to show that how the distribution of a treatment effect and how the averages of multiple treatment effects can be partially identified through this approach.

## 3.1 Introduction

To estimate average treatment effects (ATE) is of great importance to policy makers. Without the unconfoundedness assumption, economists have invested many efforts in developing econometric and statistical models to reveal causal effects from observational data (see Imbens and Wooldridge (2009) for a review of these models). However, the endogenous treatment is not the only obstacle to causal inference because the data sets used by economists are not error free (see for example Bound et al. (2001)). While there are a few studies looking into both issues simultaneously (see for example Kreider et al. (2012)), their methods are usually customized to the specific questions being answered. This paper introduces a general framework to identify the region of treatment effects in the presence of both endogenous treatment and measurement error in discrete data.

This new approach is conceptually straightforward and extremely flexible to accommodate various assumptions on the selection process and the misclassification process.

This paper adopts and generalizes the approach from Balke and Pearl (1994, 1997) and Lafférs (2013), who used linear programing to bound average treatment effects when both treatment and outcome are binary. The basic idea is to find all possible joint distributions of the response function and observables. Such a joint distribution is constrained in two ways: (1) it should respect the prior information such at the assumption on selection process and (2) it must be in line with the distribution of observables. Once a set of feasible joint distributions is identified, we are able to find a set of feasible treatment effects because the marginal distribution of response function determines the treatment effect. This approach is generalized in two aspects in this paper. First, multiple treatments and multiple outcomes are allowed. The analysis of multiple treatments is an underexplored yet important topic because multiple programs participation is common and the interaction among multiple programs is usually not well known. Allowing for multiple outcomes accommodates more data types, e.g., ordered variables, which are widely used in survey designs. Second, instead of obtaining an interval for the average treatment effect, this paper proposes set identifying a region of multiple dimensions, thus allowing many interesting questions to be answered. For example, this region could denote the distribution of the treatment effect, from which one can infer the quantile treatment effect. This region could also denote average treatment effects of multiple treatments so that the interaction among programs can be explored.

One critical component which differentiates this paper from previous work by Balke and Pearl (1994, 1997) and Lafférs (2013) is the consideration of misclassification. While they assume that the distribution of observables is consistently revealed by the observational data, this paper acknowledges the possibility that the observed distribution systematically differs from the true distribution. For example, Meyer et al. (2009) found high rates of understatement in the participation of government transfer programs in Survey of Income and Program Participation (SIPP), Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID). Feng (2013) found substantial underreport of unemployment in CPS, based on which the official unemployment rate is calculated.

Their revised unemployment rate is higher than the official rate by 2.1% on average. This paper models the misclassification process by firstly allowing any kind of errors and then gradually imposing restriction of these errors based on prior information. Essentially any assumptions on the misclassification process can be easily incorporated into this framework. As a result, there is no need to develop specific models to deal with different patterns of misclassification.

This paper contributes to the literature of program evaluation by combining two related areas: partial identification and correction for measurement error. Developed by Manski (1990, 1997, 2003), bounds analysis has gained popularity in program evaluation (see for example Ginther (2000); Gonzalez (2005); Gerfin and Schellhorn (2006); Gundersen and Kreider (2009); Kreider and Hill (2009); De Haan (2011); Gundersen et al. (2012); Kreider et al. (2012)). In their work, analytical bounds for treatment effects are derived following assumptions which are not strong enough to point identify parameters of interest. However, sharp analytical bounds, i.e., bounds which exhaust all available information, are not always easy, if not impossible, to derive because different assumptions interact with each other. For example, Manski and Pepper (2000) pointed out that it is complex to analyze the bounds of returns to schooling when the assumption of monotone treatment response (MTS) is maintained and monotone instrumental variables (MIV) are applied. More over, researchers usually have to derive the analytical bounds case by case for different combination of assumptions. The difficulty and hassle of deriving analytical bounds can be circumvented by searching algorithms which numerically minimize objective functions given a series of carefully tailored constraints. With cheaper computational power and more efficient algorithms, to perform optimization numerically becomes handy in economic studies (see for example Balke and Pearl (1997); Manski and Tamer (2002); Honoré and Tamer (2006); Molinari (2008); Ekeland et al. (2010); Lafférs (2013)).

Following this numerical optimization approach, this paper expands the searching space from the joint distribution of the response function and observables to an additive misclassification matrix. As a result, the selection process and misclassification process are considered simultaneously in a unified framework. The newly introduced additive misclassification approach is closely related to the

work by Molinari (2008), who developed the direct misclassification approach. The main motivation to modify Molinari's (2008) approach is computational tractability. The direct misclassification approach introduces a set of conditional probabilities to capture the misclassification rates. While it is easy to formulate optimization problems following this approach, it is not very computationally friendly. The main reason is that the searching space involves many non-linear constraints and is non-convex. Molinari (2008) also pointed out that her approach works best if the dimension is small, e.g., workers union status, employment status, health conditions, and health/functional status. However, to model the misclassification process at the lowest level, one needs to consider all combinations of observables, resulting in large dimensions. For example, in the case of binary outcome, binary treatment and a discrete covariate with a support of 20 values (the case that Kreider et al. (2012) considered), the dimension goes up to 80.[1] The additive misclassification approach introduced in this paper is an alternative to Molinari's (2008) approach and greatly reduces the computation intensity under certain circumstances. It will be shown that under some widely used assumptions regarding to the misclassification process, e.g., corrupted sampling and contaminated sampling (Horowitz and Manski, 1995), the optimization problem reduces to a well-understood linear programing problem.

## 3.2 The Additive Misclassification Approach

In this section, the additive misclassification approach is introduced and compared with the direct misclassification approach by Molinari (2008). To start with, let $W$ be a random discrete variable of interest, which has support $S_W = \{w_1, w_2, \ldots, w_{N_W}\}$ and distribution $P_W$. Since $W$ cannot be perfectly measured, use $W'$ to denote the observed $W$ and $P'_W$ to denote the observed distribution. Let $\mathbf{P}^W = (P_W(w_1), P_W(w_2), \ldots, P_W(w_{N_W}))^T$ and $\mathbf{P}^{W'} = (P'_W(w_1), P'_W(w_2), \ldots, P'_W(w_{N_W}))^T$. Molinari (2008) introduced a direct misclassification approach to infer $P_W$ based on $P'_W$. Her idea is that $\mathbf{P}^{W'}$ can be written as a function of $\mathbf{P}^W$ through a series of linear equations, which is captured

[1] There are 80 different vectors of the observables (outcome, treatment and covariate). Without any constraints, measurement error can happen between any two vectors, i.e., $80 \times 79 = 6320$ different types of misclassification. If it is assumed that covariates are error-free, this number reduces to $4 \times 3 \times 20 = 240$.

by a multiplicative misclassification matrix $\Pi$:

$$\mathbf{P}^{W'} = \Pi \mathbf{P}^{W},$$

where

$$\Pi = \begin{pmatrix} \mathbb{P}[W' = w_1 | W = w_1] & \mathbb{P}[W' = w_1 | W = w_2] & \cdots & \mathbb{P}[W' = w_1 | W = w_{N_W}] \\ \mathbb{P}[W' = w_2 | W = w_1] & \mathbb{P}[W' = w_2 | W = w_2] & \cdots & \mathbb{P}[W' = w_2 | W = w_{N_W}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}[W' = w_{N_W} | W = w_1] & \mathbb{P}[W' = w_{N_W} | W = w_2] & \cdots & \mathbb{P}[W' = w_{N_W} | W = w_{N_W}] \end{pmatrix}.$$

This approach is very convenient, however as will be seen later, it is sometimes not the best way to model misclassification because of the nonlinearity introduced by the multiplication. The additive misclassification approach does not use the conditional probability:

$$E = \begin{pmatrix} 0 & \mathbb{P}[W' = w_1, W = w_2] & \cdots & \mathbb{P}[W' = w_1, W = w_{N_W}] \\ \mathbb{P}[W' = w_2, W = w_1] & 0 & \cdots & \mathbb{P}[W' = w_2, W = w_{N_W}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}[W' = w_{N_W}, W = w_1] & \mathbb{P}[W' = w_{N_W}, W = w_2] & \cdots & 0 \end{pmatrix}. \quad (3.1)$$

The additive misclassification matrix $E$ is obtained from the direct misclassification matrix $\Pi$ by two steps: firstly multiply the $i$th column of $\Pi$ by $P_W(w_i)$ and secondly replace diagonal elements by zeros (because diagonal cells represent correct measurement). Let $\mathbf{1}_{N_W}$ be a vector of ones, with $N_W$ being the size of set $S_W$, then $\mathbf{P}^{W'}$ can be expressed by $\mathbf{P}^{W}$ and $E$ by the following equation:

$$\mathbf{P}^{W'} = \mathbf{P}^{W} + E\mathbf{1}_{N_W} - E^T\mathbf{1}_{N_W}, \quad (3.2)$$

where $E\mathbf{1}_{N_W}$ are false positive probabilities and $E^T\mathbf{1}_{N_W}$ are false negative probabilities.

Let $\mathbf{P}^E = \text{vec}(E)$. The set $H^p[\mathbf{P}^E]$ is a collection of all feasible $\mathbf{P}^E$, with the superscript $p$ indicating probabilistic requirement. Formally,

$$H^p[\mathbf{P}^E] = \{\mathbf{P}^E \colon \mathbf{P}^E \in [0,1]^{N_W^2}, \mathbf{P}^W \in [0,1]^{N_W}, \quad (3.3a)$$

$$E_{ii} = 0 \ \forall \ i = 1, 2, \ldots, N_W, \quad (3.3b)$$

$$\mathbf{P}^W - E^T\mathbf{1}_{N_W} \succeq 0, \quad (3.3c)$$

$$\mathbf{P}^{W'} = \mathbf{P}^W + E\mathbf{1}_{N_W} - E^T\mathbf{1}_{N_W}\}. \quad (3.3d)$$

Condition (3.3a) is the basic requirement because $\mathbf{P}^W$ and $\mathbf{P}^E$ are probabilities. Condition (3.3b) corresponds to the definition of $E$ that all diagonal elements are zero. Condition (3.3c) imposes a cap in the misclassification probability, i.e., $\mathbb{P}[W = w_i] \geq \mathbb{P}[W' \neq w_i, W = w_i]$. Condition (3.3d) is a reproduce of equation (3.2). The constraint that all probabilities in $\mathbf{P}^W$ should add up to one is implied by the last condition and thus ignored.

**Proposition 3.1.** *$H^p[\mathbf{P}^E]$ is a compact and convex set.*

This conclusion deviates from Proposition 1 by Molinari (2008), who shows that the set of feasible $\Pi$ is connected but not convex. This deviation can be explained by the different construction of misclassification matrices $\Pi$ and $E$. While the direct misclassification approach unavoidably involves nonlinear constraints on $\Pi$ and $\mathbf{P}^W$, the additive misclassification approach only involves linear constraints on $E$ and $\mathbf{P}^W$.

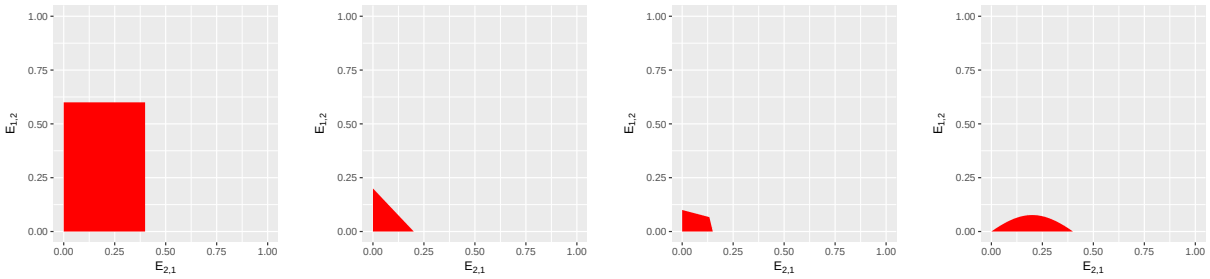
Not only $H^p[\mathbf{P}^E]$ is convex, when combined with some widely used assumptions on the misclassification process, the set of $\mathbf{P}^E$ is still convex. In Example 3.1, it is shown that the geometry of $\mathbf{P}^E$ is convex given no prior information on the misclassification. The convexity of $\mathbf{P}^E$ also holds under some widely used assumptions such as maximum misclassification rate.

**Example 3.1.** *Suppose $W$ is a binary variable and $P_{W'}(1) = 0.6$. The additive misclassification matrix is fully captured by its off-diagonal elements $E_{2,1}$ and $E_{1,2}$. Without any assumption on the misclassification, the geometry of $\mathbf{P}^E$ is plotted in Figure 3.2(a). Define misclassification rates:*

$$\pi = \mathbb{P}[W' \neq W], \quad \pi_{2,1} = \mathbb{P}[W' = 1|W = 0], \quad \pi_{1,2} = \mathbb{P}[W' = 0|W = 1].$$

*Figure 3.2(b), 3.2(c) and 3.2(d) present the geometry of $\mathbf{P}^E$ under different assumptions. In Figure 3.2(b), the overall misclassification rate is capped by 0.2. In Figure 3.2(c), the misclassification rates for both $W = 0$ and $W = 1$ are capped by 0.2. In Figure 3.2(d), the monotonicity in misclassification rate is imposed. In all cases, the geometry of $\mathbf{P}^E$ is convex.*[2]

[2] *While the convexity of $\mathbf{P}^E$ in Figure 3.2(b) and 3.2(c) holds for $W$ with any dimension, the convexity of $\mathbf{P}^E$ under the assumption of monotonicity in misclassification rate holds only when $W$ has a dimension of two.*




(a) No assumption (b) $\pi \leq 0.2$ (c) $\pi_{2,1}, \pi_{1,2} \leq 0.2$ (d) $\pi_{2,1} \leq \pi_{1,2}$

Figure 3.2: The geometry of $\mathbf{P}^E$

By construction of the additive misclassification matrix, $\mathbf{P}^W$ is a linear transformation of $\mathbf{P}^E$. As a result, any parameter of interest as a function of $\mathbf{P}^W$ is also a function of $\mathbf{P}^E$, i.e., $\tau = G(\mathbf{P}^W) = G'(\mathbf{P}^E)$. The convexity of $\mathbf{P}^E$ is a very desirable property in making inference on $\tau$ because convex optimization is in general much easier than non-convex optimization. While there are lots of standard algorithms for efficiently solving convex optimization problems, non-convex optimization problems are hard (if not impossible) to solve exactly in a reasonable time. As a result, heuristic algorithms, which does not guarantee desired solutions, are usually used in practice.

## 3.3 The Treatment Effect with Discrete Data

In this section, the linear programming model for partially identifying treatment effect (Lafférs, 2013) is utilized and seamlessly combined with the newly introduced additive misclassification approach. The resulted model is capable of taking any prior information on the selection process and misclassification process.

Given finite sets $S_Y$ and $S_Z$, let random variable $Y \in S_Y$ denote the outcome and $Z \in S_Z$ the treatment. Both outcome and treatment are likely to be measured with error. Let $Y'$ denote the observed outcome and $Z'$ the observed treatment, which may be different from their true values. Define the response function $F \in S_F = \{f : Z \to Y\}$ and the treatment effect $T = F(z_j) - F(z_i) \in S_T$. The distribution of $T$ is usually of interest. Given any specific response function $f$, the treatment effect $f(z_j) - f(z_i)$ can be calculated. As a result, in order to identify the distribution of

$T$, it suffices to identify the distribution of $F$. Let $P$ with subscript denote distribution, $\mathbb{E}$ denote expectation and $\mathbb{P}$ denote probability. Then $P_T$ and $\mathbb{E}[T]$ are functions of $P_F$:

$$P_T(t) = \sum_{\substack{f \in S_F \\ f(z_j) - f(z_i) = t}} P_F(f). \tag{3.4}$$

The distribution of treatment effect is captured by the vector $\mathbf{P}^T = (P_T(t_1), P_T(t_2), \ldots, P_T(t_{N_T}))^T$, with $N_T$ being the size of set $S_T$. Equation (3.4) indicates that $\mathbf{P}^T$ is linear in probability mass from distribution $P_F$. However, $P_F$ is never known because only one point of the response function is revealed for each unit. The difficulties in identifying treatment effects from observational data can be illustrated in Figure 3.3.

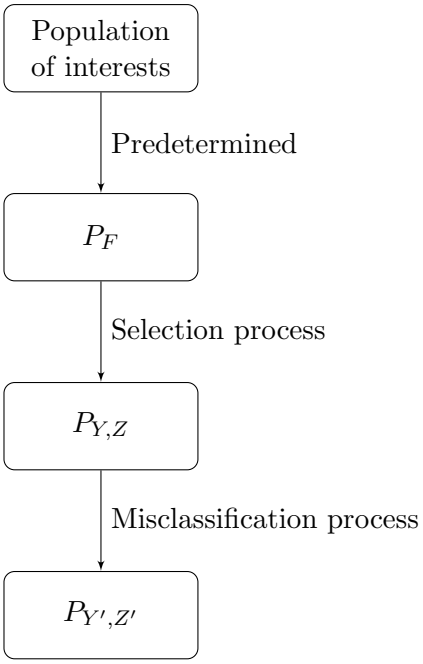



Figure 3.3: Identify treatment effects from misclassified data

The distribution of response function $F$ is predetermined and likely to affect which treatment to receive. The distribution $P_{Y,Z}$ is linked to $P_F$ by

$$\begin{aligned} P_{Y,Z}(y,z) &= \sum_{f \in S_F} P_{Y,Z,F}(y,z,f) \\ &= \sum_{f \in S_F} P_F(f) P_{Z|F}(z|f) P_{Y|Z,F}(y|z,f) \\ &= \sum_{f \in S_F} \mathbb{1}[f(z) = y] P_F(f) P_{Z|F}(z|f) \end{aligned} \tag{3.5}$$

The first equality holds because of the law of total probability. The second equality holds after repeatedly applying Bayes's rule and the third equality holds because there is no uncertainty in outcome once $Z$ and $F$ are given. Equation (3.5) highlights the data generating process of observables. Notice that $P_{Z|F}$ captures the selection process. Whenever $P_{Z|F} \neq P_Z$, endogeneity issue arises because the distribution of treatment depends on response function, which captures the unobserved heterogeneity among different units. For example, one may choose a treatment based on expected gain.

**Proposition 3.2.** *The expected outcome conditional on treatment* $\mathbb{E}[Y|Z=z]$ *is a weighted average of* $Y(z)$*, with the weights being proportional to* $\frac{P_{Z|F}(z|f)}{P_Z(z)}$.

Proposition 3.2 has two implications. First, without the assumption that $P_{Z|F}(z|f) = P_Z(z)$, the conditional expectation $\mathbb{E}[Y|Z=z]$ does not equal to the unconditional expectation $Y(z)$. As a result, one can evaluate treatment effects by taking the difference of conditional means only if the treatment is random. Second, when the selection process $P_{Z|F}(z|f)$ is known and the conditional mean $\mathbb{E}[Y|Z=z]$ can be interpreted as a weighted average of $Y(z)$. However the difference between conditional means cannot be interpreted as a weighted treatment effect because the weight $\frac{P_{Z|F}(z|f)}{P_Z(z)}$ varies with treatment in general.

The misclassification process is modeled by the additive misclassification approach similar to equations (3.1) and (3.2), except that $W$ is now a random vector $W = (Y, Z)$. Accordingly, $w, w'$ are specific vectors from $S_Y \times S_Z$.

To identify the treatment effect, let $\mathbf{P}^{FW} = \text{vec}(\boldsymbol{M}_{F,W})^T$ and $\mathbf{P}^{FWE} = [(\mathbf{P}^{FW})^T, (\mathbf{P}^{E})^T]^T$, where $\boldsymbol{M}_{F,W}$ is the matrix of all probability mass from distribution $P_{F,W}$ defined below:

$$\boldsymbol{M}_{F,W} = \begin{pmatrix} P_{F,W}(f_1, w_1) & P_{F,W}(f_1, w_2) & \dots & P_{F,W}(f_1, w_{N_W}) \\ P_{F,W}(f_2, w_1) & P_{F,W}(f_2, w_2) & \dots & P_{F,W}(f_2, w_{N_W}) \\ \vdots & \vdots & \ddots & \vdots \\ P_{F,W}(f_{N_F}, w_1) & P_{F,W}(f_{N_F}, w_2) & \dots & P_{F,W}(f_{N_F}, w_{N_W}) \end{pmatrix}.$$

The vector $\mathbf{P}^{FWE}$ must meet a minimum set of probabilistic requirement in a similar way to equation (3.3). Formally,

$$H^p[\mathbf{P}^{FWE}] = \{\mathbf{P}^{FWE} \colon \mathbf{P}^{FWE} \in [0,1]^{N_W(N_W+N_F)}, \tag{3.6a}$$

$$E_{ii} = 0 \ \ \forall \ \ i = 1, 2, \dots, N_W, \tag{3.6b}$$

$$P_{F,Y,Z}(f, y, z) = 0 \ \text{ if } \ f(z) \neq y, \tag{3.6c}$$

$$\boldsymbol{M}_{F,W}^T \mathbf{1}_{N_F} - E^T \mathbf{1}_{N_W} \succeq 0, \tag{3.6d}$$

$$\mathbf{P}^{W'} = \boldsymbol{M}_{F,W}^T \mathbf{1}_{N_F} + E \mathbf{1}_{N_W} - E^T \mathbf{1}_{N_W}\}. \tag{3.6e}$$

Conditions (3.6a), (3.6b), (3.6d) and (3.6e) are analogous to conditions (3.3a), (3.3b), (3.3c) and (3.3d) respectively. Condition (3.6c) assigns zero probabilities to some response functions because they are in conflict with observed treatment and outcome. For example, in the case of binary treatment and outcome, if response function $f_1$ is defined by $f_1(0) = 1$ and $f_1(1) = 1$, then it is known for certain that $\mathbb{P}[F = f_1, Y = 0, X = 0] = 0$ and $\mathbb{P}[F = f_1, Y = 0, X = 1] = 0$. Besides the minimum probabilistic requirement for $\mathbf{P}^{FWE}$, any prior information can impose restrictions on the values that $\mathbf{P}^{FWE}$ can take. Let $H^i[\mathbf{P}^{FWE}]$ denote the set of $\mathbf{P}^{FWE}$ which respects prior information. Then the distribution of treatment effect $\mathbf{P}^T$ is identified by

$$H[\mathbf{P}^T] = \{\boldsymbol{B}\mathbf{P}^{FWE} \colon \mathbf{P}^{FWE} \in H^p[\mathbf{P}^{FWE}] \cap H^i[\mathbf{P}^{FWE}]\}, \tag{3.7}$$

where $\boldsymbol{B}$ is a known $N_T$ by $N_W(N_W + N_F)$ matrix implicitly defined in equation (3.4).

**Proposition 3.3.** *If $H^i[\mathbf{P}^{FWE}]$ is convex, $H[\mathbf{P}^T]$ is convex.*

Though equation (3.7) provides a straightforward expression of $H[\mathbf{P}^T]$, to empirically find the geometry of $H[\mathbf{P}^T]$ is not trivial. Without loss of generality, consider the following form of $H^p[\mathbf{P}^{FWE}] \cap H^i[\mathbf{P}^{FWE}]$:[3]

$$\{\mathbf{P}^{FWE} \in \mathbb{R}^{N_W(N_W+N_F)} \colon g_i(\mathbf{P}^{FWE}) \geq 0, \;\; i = 1, 2, \dots k\}.$$

Without knowing $H[\mathbf{P}^T]$ is convex, one generally needs to check every possible vector $\zeta \in \mathbb{R}^{N_T}$ to find the region of $H[\mathbf{P}^T]$. This can be done by solving the following optimization problem:

$$\begin{aligned} Q(\zeta) = \min_{\mathbf{P}^{FWE}, \{v_i\}} & \sum_{i=1}^{k} v_i \\ \text{s.t. } & \boldsymbol{B}\mathbf{P}^{FWE} - \zeta = 0, \\ & v_i \geq 0 \;\; \forall \;\; i = 1, 2, \dots, k, \\ & g_i(\mathbf{P}^{FWE}) + v_i \geq 0 \;\; \forall \;\; i = 1, 2, \dots, k \end{aligned}$$

When $Q(\zeta) = 0$, i.e., $v_1, v_2, \dots, v_k = 0$, one can conclude that $\zeta \in H[\mathbf{P}^T]$ because all the $k$ constraints are satisfied, otherwise $\zeta \notin H[\mathbf{P}^T]$. In the case $H^i[\mathbf{P}^{FWE}]$ is convex, which will be intensively discussed in the following section, the distribution of treatment effect $H[\mathbf{P}^T]$ is also convex by Proposition 3.3. It is not necessary to check every possible vector $\zeta \in \mathbb{R}^{N_T}$. Instead, one can firstly find the range of the first dimension, secondly find the range of the second dimension conditional on the first dimension, thirdly find the range of the third dimension conditional on the first two dimensions and repeat this process until the last dimension.

## 3.4 Analysis of the Identifying Power of Specific Restrictions

Several specific restrictions are discussed in this section. In line with the previous section, these restrictions are grouped into three categories: restrictions on the predetermined distribution of

[3] Constraints with strict inequality are not considered to save the discussion of potential open set. All constraints in the form of "$\geq$", "$=$" and "$\leq$" are formatted as constraints in the form of "$\geq$" for notational simplicity.

response function $P_F$, restrictions on the selection process $P_{Z|F}$ and restrictions on the misclassification process $E$. A numerical example illustrating the identifying power of these restrictions is provided at the end of this section.

### 3.4.1 The response function

The distribution of response function $P_F$ is predetermined but unobservable. However, among the $N_Y^{N_Z}$ different response functions, some are less credible than others. Manski (1997) discussed what can be learned from the data without knowledge on treatment selection. His key assumption is monotone treatment response (MTR), i.e., $z_i \geq z_j \Rightarrow f(z_i) \geq f(z_j)$. This assumption is valid if the treatment is widely accepted to have non-positive or non-negative effect, e.g., the effect of language skills on wage (Gonzalez, 2005). To impose this assumption is equivalent to assign zero probabilities to some response functions which violate monotonicity:

$$P_F(f) = 0 \text{ if } z_i \geq z_j \Rightarrow f(z_i) < f(z_j). \tag{3.8}$$

Assumption (3.8) indicates that every individual response function is monotone in treatment, which may be too strong since no exception is allowed. To weaken MTR at individual level, one can assume that the monotonicity holds on average, hence MTR on average:

$$\mathbb{E}[F(z_i) - F(z_j)] \geq 0 \text{ if } z_i \geq z_j. \tag{3.9}$$

The assumption of MTR on average does not directly assign any zero probabilities, thus all possible response functions are allowed. Besides restriction on the average treatment effect, one may also have some beliefs on the quantile treatment effect, which can be formatted in a similar way. For example, if $z_i \geq z_j$, at least $100\alpha$ percent of units benefit from switching from $z_j$ to $z_i$:

$$Q_\alpha[F(z_i) - F(z_j)] \geq 0 \text{ if } z_i \geq z_j. \tag{3.10}$$

The assumption of quantile treatment effect may be relevant when a program is made available by voting, so at least a certain share of units can benefit from participation. Assumptions (3.9) and (3.10) are not widely used in empirical studies, probably because it is difficult to derive analytical

solutions. Not surprisingly, if Assumption (3.8) is imposed, then (3.9) holds conditional on any covariates (see for example Manski and Pepper (2009)).

### 3.4.2 The selection process

The distribution of treatment conditional on response function captures the treatment selection process. A rational unit will make a choice which maximizes her utility. Once the outcome and treatment are observed, this outcome should be the best outcome that she can obtain by choosing from all available treatments. Manski (1990) studied the identification power of selection for better outcome. In his settings, there are treatment $A$ and $B$, when the realized treatment is $B$, one can infer that $Y(B) \geq Y(A)$. Following this idea, the assumption of selection for better outcome can be generalized to multiple treatments:

$$P_{F,Y,Z}(f, y, z) = 0 \text{ if } \exists \tilde{z} \in S_Z, f(\tilde{z}) > y. \tag{3.11}$$

In other words, the probability that $F = f, Y = y, Z = z$ is zero if there exists another treatment $\tilde{z} \neq z$ such that $f(\tilde{z}) > y$. Follow the large literature of bounded rationality, it is sometimes preferred to assume that units are seeking a satisfactory solution rather than the optimal one in decision making. As a result, a probabilistic version of Assumption (3.11) is proposed below:

$$\sum_{y \in S_Y} P_{F,Y,Z}(f, y, z_i) \geq \sum_{y \in S_Y} P_{F,Y,Z}(f, y, z_j) \text{ if } f(z_i) > f(z_j). \tag{3.12}$$

Assumption (3.12) is substantially weaker than Assumption (3.11). It states that, conditioning on response function $F = f$, units are more likely to select a treatment with better outcome.

### 3.4.3 The misclassification process

The misclassification process is usually modeled by a series of restrictions on misclassification rates, i.e., the conditional probability of measurement with error. In the direct misclassification approach, every off diagonal element in the misclassification matrix $\Pi$ represents a misclassification rate. However in the additive misclassification approach, the misclassification rates need to be

calculated. The overall misclassification rate is

$$\gamma_W = \mathbb{P}[W \neq W'] = \sum_{i=1}^{N_W} \sum_{j=1}^{N_W} E_{ij}. \tag{3.13}$$

For any specific value $w_j \in S_W$, its misclassification rate is

$$\gamma_{w_j} = \frac{\mathbb{P}[W = w_j, W' \neq w_j]}{\mathbb{P}[W = w_j]} = \frac{\sum_{i=1}^{N_W} E_{ij}}{\sum_{f \in S_F} P_{F,W}(f, w_j)}. \tag{3.14}$$

The models of corrupted sampling and contaminated sampling (Horowitz and Manski, 1995) can be easily implemented by utilizing equations (3.13) and (3.14). For example, to model corrupted sampling, one can impose $\gamma_W \leq \epsilon$, where $\epsilon$ is the cap of misclassification rate obtained from other sources. To model contaminated sampling, one can impose $\gamma_{w_j} \leq \epsilon \quad \forall \quad j = 1, 2, \ldots, N_W$. Both equation (3.13) and (3.14) focus on the misclassification of vector $W = (Y, X)$. In practice, it is more likely to have prior information on the misclassification rates for variables separately, but not jointly. It turns out the misclassification rates for a single variable overall and for any of its specific values can be obtained in a similar way. Take the treatment $Z$ as an example, the misclassification rate of $Z$ is

$$\gamma_Z = \mathbb{P}[Z \neq Z'] = \sum_{(i,j) \in S} E_{ij}, \tag{3.15}$$

$$\text{with} \quad S = \{(i, j) \in \{1, 2, \ldots, N_W\}^2 \colon z_i \neq z_j\}.$$

For any specific value $z_j \in S_Z$, its misclassification rate is

$$\gamma_{z_j} = \frac{\mathbb{P}[Z = z_j, Z' \neq z_j]}{\mathbb{P}[Z = z_j]} = \frac{\sum_{(a,b) \in S} E_{ab}}{\sum_{f \in S_F} \sum_{y \in S_Y} P_{F,Y,Z}(f, y, z_j)}, \tag{3.16}$$

$$\text{with} \quad S = \{(a, b) \in \{1, 2, \ldots, N_W\}^2 \colon z_a \neq z_j, z_b = z_j\}.$$

Equations (3.15) and (3.16) allow flexible and customized assumptions on the misclassification at variable level. In the application to program evaluation, it is usually the case that some variables are more prone to misclassification (such as treatment status), while some other covariates are

relatively credible (such as age). In this case, researchers may want to impose the assumption that misclassification rates of all variables except for treatment status are zero. More over, one can go further than equation (3.16) by restricting the probability of a specific error. For example, $\mathbb{P}[Z' = 1, Z = 0] = 0$ indicates that if a unit is not treated ($Z = 0$), she will not report being treated ($Z' = 1$). In other words, the probability of false positive is ruled out and only false negative is allowed. This is consistent with the fact that socially undesirable behavior is usually underreported (Meyer et al., 2009).

### 3.4.4 Discussion

While the distribution of treatment effect $H[\mathbf{P}^T]$ is identified given any prior information captured by $H[\mathbf{P}^{FWE}]$, the difficulty of empirically finding the geometry of $\mathbf{P}^T$ depends on how $H^i[\mathbf{P}^{FWE}]$ is constructed. It is straightforward to verify that Assumptions (3.8) - (3.12) imposes linear restrictions on the vector $\mathbf{P}^{FWE}$. The models of corrupted sampling and contaminated sampling utilizing equations (3.13) - (3.16) also impose linear restrictions on the vector $\mathbf{P}^{FWE}$. Since linear optimization is a special case of mathematical optimization and runs in polynomial time, it is easy to find the geometry of $\mathbf{P}^T$ even when the number of dimensions is large.

However, the optimization becomes substantially harder if some assumptions introduce non-linear constraints and lead to a non-convex set $H^i[\mathbf{P}^{FWE}]$. For example, this can happen when the assumption is formatted as comparison between two ratios, where both numerator and denominator involve vector $\mathbf{P}^{FWE}$. The assumption of monotonicity in correct reporting (see case (d) in Example 3.1) falls into this category. Another possibility is the application of mean independent or monotone instrumental variables when the instrumental variables are also misclassified. In both cases, the non-linearity arises because conditional probabilities are compared directly. If one is willing to assume that the variables being conditioned are error free, then the denominators in those ratios do not involve vector $\mathbf{P}^{FWE}$ and the constraints become linear in vector $\mathbf{P}^{FWE}$.

### 3.4.5 A numerical example

In this subsection, the identifying power of various assumptions on $P_F$, $P_{Z|F}$ and the misclassification matrix $E$ are examined. Consider $S_Y = \{0, 1\}$, $S_Z = \{0, 1, 2\}$ and $P_{Y',Z'}(y, z) = 1/6$. The treatment effect of $Z = 1$ is $T_1 = Y(1) - Y(0)$ and the treatment effect of $Z = 2$ is $T_2 = Y(2) - Y(0)$. Suppose one is interested in

- the distribution of $T_1$,
- the joint feasible set of $\mathbb{E}[T_1]$ and $\mathbb{E}[T_2]$.

Following (3.7), all possible distributions for $T_1$ constitute the following set:

$$\begin{aligned} H[\mathbf{P}^{T_1}] = \{(P_{T_1}(-1), P_{T_1}(0), P_{T_1}(1))^T : P_{T_1}(i) = & \sum_{\substack{f \in S_F \\ f(1)-f(0)=i}} P_F(f), \quad i = -1, 0, 1, \\ & P_{Y',Z'}(y, z) = 1/6 \ \ \forall y = 0, 1, \quad z = 0, 1, 2, \\ & \mathbf{P}^{FWE} \in H^p[\mathbf{P}^{FWE}] \cap H^i[\mathbf{P}^{FWE}]\}, \end{aligned}$$

where the first condition defines the probabilities of each possible treatment effect (in this example, the treatment effect can only be -1, 0 or 1) and the second and the third conditions makes sure that $\mathbf{P}^{FWE}$ is consistent with all prior information.

Analogous to $\mathbf{P}^{T_1}$, let $\mathbf{P}^{ATE} = (\mathbb{E}[T_1], \mathbb{E}[T_2])^T$ be a vector of two different average treatment effects. It is trivial to verify that $\mathbf{P}^{ATE} = \boldsymbol{B'}\mathbf{P}^{FWE}$, with $\boldsymbol{B'}$ being a $2 \times N_W(N_W + N_F)$ matrix. So equation (3.7) and Proposition 3.2 also apply to $\mathbf{P}^{ATE}$. All possible vectors of $\mathbf{P}^{ATE}$ constitute the following set:

$$\begin{aligned} H[\mathbf{P}^{ATE}] = \{(\mathbb{E}[T_1], \mathbb{E}[T_2])^T : \ \mathbb{E}[T_j] = & \sum_{i \in \{-1,0,1\}} i \sum_{\substack{f \in S_F \\ f(j)-f(0)=i}} P_F(f), \quad j = 1, 2, \\ & P_{Y',Z'}(y, z) = 1/6 \ \ \forall y = 0, 1, \quad z = 0, 1, 2, \\ & \mathbf{P}^{FWE} \in H^p[\mathbf{P}^{FWE}] \cap H^i[\mathbf{P}^{FWE}]\}. \end{aligned}$$

A large number of different $H^i[\mathbf{P}^{FWE}]$ are considered. Each $H^i[\mathbf{P}^{FWE}]$ is obtained by specifying the assumption on the distribution of response function $P_F$, the selection process $P_{Z|F}$ and the

misclassification process $E$. For $P_F$, three assumptions are considered: non-negative median treatment effect, MTR on averate and MTR at individual level.[4] See equations (3.8) - (3.10) for details. For $P_{Z|F}$, two assumptions are considered: select for better outcome probabilistically and select for better outcome deterministically. See equations (3.11) - (3.12) for details. For $E$, three assumptions are considered: corrupted sampling, contaminated sampling and asymmetric misclassification. See equations (3.13) - (3.16) for details. When the misclassification is asymmetric, positive errors are not allowed for $Z$ and negative errors are not allowed for $Y$.

Figure 3.4 - 3.5 show the identified regions of $\mathbf{P}^{TE1}$ and $\mathbf{P}^{ATE}$ under various assumptions on $P_F$ and $P_{Z|F}$. For all plots in these two figures, the corrupted sampling model is maintained. The header row specifies the assumption on $P_{Z|F}$, which gets stronger from the left to the right. The header column specifies the assumption on $P_F$. The largest identified region is observed in the top-left corner (no assumption on $P_F$ and $P_{Z|F}$) and the smallest identified region is observed in the bottom-right corner (MTR at individual level and selection for better outcome deterministically). The effects of different assumptions are not always trivial, but some are very straightforward. For example, when MTR on average is imposed, we will have $P_{T_1}(1) \geq P_{T_1}(-1)$ and $P_{T_1}(1) + P_{T_1}(0) + P_{T_1}(-1) = 1$. Equivalently, $2P_{T_1}(1) + P_{T_1}(0) \geq 1$, which explains the third row of Figure 3.4. If MTR at the individual level is imposed, then $P_{T_1}(-1)$ is forced to zero and we will have $P_{T_1}(1) + P_{T_1}(0) = 1$, which explains the fourth row of Figure 3.4.

Figure 3.6 - 3.7 show the identified regions of $\mathbf{P}^{TE1}$ and $\mathbf{P}^{ATE}$ under various assumptions on the misclassification process. For all plots in these two figures, the assumptions of MTR on average and selection for better outcome probabilistically are maintained. The header row specifies which variable is subject to misclassification and the header column specifies the type of misclassification. In both Figure 3.6 and 3.7, the red areas are the same in all eight plots because they are the identified regions without measurement error and, as a result, are not affected by the assumptions on the patterns of measurement error.

[4] The following definition of percentile is adopted: $Q_\alpha(X) = \inf\{t : F_X(t) \geq \alpha\}$.

## 3.5 Conclusion

Endogeneity and measurement error are almost unavoidable in survey data and pose threats to reliable estimation of treatment effects. In this chapter, a unifying framework is proposed to address these two problems simultaneously. This new framework is based on a novel additive misclassification matrix such that many widely adopted assumptions on the patterns of measurement error can be easily formulated as linear constraints. As a result, the bounds on treatment effects can be obtained by solving linear programing problems, whose solutions have been well studied and algorithms have been widely available. Compared with conventional analytical approach, the newly proposed numerical approach is general enough to be applied to various different problems and guarantees sharp bounds.

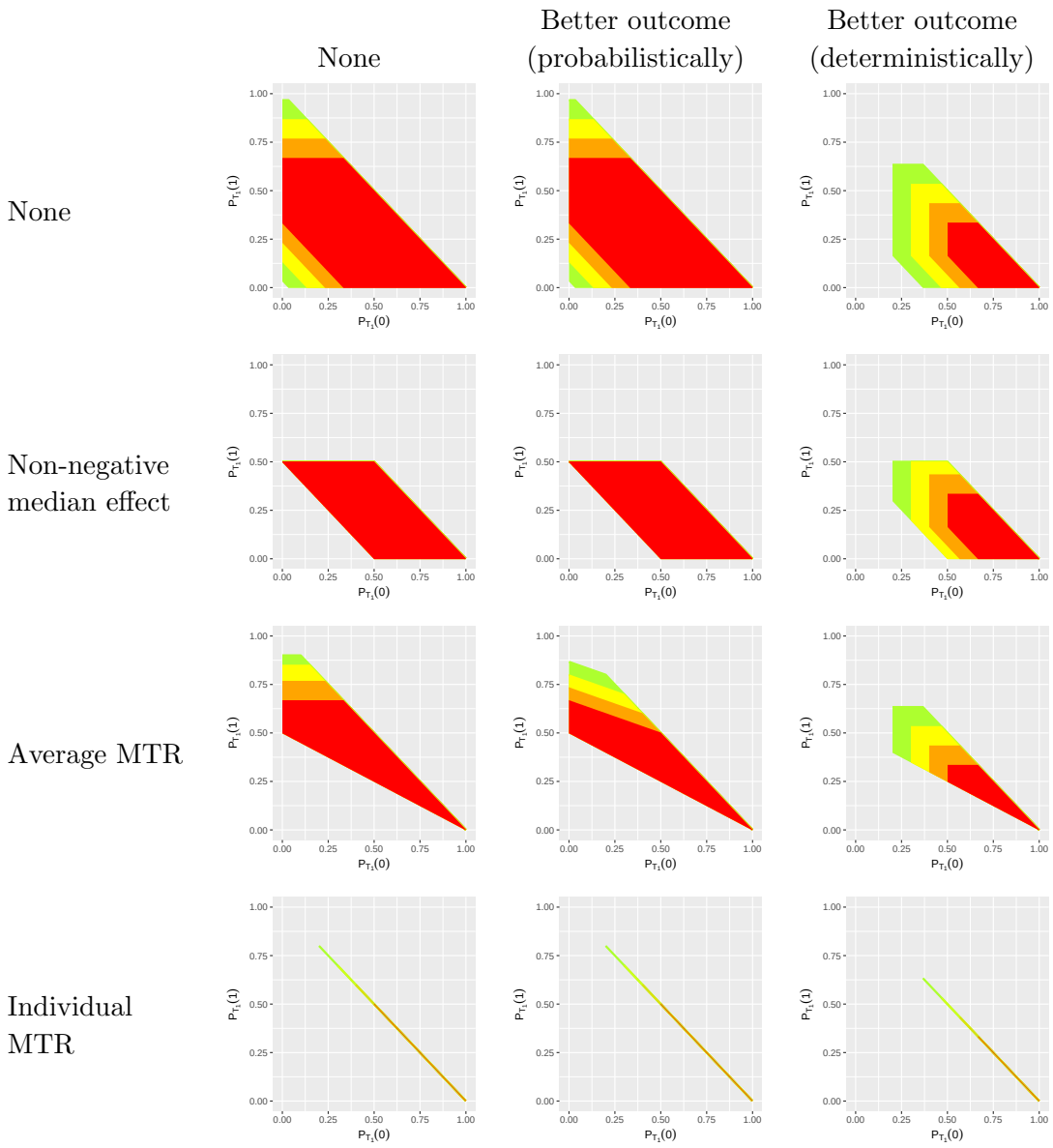



Figure 3.4: The geometry of $\mathbf{P}^{TE1}$ under different assumptions on $P_F$ and $P_{Z|F}$

Note: The header row specifies the assumptions on $P_{Z|F}$ and the header column specifies the assumptions on $P_F$. Each cell is a plot of the geometry of $\mathbf{P}^{TE1}$ under a combination of assumptions on $P_{Z|F}$ and $P_F$. The X-axis denotes the probability of zero treatment effect. The Y-axis denotes the probability of unit treatment effect. In each plot, multiple identified regions under the corrupted sampling model are presented, with different colors denoting the maximum misclassification rates: 0% in red, 10% in orange, 20% in yellow and 30% in green.

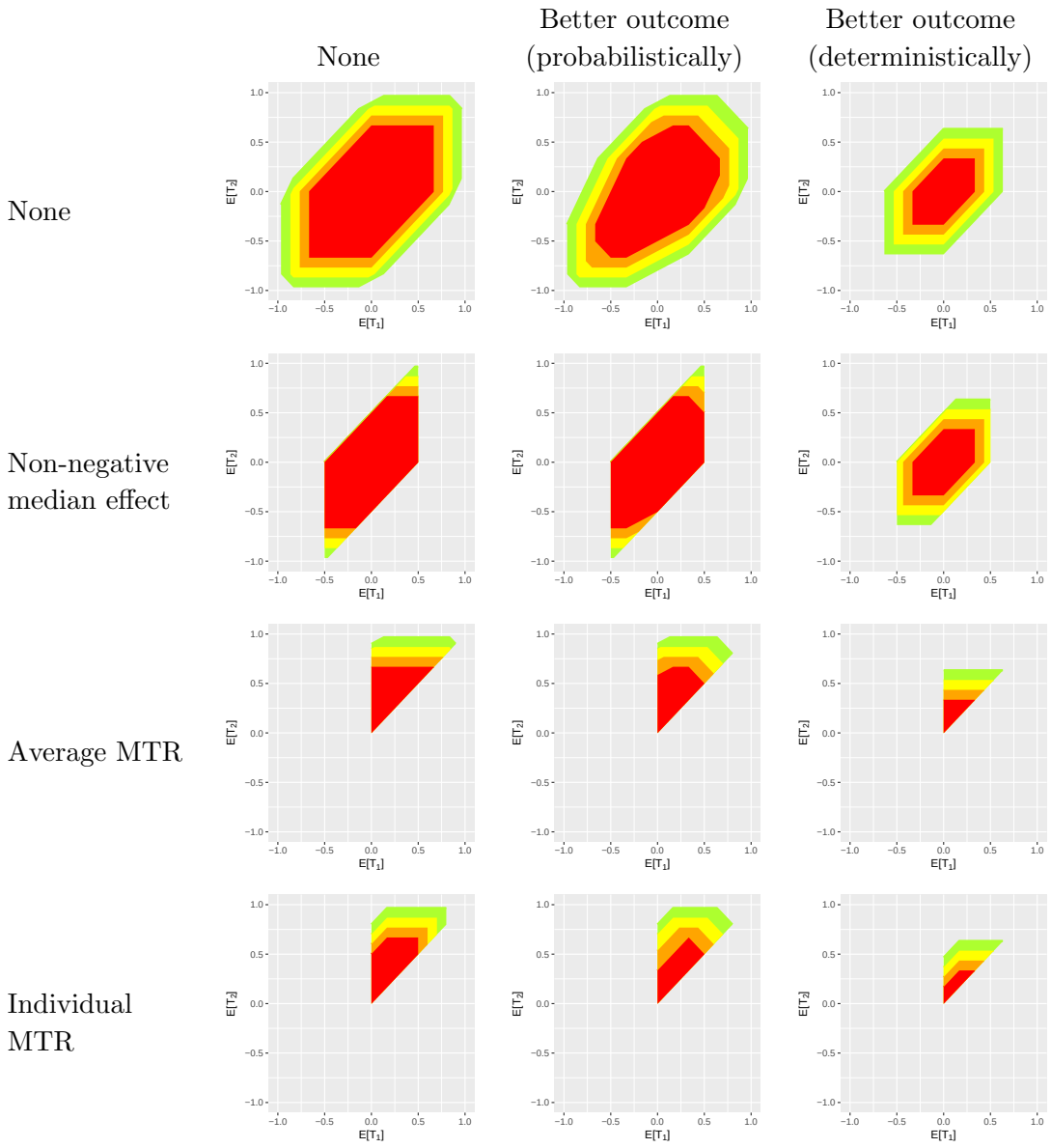



Figure 3.5: The geometry of $\mathbf{P}^{ATE}$ under different assumptions on $P_F$ and $P_{Z|F}$

Note: The header row specifies the assumptions on $P_{Z|F}$ and the header column specifies the assumptions on $P_F$. Each cell is a plot of the geometry of $\mathbf{P}^{ATE}$ under a combination of assumptions on $P_{Z|F}$ and $P_F$. The X-axis denotes the average treatment effect of the first treatment. The Y-axis denotes the average treatment effect of the second treatment. In each plot, multiple identified regions under the corrupted sampling model are presented, with different colors denoting the maximum misclassification rates: 0% in red, 10% in orange, 20% in yellow and 30% in green.

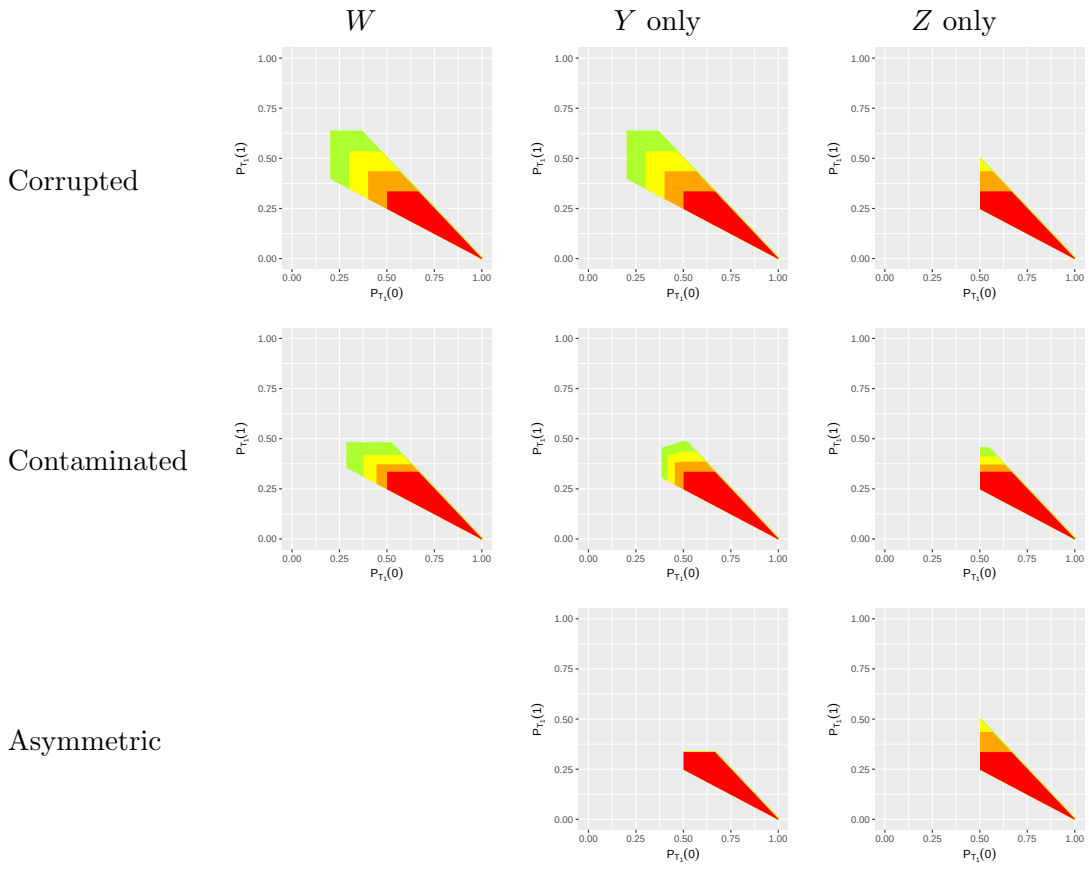



Figure 3.6: The geometry of $\mathbf{P}^{TE1}$ under different assumptions on $E$

Note: The header row specifies which variable is subject to misclassification and the header column specifies the type of misclassification. Each cell is a plot of the geometry of $\mathbf{P}^{TE1}$ based on the assumptions of MTR on average and select for better outcome probabilistically. The X-axis denotes the probability of zero treatment effect. The Y-axis denotes the probability of unit treatment effect. Since a complete ordering does not exist for $W = (Y, Z)$, the left bottom is left blank. In each plot, multiple identified regions are presented, with different colors denoting the maximum misclassification rates: 0% in red, 10% in orange, 20% in yellow and 30% in green.

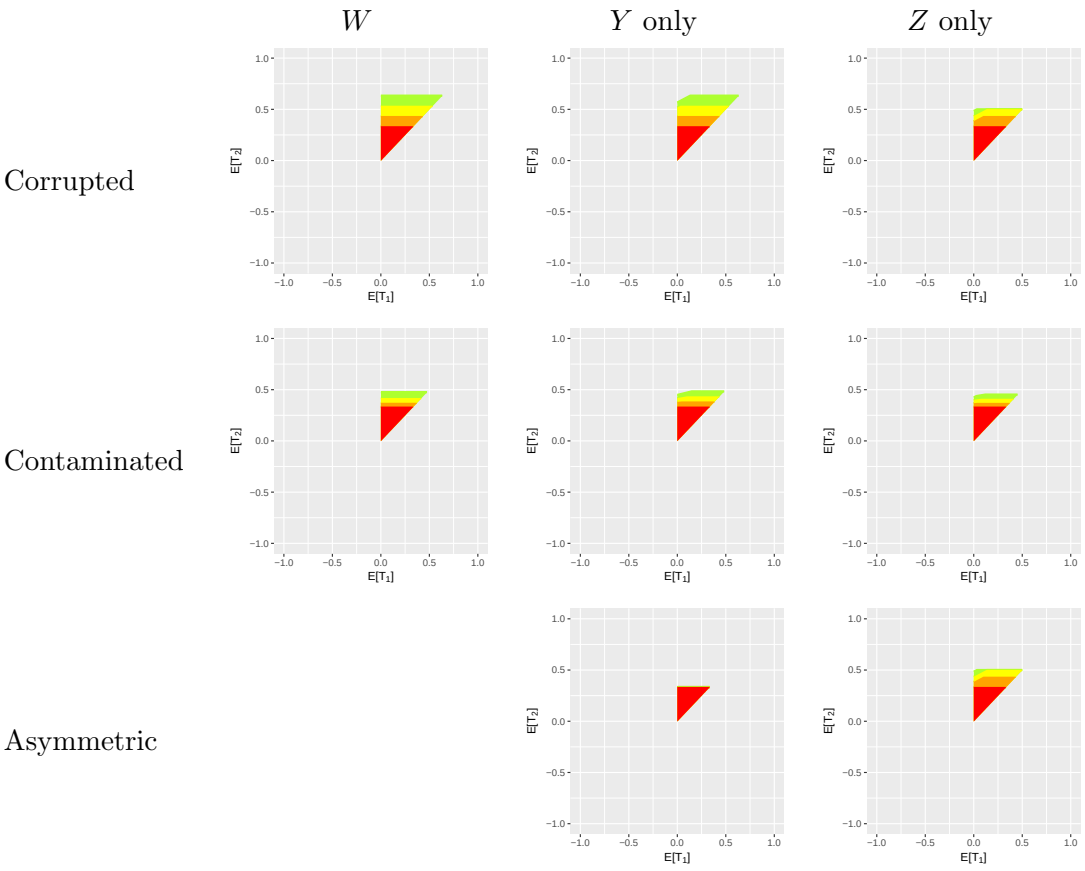



Figure 3.7: The geometry of $\mathbf{P}^{ATE}$ under different assumptions on $E$

Note: The header row specifies which variable is subject to misclassification and the header column specifies the type of misclassification. Each cell is a plot of the geometry of $\mathbf{P}^{ATE}$ based on the assumptions of MTR on average and select for better outcome probabilistically. The X-axis denotes the average treatment effect of the first treatment. The Y-axis denotes the average treatment effect of the second treatment. Since a complete ordering does not exist for $W = (Y, Z)$, the left bottom is left blank. In each plot, multiple identified regions are presented, with different colors denoting the maximum misclassification rates: 0% in red, 10% in orange, 20% in yellow and 30% in green.

# BIBLIOGRAPHY

Anderson, T. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, pages 46–63.

Andrews, D. W., Moreira, M. J., and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

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

Arai, Y. and Ichimura, H. (2013). Optimal bandwidth selection for differences of nonparametric estimators with an application to the sharp regression discontinuity design. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Arellano, M. (1987). Practitionerscorner: Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Barreca, A. I., Lindo, J. M., and Waddell, G. R. (2016). Heaping-induced bias in regression-discontinuity designs. *Economic Inquiry*, 54(1):268–293.

Bartalotti, O. (2018). Regression discontinuity and heteroskedasticity robust standard errors: Evidence from a fixed-bandwidth approximation. *Journal of Econometric Methods*.

Bartalotti, O. and Brummet, Q. (2017). Regression discontinuity designs with clustered data. In *Regression Discontinuity Designs: Theory and Applications*, pages 383–420. Emerald Publishing Limited.

Bartalotti, O., Brummet, Q., and Dieterle, S. (2017a). A correction for regression discontinuity designs with group-specific mismeasurement of the running variable. *Working paper*.

Bartalotti, O., Calhoun, G., and He, Y. (2017b). Bootstrap confidence intervals for sharp regression discontinuity designs. In *Regression Discontinuity Designs: Theory and Applications*, pages 421–453. Emerald Publishing Limited.

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

Brownstone, D. and Valletta, R. (2001). The bootstrap and multiple imputations: harnessing increased computing power for improved statistical tests. *The Journal of Economic Perspectives*, 15(4):129–141.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2017). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, (just-accepted).

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2012). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*.

Cameron, A. C. and Miller, D. L. (2015). A practitioners guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Card, D., Johnston, A., Leung, P., Mas, A., and Pei, Z. (2015a). The effect of unemployment benefits on the duration of unemployment insurance receipt: New evidence from a regression kink design in missouri, 2003-2013. *American Economic Review*, 105(5):126–30.

Card, D., Lee, D., Pei, Z., and Weber, A. (2014). Local polynomial order in regression discontinuity designs. *Working paper*.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015b). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.

Card, D. and Yakovlev, E. (2014). The causal effect of serving in the army on health: Evidence from regression kink design and russian data. *Working paper*.

Chen, K., Fan, J., and Jin, Z. (2008). Design-adaptive minimax local linear regression for longitudinal/clustered data. *Statistica Sinica*, pages 515–534.

Chen, K. and Jin, Z. (2005). Local polynomial regression analysis of clustered data. *Biometrika*, 92(1):59–74.

Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Chernozhukov, V. and Hansen, C. (2008a). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398.

Chernozhukov, V. and Hansen, C. (2008b). The reduced form: A simple approach to inference with weak instruments. *Economics Letters*, 100(1):68–71.

Chiang, H. D., Hsu, Y.-C., and Sasaki, Y. (2017). A unified robust bootstrap method for sharp/fuzzy mean/quantile regression discontinuity/kink designs. *Working paper*.

Chiang, H. D. and Sasaki, Y. (2016). Causal inference by quantile regression kink designs. *Working paper*.

Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169.

De Haan, M. (2011). The effect of parents schooling on childs schooling: A nonparametric bounds analysis. *Journal of Labor Economics*, 29(4):859–892.

Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics*, 30(3):422–446.

Dong, Y. (2016). Jump or kink? regression probability jump and kink design for treatment effect evaluation. *Working paper*.

Dong, Y. and Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5):1081–1092.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica: Journal of the Econometric Society*, pages 1365–1387.

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, 36(4):767–808.

Dufour, J.-M. and Taamouti, M. (2005). Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica*, 73(4):1351–1365.

Dufour, J.-M. and Taamouti, M. (2007). Further results on projection-based inference in iv regressions with weak, collinear or missing instruments. *Journal of Econometrics*, 139(1):133–153.

Ekeland, I., Galichon, A., and Henry, M. (2010). Optimal transportation and the falsifiability of incompletely specified economic models. *Economic Theory*, 42(2):355–374.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004.

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

Feir, D., Lemieux, T., and Marmer, V. (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, 34(2):185–196.

Feng, S. (2013). Misclassification errors and the underestimation of the us unemployment rate. *The American Economic Review*, 103(2):1054–1070.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49(2):361–376.

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

Gelman, A. and Imbens, G. (2017). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, (just-accepted).

Gerfin, M. and Schellhorn, M. (2006). Nonparametric bounds on the effect of deductibles in health care insurance on doctor visits–swiss evidence. *Health economics*, 15(9):1011–1020.

Ginther, D. K. (2000). Alternative estimates of the effect of schooling on earnings. *Review of Economics and Statistics*, 82(1):103–116.

Gonzalez, L. (2005). Nonparametric bounds on the returns to language skills. *Journal of Applied Econometrics*, 20(6):771–795.

Gundersen, C. and Kreider, B. (2009). Bounding the effects of food insecurity on childrens health outcomes. *Journal of health economics*, 28(5):971–983.

Gundersen, C., Kreider, B., and Pepper, J. (2012). The impact of the national school lunch program on child health: A nonparametric bounds analysis. *Journal of Econometrics*, 166(1):79–91.

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

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika*, 75(4):661–671.

Honoré, B. E. and Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74(3):611–629.

Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959.

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

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803.

Kreider, B. and Hill, S. C. (2009). Partially identifying treatment effects with an application to covering the uninsured. *Journal of Human Resources*, 44(2):409–449.

Kreider, B., Pepper, J. V., Gundersen, C., and Jolliffe, D. (2012). Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported. *Journal of the American Statistical Association*, 107(499):958–975.

Lafférs, L. (2013). A note on bounding average treatment effects. *Economics Letters*, 120(3):424–428.

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

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

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

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Ludwig, J. and Miller, D. L. (2005). Does head start improve children's life chances? evidence from a regression discontinuity design. Technical report, National Bureau of Economic Research.

Ludwig, J. and Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1):159–208.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis*, pages 437–461. Springer.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. (1997). Monotone treatment response. *Econometrica: Journal of the Econometric Society*, pages 1311–1334.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010.

Manski, C. F. and Pepper, J. V. (2009). More on monotone instrumental variables. *The Econometrics Journal*, 12(s1):S200–S216.

Manski, C. F. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546.

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

Meyer, B. D., Mok, W. K., and Sullivan, J. X. (2009). The under-reporting of transfers in household surveys: its nature and consequences. Technical report, National Bureau of Economic Research.

Mikusheva, A. et al. (2013). Survey on statistical inferences in weakly-identified instrumental variable models. *Applied Econometrics*, 29(1):117–131.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117.

Moreira, M. J. (2002). *Tests with correct size in the simultaneous equations model*. PhD thesis, University of California, Berkeley.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, 152(2):131–140.

Otsu, T., Xu, K.-L., and Matsushita, Y. (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics*, 186(1):94–112.

Porter, J. (2003). Estimation in the regression discontinuity model. *Working paper*, pages 5–19.

Qu, Z., Yoon, J., et al. (2015). Uniform inference on quantile effects under sharp regression discontinuity designs. Technical report, Boston University-Department of Economics.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sekhon, J. S. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In *Regression discontinuity designs: Theory and applications*, pages 1–28. Emerald Publishing Limited.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

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

Trochim, W. and Spiegelman, C. (1980). The relative assignment variable approach to selection bias in pretest-posttest group designs. *Proceedings of the Survey Research Section*, pages 376–80.

Trochim, W. M. (1984). *Research design for program evaluation: The regression-discontinuity approach*, volume 6. SAGE Publications, Inc.

Turner, L. J. (2017). The economic incidence of federal student grant aid. *Working paper*.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90(1):43–52.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *The American Economic Review*, 93(2):133–138.

# APPENDIX A. ADDITIONAL MATERIAL FOR CHAPTER 1

## A.1 Additional simulation results

Table A.1: Percent Rejected under $H_0 : \tau = 1$ at Nominal Level of 10%

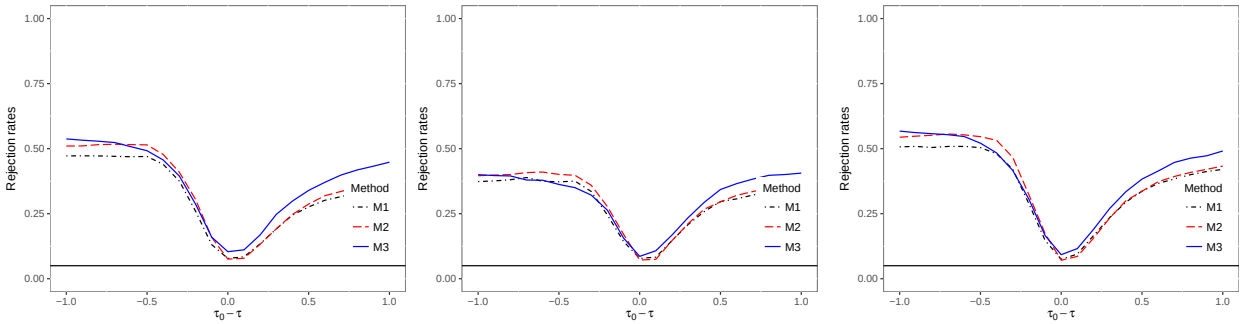
| $\rho$ | $t_j$ | $t_k$ | $AR_j$ | $AR_k$ | $AR$ | $LM$ | $CLR$ |
|---|---|---|---|---|---|---|---|
| | | | Panel A: $\Upsilon_j = \Upsilon_k = 1$ | | | | |
| 0.0 | 0.4 | 0.8 | 10.8 | 10.1 | 10.6 | 10.5 | 10.4 |
| -0.9 | 21.4 | 18.8 | 10.2 | 10.2 | 10.3 | 10.0 | 10.1 |
| 0.9 | 19.1 | 20.0 | 10.1 | 9.3 | 10.0 | 10.5 | 7.4 |
| | | | Panel B: $\Upsilon_j = \Upsilon_k = 10$ | | | | |
| 0.0 | 4.6 | 4.8 | 9.8 | 9.6 | 10.1 | 9.8 | 10.0 |
| -0.9 | 11.5 | 11.3 | 10.1 | 10.8 | 9.9 | 10.4 | 10.8 |
| 0.9 | 10.5 | 10.4 | 9.5 | 9.8 | 9.4 | 9.8 | 10.4 |
| | | | Panel C: $\Upsilon_j = \Upsilon_k = 100$ | | | | |
| 0.0 | 9.0 | 9.3 | 9.6 | 9.7 | 9.4 | 10.2 | 10.5 |
| -0.9 | 8.3 | 9.8 | 10.2 | 10.4 | 10.0 | 10.1 | 10.2 |
| 0.9 | 10.2 | 10.1 | 11.2 | 10.5 | 11.2 | 10.1 | 10.1 |

Table A.2: Percent Rejected at Nominal Level of 5% with $N = 500$

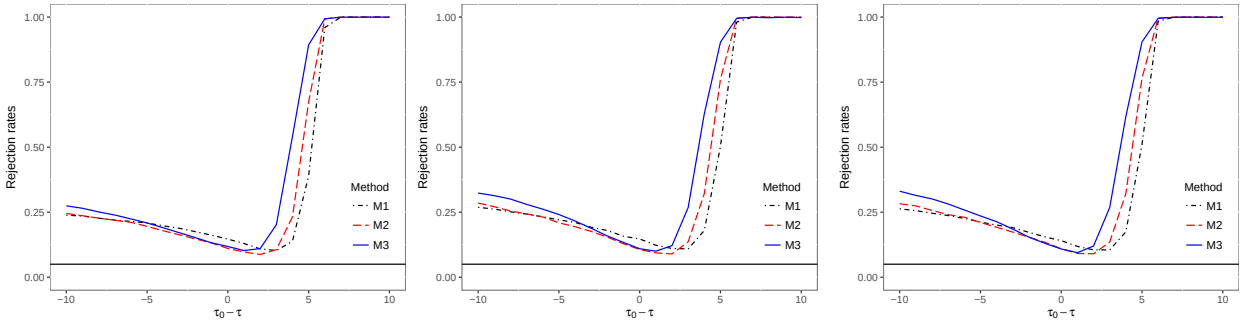
| Method | $M1$ | $M2$ | $M3$ | $M1$ | $M2$ | $M3$ |
|---|---|---|---|---|---|---|
| Bias correction | No | No | No | Yes | Yes | Yes |
| Test | Panel A: $\mu_1^Y$ and $\mu_1^T$ | | | | | |
| $AR$ | 9.5 | 9.6 | 16.9 | 7.8 | 7.5 | 10.4 |
| $LM$ | 9.8 | 10.0 | 12.4 | 7.8 | 7.2 | 8.6 |
| $CLR$ | 9.6 | 9.8 | 13.9 | 7.8 | 7.2 | 9.4 |
| | Panel B: $\mu_2^Y$ and $\mu_1^T$ | | | | | |
| $AR$ | 16.1 | 11.9 | 14.1 | 14.6 | 11.1 | 11.8 |
| $LM$ | 15.4 | 12.2 | 12.2 | 14.7 | 10.6 | 10.9 |
| $CLR$ | 14.3 | 12.8 | 11.7 | 14.1 | 10.8 | 10.4 |
| | Panel C: $\mu_1^Y$ and $\mu_2^T$ | | | | | |
| $AR$ | 9.6 | 9.8 | 17.0 | 7.8 | 7.8 | 10.0 |
| $LM$ | 10.1 | 10.0 | 12.6 | 8.0 | 7.7 | 8.7 |
| $CLR$ | 9.6 | 10.3 | 13.5 | 7.6 | 7.6 | 9.2 |
| | Panel D: $\mu_2^Y$ and $\mu_2^T$ | | | | | |
| $AR$ | 15.8 | 12.2 | 14.5 | 14.5 | 11.3 | 12.2 |
| $LM$ | 15.4 | 12.2 | 12.5 | 14.7 | 10.2 | 11.3 |
| $CLR$ | 14.3 | 12.3 | 11.8 | 14.1 | 10.4 | 11.0 |

Table A.3: Percent Rejected at Nominal Level of 5% with $N = 10000$

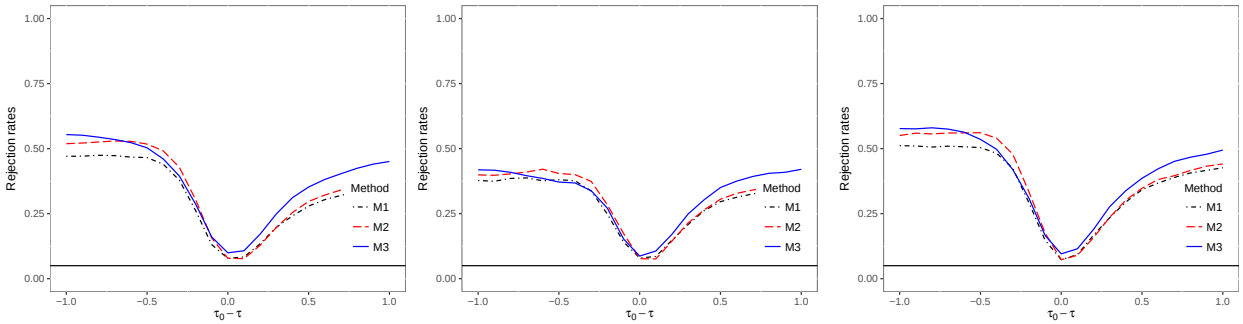
| Method | $M1$ | $M2$ | $M3$ | $M1$ | $M2$ | $M3$ |
|---|---|---|---|---|---|---|
| Bias correction | No | No | No | Yes | Yes | Yes |
| Test | Panel A: $\mu_1^Y$ and $\mu_1^T$ | | | | | |
| $AR$ | 22.4 | 48.8 | 13.2 | 9.1 | 20.2 | 7.0 |
| $LM$ | 22.8 | 51.6 | 11.8 | 11.1 | 22.9 | 6.9 |
| $CLR$ | 22.9 | 52.3 | 12.0 | 11.5 | 23.2 | 7.0 |
| | Panel B: $\mu_2^Y$ and $\mu_1^T$ | | | | | |
| $AR$ | 5.8 | 7.1 | 6.5 | 5.5 | 5.4 | 4.8 |
| $LM$ | 5.2 | 7.2 | 6.6 | 5.2 | 5.1 | 5.6 |
| $CLR$ | 5.4 | 6.7 | 6.8 | 5.2 | 5.3 | 6.0 |
| | Panel C: $\mu_1^Y$ and $\mu_2^T$ | | | | | |
| $AR$ | 22.1 | 50.5 | 13.4 | 8.9 | 20.6 | 6.8 |
| $LM$ | 22.9 | 51.7 | 11.9 | 11.1 | 23.8 | 6.4 |
| $CLR$ | 23.0 | 52.4 | 12.1 | 11.4 | 24.0 | 6.9 |
| | Panel D: $\mu_2^Y$ and $\mu_2^T$ | | | | | |
| $AR$ | 5.4 | 6.6 | 7.0 | 5.5 | 5.4 | 5.1 |
| $LM$ | 5.2 | 6.8 | 6.8 | 5.0 | 5.0 | 5.8 |
| $CLR$ | 5.2 | 6.8 | 6.7 | 5.3 | 5.2 | 6.1 |



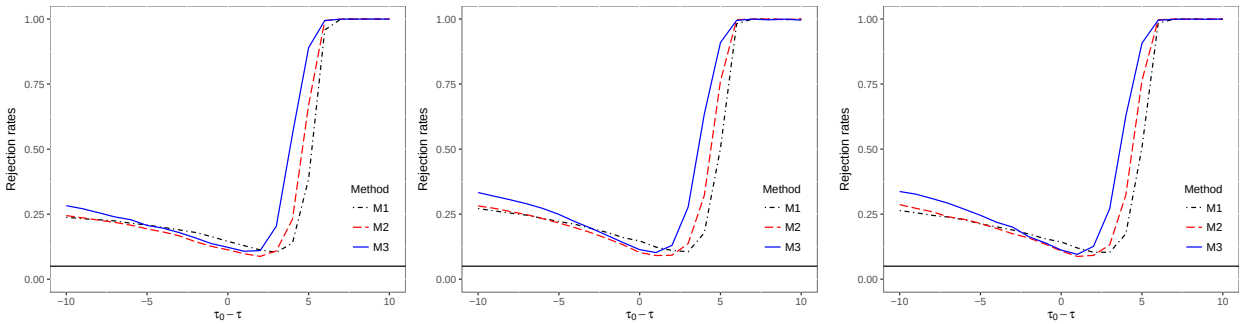

(a) AR test with $\mu_1^Y$ and $\mu_1^T$ (b) LM test with $\mu_1^Y$ and $\mu_1^T$ (c) CLR test with $\mu_1^Y$ and $\mu_1^T$




(d) AR test with $\mu_2^Y$ and $\mu_1^T$ (e) LM test with $\mu_2^Y$ and $\mu_1^T$ (f) CLR test with $\mu_2^Y$ and $\mu_1^T$

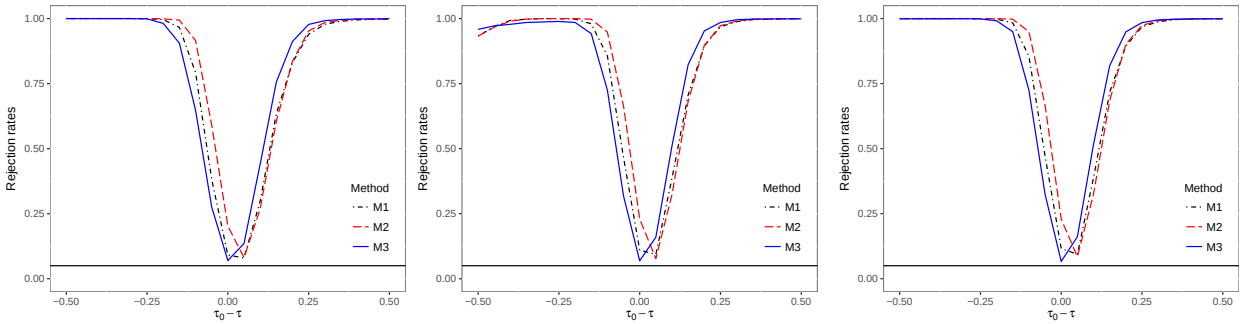



(g) AR test with $\mu_1^Y$ and $\mu_2^T$ (h) LM test with $\mu_1^Y$ and $\mu_2^T$ (i) CLR test with $\mu_1^Y$ and $\mu_2^T$

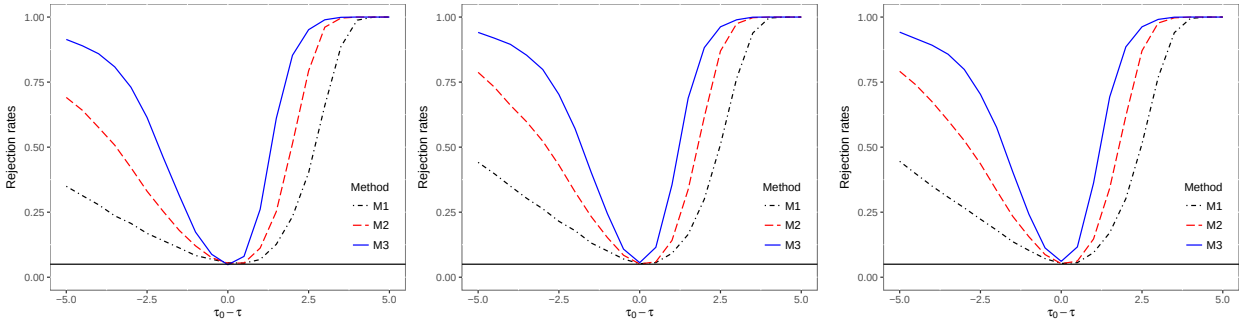



(j) AR test with $\mu_2^Y$ and $\mu_2^T$ (k) LM test with $\mu_2^Y$ and $\mu_2^T$ (l) CLR test with $\mu_2^Y$ and $\mu_2^T$

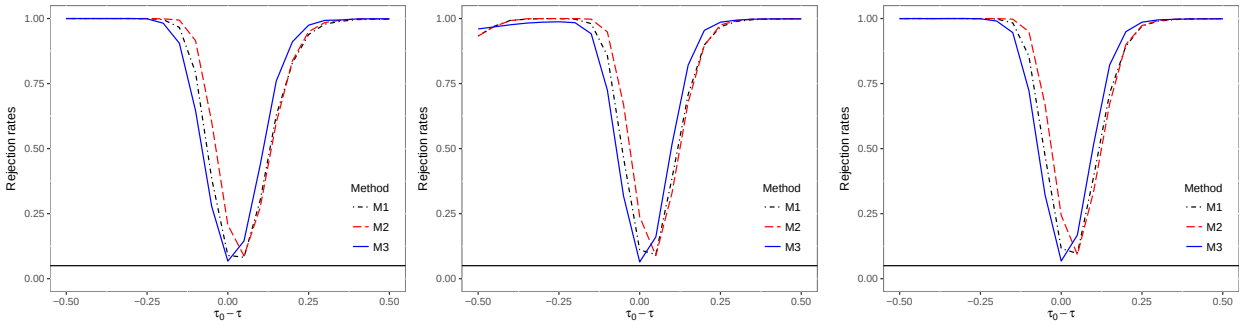
Figure A.2: Power of Bias-corrected Tests at Nominal Level of 5% with $N = 500$



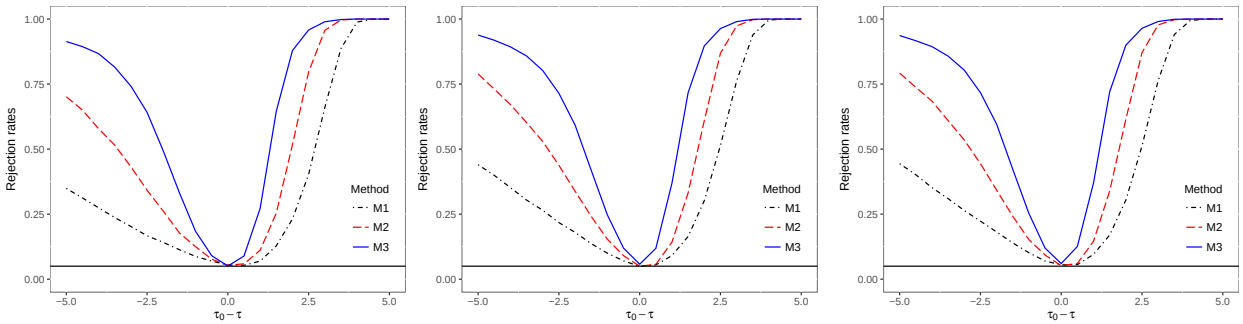

(a) AR test with $\mu_1^Y$ and $\mu_1^T$ (b) LM test with $\mu_1^Y$ and $\mu_1^T$ (c) CLR test with $\mu_1^Y$ and $\mu_1^T$




(d) AR test with $\mu_2^Y$ and $\mu_1^T$ (e) LM test with $\mu_2^Y$ and $\mu_1^T$ (f) CLR test with $\mu_2^Y$ and $\mu_1^T$




(g) AR test with $\mu_1^Y$ and $\mu_2^T$ (h) LM test with $\mu_1^Y$ and $\mu_2^T$ (i) CLR test with $\mu_1^Y$ and $\mu_2^T$




(j) AR test with $\mu_2^Y$ and $\mu_2^T$ (k) LM test with $\mu_2^Y$ and $\mu_2^T$ (l) CLR test with $\mu_2^Y$ and $\mu_2^T$

Figure A.4: Power of Bias-corrected Tests at Nominal Level of 5% with $N = 10000$

## A.2 Proofs

### A.2.1 Proof of Lemma 1.1

The distribution of $W_n$ is $W_n \sim N(\mu, \Omega_n)$ with

$$\mu = (\tau\Pi, \tau\Pi' + \tau'\Pi, \Pi, \Pi').$$

Given a single observation of $W_n$, $w$, and a known variance $\Omega_n$, $w$ is a sufficient statistic for $\theta$ because the factorization theorem naturally holds for

$$f(w|\theta) = (2\pi)^{-1/2}|\Omega|^{-1/2} \exp\left(-\frac{1}{2}(w-\mu)\Omega^{-1}(w-\mu)'\right).$$

Note that $W_n$ is a function of $S_n$ and $T_n$

$$W_n^T = [S_n^T : T_n^T][B_0(B_0^T\Omega_n B_0)^{-\frac{1}{2}} : \Omega_n^{-1}A_0(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}]^{-1},$$

hence $S_n$ and $T_n$ are sufficient statistics for $\theta$, and part (a) of this lemma holds.

To prove part (b), firstly note that $S_n$ and $T_n$ are jointly normal. Their mean and variance are

$$\begin{aligned}
\mathbb{E}(S_n) &= (B_0^T\Omega_n B_0)^{-\frac{1}{2}}(B_0 - B + B)^T\,\mathbb{E}(W_n) = (B_0^T\Omega_n B_0)^{-\frac{1}{2}}(B_0 - B)^T\mu,\\
\mathbb{V}(S_n) &= (B_0^T\Omega_n B_0)^{-\frac{1}{2}}B_0^T\,\mathbb{V}(W_n^T)B_0(B_0^T\Omega_n B_0)^{-\frac{1}{2}} = I_2,\\
\mathbb{E}(T_n) &= (A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}A_0^T\Omega_n^{-1}\,\mathbb{E}(W_n) = (A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}A_0^T\Omega_n^{-1}\mu,\\
\mathbb{V}(T_n) &= (A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}A_0^T\Omega_n^{-1}\,\mathbb{V}(W_n^T)\Omega_n^{-1}A_0(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}} = I_2.
\end{aligned}$$

In addition, the covariance between $S_n$ and $T_n$ is

$$\begin{aligned}
\mathbb{C}\text{ov}(S_n, T_n) &= \mathbb{C}\text{ov}\left((B_0^T\Omega_n B_0)^{-\frac{1}{2}}B_0^T W_n, (A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}A_0^T\Omega_n^{-1}W_n\right)\\
&= (B_0^T\Omega_n B_0)^{-\frac{1}{2}}B_0^T\,\mathbb{V}(W_n)\Omega_n^{-1}A_0(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}\\
&= (B_0^T\Omega_n B_0)^{-\frac{1}{2}}B_0^T A_0(A_0^T\Omega_n^{-1}A_0)^{-\frac{1}{2}}\\
&= 0,
\end{aligned}$$

where the last equality holds because $B_0$ and $A_0$ are designed to be orthogonal. As a result, $S_n$ and $T_n$ are independent. □

### A.2.2 Proof of Theorem 1.1

Firstly notice that since $\widehat{W}_n$ is asymptotically normal, $\widehat{S}_n$ and $\widehat{T}_n$ are asymptotically joint normal as they are linear transformations of $\widehat{W}_n$. We show that $\widehat{S}_n \to^d S_n$, $\widehat{T}_n \to^d T_n$, and $\widehat{S}_n$ and $\widehat{T}_n$ are asymptotically uncorrelated.

$$\begin{aligned}
\widehat{S}_n &= B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T \widehat{W}_n \\
&= (B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T (\widehat{W}_n - \mu) + (B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T \mu \\
&= (B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T (\widehat{W}_n - \mu) + (B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} (B_0 - B)^T \mu.
\end{aligned}$$

By Cramer-Wold Device, we have $(B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T (\widehat{W}_n - \mu) \to^d N(0, I_2)$; by Slutsky Theorem, we have $(B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} (B_0 - B)^T \mu \to^p (B_0^T \Omega_n B_0)^{-\frac{1}{2}} (B_0 - B)^T \mu$. Hence $\widehat{S}_n \to^d S_n$. Analogously,

$$\begin{aligned}
\widehat{T}_n &= (A_0^T \widehat{\Omega}_n^{-1} A_0)^{-\frac{1}{2}} A_0^T \widehat{\Omega}_n^{-1} \widehat{W}_n \\
&= (A_0^T \widehat{\Omega}_n^{-1} A_0)^{-\frac{1}{2}} A_0^T \widehat{\Omega}_n^{-1} (\widehat{W}_n - \mu) + (A_0^T \widehat{\Omega}_n^{-1} A_0)^{-\frac{1}{2}} A_0^T \widehat{\Omega}_n^{-1} \mu,
\end{aligned}$$

where its first part converge in distribution to standard normal and the second part converge in probability to the mean of $T_n$. As a result, $\widehat{T}_n \to^d T_n$. Finally,

$$\begin{aligned}
\mathbb{C}\text{ov}(\widehat{S}_n, \widehat{T}_n) &= \mathbb{C}\text{ov}\left( (B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T \widehat{W}_n, (A_0^T \widehat{\Omega}_n^{-1} A_0)^{-\frac{1}{2}} A_0^T \widehat{\Omega}_n^{-1} \widehat{W}_n \right) \\
&= (B_0^T \widehat{\Omega}_n B_0)^{-\frac{1}{2}} B_0^T \, \mathbb{V}(\widehat{W}_n) \widehat{\Omega}_n^{-1} A_0 (A_0^T \widehat{\Omega}_n^{-1} A_0)^{-\frac{1}{2}} \\
&\to^p 0
\end{aligned}$$

because $\widehat{\Omega}_n \to^p \mathbb{V}(\widehat{W}_n)$ and $B_0^T A_0 = 0$. Part (a) of Theorem 1.1 holds.

The statistic $\phi(\cdot, \cdot, \cdot, \tau_0, \tau_0')$ is, by definition, a continuous function. The critical value function $c_\phi(\cdot, \cdot, \tau_0, \tau_0', \alpha)$ is also a continuous function because the conditional distribution of $S_n$ given $T_n$ is absolutely continuous with a density that is smooth function of $T_n$. Hence part (b) of this theorem holds by continuous mapping theorem.

Part (c) follows immediately from part (b) because under the null,

$$Pr\big(\psi(\widehat{S}_n, \widehat{T}_n, \widehat{\Omega}_n, \tau_0, \tau_0') > c_\psi(\widehat{T}_n, \widehat{\Omega}_n, \tau_0, \tau_0', \alpha)\big) \to^p Pr\big(\psi(S_n, T_n, \Omega_n, \tau_0, \tau_0') > c_\psi(T_n, \Omega_n, \tau_0, \tau_0', \alpha)\big) = \alpha,$$

where the equality holds by definition of the critical value function. □

### A.2.3 Proof of Lemma 1.4

Assumption 1.1 ensures that $f_{Y(0)|X}(y|x) = \int_{Y(1)}\int_{T(1)}\int_{T(0)} f_{S|X}(s|x)ds$ is continuous at the threshold, implying the continuity of $y(0,x,p)$ because

$$y(0,x,p) = \min_a \int_{-\infty}^{a} f_{Y_0|X}(u|x)du = q.$$

Again, Assumption 1.1 ensures that $\frac{\partial f_{Y(0)|X}(y|x)}{\partial x} = \int_{Y(1)}\int_{T(1)}\int_{T(0)} \frac{\partial f_{S|X}(s|x)}{\partial x}ds$ is continuous at the threshold. Note that

$$\frac{\partial y(0,x,p)}{\partial x} = -\frac{\frac{\partial F_{Y(0)|X}(y|x)}{\partial x}}{f_{Y(0)|X}(y|x)}\bigg|_{y=y(0,x,p)} = -\frac{\int_{-\infty}^{y}\frac{\partial f_{Y(0)|X}(u|x)}{\partial x}du}{f_{Y(0)|X}(y|x)}\bigg|_{y=y(0,x,p)},$$

so $y(0,x,p)$ has continuous first order derivative with respect to the running variable at the threshold. Similar argument can be made to $y(1,x,p)$. Hence part (a) of this lemma holds.

Let $\tau(x,p) = y(1,x,p) - y(0,x,p)$ be the quantile treatment effect and $\tau'(x,p)$ be its first order derivative with respect to the running variable. Theorem 1 from Chernozhukov and Hansen (2005) ensures that[1]

$$\Pr[Y_i \le y(T_i, X_i, p)|X_i] = p \ \ \forall p \in (0,1).$$

Notice that $y(T_i,X_i,p) = y(0,X_i,p) + T_i\big(y(1,X_i,p) - y(0,X_i,q)\big) = y(0,X_i,p) + T_i\tau(X_i,p)$. As a result, $\Pr[Y_i - T_i\tau(X_i,p) \le y(0,X_i,p)|X_i] = p$, or equivalently, $q_p\big(Y_i - T_i\tau(X_i,p)|X_i\big) = y(0,X_i,p)$.

Since $q_p\big(Y_i - T_i\tau(X_i,p)|X_i\big)$ has the same smoothness properties as $y(0,X_i,p)$, it suffices to show that

$$\lim_{x\to 0} q_p(Y_i^*|X_i = x) = q_p\big(Y_i - T_i\tau(X_i,p)|X_i = 0\big) \tag{A.1}$$

and

$$\lim_{x\to 0}\frac{\partial q_p(Y_i^*|X_i=x)}{\partial x} = \frac{\partial q_p\big(Y_i - T_i\tau(X_i,p)|X_i = x\big)}{\partial x}\bigg|_{x=0}. \tag{A.2}$$

Equality (A.1) is trivial by the definition of $Y_i^*$. Equality (A.2) holds following the proof below.

[1] They result is conditioning on instrumental variable $Z$, which is a fully determined by $X$ and is thus dropped.

$$
\begin{aligned}
&\frac{\partial q_p\big(Y_i - T_i\tau(X_i,p)|X_i = x\big)}{\partial x} \\
=& -\left.\frac{\frac{\partial F_{Y-T\tau(X,p)|X}(y|x)}{\partial x}}{f_{Y-T\tau(X,p)|X}(y|x)}\right|_{y=y(0,x,p)} \\
=& -\left.\frac{\frac{\partial \int_{-\infty}^{y} f_{Y-T\tau(X,p)|X}(u|x)du}{\partial x}}{f_{Y-T\tau(X,p)|X}(y|x)}\right|_{y=y(0,x,p)} \\
=& -\left.\frac{\frac{\partial \int_{-\infty}^{y} \Pr[T=0|x] f_{y(0,X,\epsilon_0)|(X,T=0)}(u|x) + \Pr[T=1|x] f_{y(1,X,\epsilon_1)-\tau(X,p)|(X,T=1)}(u|x)du}{\partial x}}{f_{Y-T\tau(X,p)|X}(y|x)}\right|_{y=y(0,x,p)}.
\end{aligned}
$$

Analogously,

$$
\frac{\partial q_p(Y_i^*|X_i = x)}{\partial x} = -\left.\frac{\frac{\partial \int_{-\infty}^{y} \Pr[T=0|x] f_{y(0,X,\epsilon_0)|(X,T=0)}(u|x) + \Pr[T=1|x] f_{y(1,X,\epsilon_1)-(\tau(p)+X\tau'(p))|(X,T=1)}(u|x)du}{\partial x}}{f_{Y-T(\tau(p)+X\tau'(p))|X}(y|x)}\right|_{y=y(0,x,p)}
$$

After comparing the difference between $\frac{\partial q_p(Y_i^*|X_i=x)}{\partial x}$ and $\frac{\partial q_p\left(Y_i - T_i\tau(X_i,p)|X_i=x\right)}{\partial x}$, it can be shown that

$$
\lim_{x\to 0} f_{Y-T(\tau(p)+X\tau'(p))|X}(y|x) = f_{Y-T\tau(X,p)|X}(y|x)\big|_{x=0},
$$

and

$$
\begin{aligned}
&\lim_{x\to 0} \frac{\partial f_{y(1,X,\epsilon)-(\tau(p)+X\tau'(p))|(X,T=1)}(u|x)}{\partial x} \\
=& \lim_{x\to 0} \frac{\partial f_{y(1,X,\epsilon_1)|(X,T=1)}(u + (\tau(p) + x\tau'(p))|x)}{\partial x}\tau'(p) \\
=& \frac{\partial f_{y(1,X,\epsilon_1)|(X=x,T=1)}(u + \tau(0,p))}{\partial x}\tau'(0,p) \\
=& \left.\frac{\partial f_{y(1,X,\epsilon)|(X,T=1)}(u + \tau(x,p)|x)}{\partial x}\tau'(x,p)\right|_{x=0} \\
=& \left.\frac{\partial f_{y(1,X,\epsilon)-\tau(X,p)|(X,T=1)}(u|x)}{\partial x}\right|_{x=0}.
\end{aligned}
$$

As a result, equation (A.2) holds. □

## A.3 Additional mathematic notes

### A.3.1 The statistic for likelihood ratio test

Given $\Omega_n$, the log likelihood function of observing $W_n$ is

$$\ln L(W_n|\tau,\tau',\Pi,\Pi') = -\ln(2\pi) - \frac{1}{2}\ln(|\Omega_n|) - \frac{1}{2}(W_n-\mu)^T\Omega_n^{-1}(W_n-\mu),$$

with $\mu = (\tau\Pi, \tau'\Pi + \tau\Pi', \Pi, \Pi')^T$. To remove nuisance parameter $(\Pi, \Pi')$, let $\tilde{\mu} = A(\Pi,\Pi')^T$ and assume $(\tau,\tau')$ is fixed. Then to maximize $\ln L(W_n|\tau,\tau',\pi,\pi')$ is equivalent to the following restricted optimization problem:

$$\begin{aligned} \max_{\tilde{\mu}} \quad & \ln L(W_n|\tilde{\mu}) = -ln(2\pi) - \frac{1}{2}ln(|\Omega_n|) - \frac{1}{2}(W_n-\tilde{\mu})^T\Omega_n^{-1}(W_n-\tilde{\mu}) \\ s.t. \quad & B^T\tilde{\mu} = 0. \end{aligned}$$

With Lagrange multiplier method, one can obtain $\tilde{\mu}^* = \big(I_4 - \Omega_n B(B^T\Omega_n B)^{-1}B^T\big)W_n$. As a result, the concentrated log likelihood function is

$$\ln L(W_n|\tau,\tau') = -\ln(2\pi) - \frac{1}{2}\ln(|\Omega_n|) - \frac{1}{2}W_n^T B(B^T\Omega_n B)^{-1}B^T W_n.$$

Hence the likelihood ratio statistic is

$$LR_0 = W_n^T B_0(B_0^T\Omega_n B_0)^{-1}B_0^T W_n - \min_{(\tau,\tau')} W_n^T B(B^T\Omega_n B)^{-1}B^T W_n.$$

### A.3.2 The statistic for Lagrange multiplier test

The first order derivative of log likelihood with respect to parameters $(\tau,\tau')^T$ is

$$\frac{\partial \ln L(W_n|\tau,\tau',\Pi,\Pi')}{\partial(\tau,\tau')^T} = (W_n-\mu)^T\Omega_n^{-1}\begin{pmatrix} \Pi & 0 \\ \Pi' & \Pi \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Note that when evaluated at $(\tau_0, \tau_0', \widehat{\Pi}, \widehat{\Pi}')$, the mean is $\tilde{\mu}^* = \big(I_4 - \Omega_n B_0 (B_0^T \Omega_n B_0)^{-1} B_0^T\big) W_n$, hence

$$\begin{aligned}
\left.\frac{\partial \ln L(W_n|\tau, \tau', \Pi, \Pi')}{\partial(\tau, \tau')^T}\right|_{\tau_0, \tau_0', \widehat{\Pi}, \widehat{\Pi}'} &= \big(W_n - \big(I_4 - \Omega_n B_0 (B_0^T \Omega_n B_0)^{-1} B_0^T\big) W_n\big)^T \Omega_n^{-1} \begin{pmatrix} \widehat{\Pi} & 0 \\ \widehat{\Pi}' & \widehat{\Pi} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \\
&= W_n^T B_0 (B_0^T \Omega_n B_0)^{-1} B_0^T \begin{pmatrix} \widehat{\Pi} & 0 \\ \widehat{\Pi}' & \widehat{\Pi} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \\
&= W_n^T B_0 (B_0^T \Omega_n B_0)^{-1} \widehat{\mathbf{\Pi}} \\
&= S_n^T (B_0^T \Omega_n B_0)^{-\frac{1}{2}} \widehat{\mathbf{\Pi}}.
\end{aligned}$$

So the statistic for Lagrange multiplier test is

$$LM_0 = S_n^T (B_0^T \Omega_n B_0)^{-\frac{1}{2}} \widehat{\mathbf{\Pi}} \big(\widehat{\mathbf{\Pi}}^T (B_0^T \Omega_n B_0)^{-1} \widehat{\mathbf{\Pi}}\big)^{-1} \widehat{\mathbf{\Pi}}^T (B_0^T \Omega_n B_0)^{-\frac{1}{2}} S_n,$$

which can be further simplified to $S_n^T S_n$ given that $\widehat{\mathbf{\Pi}}$ is non-singular.

The maximum likelihood estimators for nuisance parameters are obtained by solving the following first order condition:

$$\begin{aligned}
\frac{\partial \ln L(W_n|\tau_0, \tau_0', \Pi, \Pi')}{\partial(\Pi, \Pi')^T} &= (W_n - \mu)^T \Omega_n^{-1} A_0 \\
&= \left(W_n - A_0 \begin{pmatrix} \Pi \\ \Pi' \end{pmatrix}\right)^T \Omega_n^{-1} A_0 \\
&= 0.
\end{aligned}$$

The solution is $(\widehat{\Pi}, \widehat{\Pi}')^T = T_n^T (A_0^T \Omega_n^{-1} A_0)^{-\frac{1}{2}}$ and is independent with $S_n$. As a result, $LM_0$ follows chi-squared distribution with two degrees of freedom under the null hypothesis.

### A.3.3 The estimation of $\hat{\Omega}_n$

By definition, $\widehat{\Omega}_n$ is the variance estimator for $\widehat{W}_n$. Since $\widehat{W}_n$ is the difference between estimators from two sides, which are independent, we have

$$\widehat{\Omega}_n = \mathbb{V}[\widehat{W}_n] = \mathbb{V}[\widehat{W}_n^+] + \mathbb{V}[\widehat{W}_n^-].$$

The same steps can be applied to the calculation of both $\mathbb{V}[\widehat{W}_n^+]$ and $\mathbb{V}[\widehat{W}_n^-]$. The following discussion focuses on $\mathbb{V}[\widehat{W}_n^+]$ only. $\widehat{W}_n^+$ is a vector of bias-corrected intercepts and slopes, i.e.,

$$\widehat{W}_n^+ = \begin{pmatrix} \widehat{\mu}_{Y_+}(h_{Y,0}) - B_{+,0}\widehat{\mu}^{(2)}_{Y_+}(h_{Y,2})h_{Y,0}^2 \\ \widehat{\mu}^{(1)}_{Y_+}(h_{Y,1}) - B_{+,1}\widehat{\mu}^{(2)}_{Y_+}(h_{Y,2})h_{Y,1} \\ \widehat{\mu}_{T_+}(h_{T,0}) - B_{+,0}\widehat{\mu}^{(2)}_{T_+}(h_{T,2})h_{T,0}^2 \\ \widehat{\mu}^{(1)}_{T_+}(h_{T,1}) - B_{+,1}\widehat{\mu}^{(2)}_{T_+}(h_{T,2})h_{T,1} \end{pmatrix}.$$

CCT's Lemma SA4 provides formula for the diagonal elements of $\mathbb{V}[\widehat{W}_n^+]$. For off-diagonal elements, one can make use of the covariance terms provided by CCT's Theorem A2.

# APPENDIX B. ADDITIONAL MATERIAL FOR CHAPTER 2

This appendix adopts CCT's notation where possible and utilizes some conclusions from that paper. Let $e_p$ be the selection vector with 1 in element $p+1$ and 0 everywhere else and assume, with some abuse of notation, that the dimension of $e_p$ adapts to make matrix and vector operations conformable. Much of the theory in this appendix applies to both sides of the cutoff symmetrically, so I use "$\bullet$" as a placeholder for either $+$ or $-$ in equations. Further let $r_p(x) = (1,\ x, \ldots,\ x^p)'$, $\mathbb{1}_+(x) = \mathbb{1}\{x \geq 0\}$, $\mathbb{1}_-(x) = \mathbb{1}\{x < 0\}$, $m = \min(h, b)$ and $\nu \leq p < q$. Define the following terms related to local polynomial regression:

$$\Gamma_{\bullet,p}(h) = \tfrac{1}{n}\sum_{i=1}^{n} r_p(X_i/h) r_p(X_i/h)' K_{\bullet,h}(X_i)$$

$$\Gamma_{\bullet,q}(b) = \tfrac{1}{n}\sum_{i=1}^{n} r_q(X_i/b) r_q(X_i/b)' K_{\bullet,b}(X_i)$$

$$\mathfrak{B}_{\bullet,\nu,p,q}(h) = \nu! e_\nu' \big(\Gamma_{\bullet,p}(h)\big)^{-1} \tfrac{1}{n}\sum_{i=1}^{n} (X_i/h)^q r_p(X_i/h) K_{\bullet,h}(X_i).$$

When $nh \to \infty$, $nm \to \infty$ and $h \to 0$, CCT's Lemma SA.1 and SA.2 imply that these terms have well-defined limits under Assumptions 2.1 and 2.2.

Let $\hat{\beta}_{Z\bullet,p}(h)$ be the coefficient estimators from the weighted regression of $Z_i$ on $r_p(X_i)$:

$$\hat{\beta}_{Z\bullet,p}(h) = H_p(h)\Gamma_{\bullet,p}(h)^{-1} \tfrac{1}{n}\sum_{i=1}^{n} r_p(X_i/h) Z_i K_{\bullet,h}(X_i)$$

with $H_p(h) = diag(1, h^{-1}, \ldots, h^{-p})$. These coefficients are related to the quantities of interest by

$$\hat{\mu}^{(\nu)}_{Z\bullet,p}(h) = \nu! e_\nu' \hat{\beta}_{Z\bullet,p}(h)$$

and

$$\hat{\zeta}_{\nu,p}(h) = \frac{\hat{\mu}^{(\nu)}_{Y+,p}(h) - \hat{\mu}^{(\nu)}_{Y-,p}(h)}{\hat{\mu}^{(\nu)}_{T+,p}(h) - \hat{\mu}^{(\nu)}_{T-,p}(h)}$$

for $\nu = 0, \ldots, p$.

### B.0.1 Proof of Theorem 2.1

Based on the bias calculated from Algorithm 2.1, the difference between the bias-corrected estimator and the true treatment effect is

$$\hat{\zeta}_{\nu,p}(h) - \Delta^*_{\nu,p,q}(h,b) - \zeta_\nu = (\hat{\zeta}_{\nu,p}(h) - \zeta) - (\mathbb{E}^* \frac{\hat{\tau}^*_{Y,\nu,p}(h)}{\hat{\tau}^*_{T,\nu,p}(h)} - \frac{\tau^*_{Y,\nu}}{\tau^*_{T,\nu}}).$$

The first two terms on the right side can be written as

$$\begin{aligned}\hat{\zeta}_{\nu,p}(h) - \zeta_\nu =& \frac{1}{\tau_{T,\nu}}(\hat{\tau}_{Y,\nu,p}(h) - \tau_{Y,\nu}) - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}_{T,\nu,p}(h) - \tau_{T,\nu}) \\ &+ \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}\hat{\tau}_{T,\nu,p}}(\hat{\tau}_{T,\nu,p}(h) - \tau_{T,\nu})^2 - \frac{1}{\tau_{T,\nu}\hat{\tau}_{T,\nu,p}}(\hat{\tau}_{Y,\nu,p}(h) - \tau_{Y,\nu})(\hat{\tau}_{T,\nu,p}(h) - \tau_{T,\nu}) \\ =& \frac{1}{\tau_{T,\nu}}(\hat{\tau}_{Y,\nu,p}(h) - \tau_{Y,\nu}) - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}_{T,\nu,p}(h) - \tau_{T,\nu}) + R_n,\end{aligned}$$

with $R_n = O_p(\frac{1}{nh^{1+2\nu}} + h^{2(p+1-\nu)})$ (CCT's Lemma A.2). Similarly, the last two terms on the right side can be written as

$$\mathbb{E}^* \frac{\hat{\tau}^*_{Y,\nu,p}(h)}{\hat{\tau}^*_{T,\nu,p}(h)} - \frac{\tau^*_{Y,\nu}}{\tau^*_{T,\nu}} = \frac{1}{\tau^*_{T,\nu}}(\mathbb{E}^* \hat{\tau}^*_{Y,\nu,p}(h) - \tau^*_{Y,\nu}) - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\mathbb{E}^* \hat{\tau}^*_{T,\nu,p}(h) - \tau^*_{T,\nu}) + R^*_n,$$

with $R^*_n = O_p(\frac{1}{nh^{1+2\nu}} + h^{2(p+1-\nu)})$. By construction of the wild bootstrap DGP,

$$Z^*_i = \begin{cases} r_q(X_i/b)' H_q(b)^{-1} \beta^*_{Z+,q} + \varepsilon^*_i & X_i \geq 0 \\ r_q(X_i/b)' H_q(b)^{-1} \beta^*_{Z-,q} + \varepsilon^*_i & X_i < 0, \end{cases}$$

with $\beta^*_{Z+,q}$ and $\beta^*_{Z-,q}$ being the true parameters in the bootstrap data. Equivalently, $\mu^{*(\nu)}_{Z^*_\bullet} = \nu! e'_\nu \beta^*_{Z^*_\bullet,q}$ is the true treatment effect in the bootstrap data. CCT's Lemma SA.3 indicates that

$$\mathbb{E}^* \hat{\mu}^{*(\nu)}_{Z^*_\bullet,p}(h) - \mu^{*(\nu)}_{Z^*_\bullet} = h^{1+p-\nu} \mu^{*(1+p)}_{Z^*_\bullet} \mathfrak{B}_{\bullet,\nu,p,1+p}(h)/(1+p)! + O_p(h^{2+p-\nu}),$$

which allows for an analytical form of the bias in the bootstrap data:

$$\mathbb{E}^* \hat{\tau}^*_{Z,\nu,p}(h) - \tau^*_{Z,\nu} = h^{1+p-\nu}\big(\mu^{*(1+p)}_{Z+} \mathfrak{B}_{+,\nu,p,p+1}(h) - \mu^{*(1+p)}_{Z-} \mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! + O_p(h^{2+p-\nu}).$$

Notice that CCT's bias term is only slightly different from this. They use the following formula for bias correction:

$$\hat{\tau}^{bc}_{Z,\nu,p,q}(h,b) = \hat{\tau}_{Z,\nu,p}(h) - h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{Z+,q} \mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}^{(1+p)}_{Z-,q} \mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!.$$

Built on above preparations, it can be shown that

$$\hat{\zeta}_{\nu,p}(h)-\Delta^*_{\nu,p,q}(h,b)-\zeta_\nu = \frac{1}{\tau_{T,\nu}}(\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b)-\tau_{Y,\nu})-\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}^{bc}_{T,\nu,p,q}(h,b)-\tau_{T,\nu})+R_n-R^*_n-R^{*bc}_n+O_p(h^{2+p-\nu}), \tag{B.1}$$

where $R^{*bc}_n$ is defined by:

$$\begin{aligned}
R^{*bc}_n =&\frac{1}{\tau^*_{T,\nu}}h^{1+p-\nu}\big(\mu^{*(1+p)}_{Y+}\mathfrak{B}_{+,\nu,p,p+1}(h)-\mu^{*(1+p)}_{Y-}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&-\frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T.\nu}}h^{1+p-\nu}\big(\mu^{*(1+p)}_{T+}\mathfrak{B}_{+,\nu,p,p+1}(h)-\mu^{*(1+p)}_{T-}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&-\frac{1}{\tau_{T,\nu}}h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{Y+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{Y-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&+\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{T+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{T-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
=&\frac{1}{\hat{\tau}_{T,\nu,q}(b)}h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{Y+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{Y-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&-\frac{\hat{\tau}_{Y,\nu,q}(b)}{\hat{\tau}^2_{T,\nu,q}(b)}h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{T+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{T-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&-\frac{1}{\tau_{T,\nu}}h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{Y+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{Y-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&+\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{T+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{T-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
=&(\frac{1}{\hat{\tau}_{T,\nu,q}(b)}-\frac{1}{\tau_{T,\nu}})h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{Y+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{Y-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
&-(\frac{\hat{\tau}_{Y,\nu,q}(b)}{\hat{\tau}^2_{T,\nu,q}(b)}-\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}})h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{T+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{T-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!\\
=&h^{1+p-\nu}O_p(\frac{1}{\sqrt{nb^{1+2\nu}}}+b^{1+q-\nu})O_p(1+\frac{1}{\sqrt{nb^{3+2p}}}).
\end{aligned}$$

The second equality holds because $\mu^{*(1+p)}_{Z\bullet}=\hat{\mu}^{(1+p)}_{Z\bullet,q}(b)$ and $\tau^*_{Z,\nu}=\hat{\tau}_{Z,\nu,q}(b)$ almost surely because the bootstrap DGP is obtained by fitting a local polynomials of order $q$. The last equality holds because of similar argument in CCT's Theorem A.2. Asymptotic normality of $\hat{\zeta}_{\nu,p}(h)-\Delta^*_{\nu,p,q}(h,b)-\zeta_\nu$ then follows from normality of $\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b)-\tau_{Y,\nu}$, $\hat{\tau}^{bc}_{T,\nu,p,q}(h,b)-\tau_{T,\nu}$ (CCT's Theorem 1) and the fact that remaining terms $R_n$, $R^*_n$, $R^{*bc}_n$ and $O_p(h^{2+p-\nu})$ are negligible.

CCT have shown that $V^{bc}(h,b) = O_p(\frac{1}{nh^{1+2\nu}} + \frac{h^{2(1+p-\nu)}}{nb^{3+2p}})$ (Lemma SA.4) and $R_n^2 = o_p(V^{bc}(h,b))$ (Theorem A.2). In addition, because $O_p(h^{2+p-\nu}) = o_p(R_n^{*bc})$, it suffices to show that

$$\begin{aligned}
\frac{R_n^{*bc2}}{V^{bc}(h,b)} =& O_p\big(\min\{nh^{1+\nu}, \frac{nb^{3+2p}}{h^{2(1+p-\nu)}}\}\big)h^{2(1+p-\nu)}O_p\big(\frac{1}{nb^{1+2\nu}} + b^{2(1+q-\nu)}\big)O_p\big(1+\frac{1}{nb^{3+2p}}\big)\\
=& O_p\big(\min\{nh^{3+2p}, nb^{3+2p}\}\big)O_p\big(\frac{1}{nb^{1+2\nu}} + b^{2(1+q-\nu)}\big)O_p\big(1+\frac{1}{nb^{3+2p}}\big)\\
=& O_p\big(b^{2+2(p-\nu)}\min\{(\frac{h}{b})^{3+2p},1\} + nb^{2(1+q-\nu)}\min\{nh^{3+2p}, nb^{3+2p}\}\big)O_p\big(1+\frac{1}{nb^{3+2p}}\big)\\
=& O_p\big(b^{2+2(p-\nu)}\min\{(\frac{h}{b})^{3+2p},1\} + nb^{2(q-p)}b^{2(1+p-\nu)}\min\{nh^{3+2p}, nb^{3+2p}\}\big)\\
&+ O_p\big(\frac{1}{nb^{1+2\nu}}\min\{(\frac{h}{b})^{3+2p},1\} + b^{2(1+q-\nu)}\min\{(\frac{h}{b})^{3+2p},1\}\big)\\
=& o_p(1),
\end{aligned}$$

provided that $n\min\{h^{3+2p}, b^{3+2p}\}\max\{h^2, b^{2(q-p)}\} \to 0$ and $n\min\{h, b^{1+2\nu}\} \to \infty$. □.

### B.0.2 Proof of Theorem 2.2

Repeat the steps from Theorem 2.1's proof for the iterated bootstrap to get

$$\hat{\zeta}^*_{\nu,p}(h) - \Delta^{**}_{\nu,p,q}(h,b) - \zeta^*_\nu = \frac{1}{\tau^*_{T,\nu}}(\hat{\tau}^{*bc}_{Y,\nu,p,q}(h,b) - \tau^*_{Y,\nu}) - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\hat{\tau}^{*bc}_{T,\nu,p,q}(h,b) - \tau^*_{T,\nu}) + R^*_n - R^{**}_n - R^{**bc}_n + O_p(h^{2+p-\nu}),$$

As is proved in previous section, the higher order terms do not contribute to its asymptotic variance and can be ignored. It will be firstly shown that the variance of $\frac{1}{\tau^*_{T,\nu}}(\hat{\tau}^{*bc}_{Y,\nu,p,q}(h,b) - \tau^*_{Y,\nu}) - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\hat{\tau}^{*bc}_{T,\nu,p,q}(h,b) - \tau^*_{T,\nu})$ converges to that of $\frac{1}{\tau_{T,\nu}}(\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b) - \tau_{Y,\nu}) - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}^{bc}_{T,\nu,p,q}(h,b) - \tau_{T,\nu})$, then its asymptotic normality will be proved.

**Proof for variance convergence in probability.** Rewrite bias-corrected estimator for $Z$:

$$\begin{aligned}
\hat{\tau}^{bc}_{Z,\nu,p,q}(h,b) - \tau_{Z,\nu} =& (\hat{\tau}_{Z,\nu,p}(h) - \mathbb{E}\,\hat{\tau}_{Z,\nu,p}(h)) + (\mathbb{E}\,\hat{\tau}_{Z,\nu,p}(h) - \tau_{Z,\nu}) - (\mathbb{E}^*\,\hat{\tau}^*_{Z,\nu,p}(h) - \tau^*_{Z,\nu}) \\
=& \hat{\tau}_{Z,\nu,p}(h) - \mathbb{E}\,\hat{\tau}_{Z,\nu,p}(h) \\
&+ h^{1+p-\nu}(\hat{\mu}^{(q)}_{Z-,q}(b) - \mu^{(q)}_{Z-})\mathfrak{B}_{-,\nu,p,p+1}(h)/(1+p)! \\
&- h^{1+p-\nu}(\hat{\mu}^{(q)}_{Z+,q}(b) - \mu^{(q)}_{Z+})\mathfrak{B}_{+,\nu,p,p+1}(h)/(1+p)! \\
&+ O_p(h^{2+p-\nu}) \\
=& \nu! e_\nu' \Gamma_{+,p}(h)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_p(X_i/h)K_{+,h}(X_i)\varepsilon_{Zi}\Big) \\
&- \nu! e_\nu' \Gamma_{-,p}(h)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_p(X_i/h)K_{-,h}(X_i)\varepsilon_{Zi}\Big) \\
&+ \frac{q! e_q' h^{1+p-\nu}}{(1+p)!b^q}\Gamma_{-,q}(b)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_q(X_i/b)K_{-,b}(X_i)\varepsilon_{Zi}\Big)\mathfrak{B}_{-,\nu,p,p+1}(h) \\
&- \frac{q! e_q' h^{1+p-\nu}}{(1+p)!b^q}\Gamma_{+,q}(b)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_q(X_i/b)K_{+,b}(X_i)\varepsilon_{Zi}\Big)\mathfrak{B}_{+,\nu,p,p+1}(h) \\
&+ O_p(h^{2+p-\nu}) \\
=& \sum_{i=1}^n W(X_i)\varepsilon_{Zi} + O_p(h^{2+p-\nu})
\end{aligned}$$

with

$$\begin{aligned}
W(X_i) &= W_+(X_i) - W_-(X_i) \\
W_\bullet(X_i) &= \tfrac{1}{n}\nu! e_\nu' \Gamma_{\bullet,p}(h)^{-1} r_p(X_i/h)K_{\bullet,h}(X_i) - \tfrac{1}{n}\frac{q! e_q' h^{1+p-\nu}}{(1+p)!b^q}\Gamma_{\bullet,q}(b)^{-1} r_q(X_i/b)K_{\bullet,b}(X_i).
\end{aligned}$$

With this simplified notation, we have

$$\frac{1}{\tau_{T,\nu}}(\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b) - \tau_{Y,\nu}) - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}^{bc}_{T,\nu,p,q}(h,b) - \tau_{T,\nu}) = \sum_{i=1}^n W(X_i)(\frac{1}{\tau_{T,\nu}}\varepsilon_{Yi} - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}\varepsilon_{Ti}) + O_p(h^{2+p-\nu}),$$

which has variance

$$\mathbb{V}\,\big(\sum_{i=1}^n W(X_i)(\frac{1}{\tau_{T,\nu}}\varepsilon_{Yi} - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}\varepsilon_{Ti})\big) = \sum_{i=1}^n W(X_i)^2(\frac{1}{\tau^2_{T,\nu}}\sigma^2_{Yi} + \frac{\tau^2_{Y,\nu}}{\tau^4_{T,\nu}}\sigma^2_{Ti} - \frac{2\tau_{Y,\nu}}{\tau^3_{T,\nu}}\sigma_{Yi,Ti}).$$

Apply similar steps to the iterated bootstrap, we have

$$\frac{1}{\tau^*_{T,\nu}}(\hat{\tau}^{*bc}_{Y,\nu,p,q}(h,b) - \tau^*_{Y,\nu}) - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\hat{\tau}^{*bc}_{T,\nu,p,q}(h,b) - \tau^*_{T,\nu}) = \sum_{i=1}^n W(X_i)(\frac{1}{\tau^*_{T,\nu}}\varepsilon^*_{Yi} - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}\varepsilon^*_{Ti}),$$

which, by the construction of wild bootstrap, has variance

$$\mathbb{V}^*\big(\sum_{i=1}^n W(X_i)(\frac{1}{\tau^*_{T,\nu}}\varepsilon^*_{Yi}-\frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}\varepsilon^*_{Ti})\big)=\sum_{i=1}^n W(X_i)^2(\frac{1}{\tau^{*2}_{T,\nu}}\hat{\varepsilon}^2_{Yi}+\frac{\tau^{*2}_{Y,\nu}}{\tau^{*4}_{T,\nu}}\hat{\varepsilon}^2_{Ti}-\frac{2\tau^*_{Y,\nu}}{\tau^{*3}_{T,\nu}}\hat{\varepsilon}_{Yi}\hat{\varepsilon}_{Ti}).$$

By the standard argument on the convergence of residuals to the population error, it is ensured that $\sum_{i=1}^n W(X_i)^2\hat{\varepsilon}^2_{Yi} \to^p \sum_{i=1}^n W(X_i)^2\sigma^2_{Yi}$, $\sum_{i=1}^n W(X_i)^2\hat{\varepsilon}^2_{Ti} \to^p \sum_{i=1}^n W(X_i)^2\sigma^2_{Ti}$ and $\sum_{i=1}^n W(X_i)^2\hat{\varepsilon}_{Yi}\hat{\varepsilon}_{Ti} \to^p \sum_{i=1}^n W(X_i)^2\sigma_{Yi,T_i}$. Combined with the fact that $\tau^*_{Z,\nu}=\hat{\tau}_{Z,q}(b) \to^p \tau_Z$, the proof for convergence of variance is complete.

**Proof for asymptotic normality.** Conditional on the regressors and residuals, $\{W(X_i)(\frac{1}{\tau^*_T}\hat{\varepsilon}_{Yi}-\frac{\tau^*_Y}{\tau^{*2}_T}\hat{\varepsilon}_{Ti})e^*_i\}$ is a sequence of independent and mean zero random variables. In addition, it consists of four parts based on the definition of $W(X_i)$. It can be shown that each part is asymptotically normal by Lindeberg-Feller CLT. The proof below is an example showing that the first part $\frac{1}{n}\nu!e'_\nu\Gamma_{\bullet,p}(h)^{-1}r_p(X_i/h)K_{\bullet,h}(X_i)(\frac{1}{\tau^*_T}\hat{\varepsilon}_{Yi}-\frac{\tau^*_Y}{\tau^{*2}_T}\hat{\varepsilon}_{Ti})e^*_i$ is asymptotically normal.

The Liapunov's condition requires that

$$\frac{1}{s_n^{2+\delta}}\sum_{i=1}^n \mathbb{E}|H_i(Xi)|^{2+\delta} \to^p 0$$

with

$$H_i(X_i)=\tfrac{1}{n}\nu!e'_\nu\Gamma_{\bullet,p}(h)^{-1}r_p(X_i/h)K_{\bullet,h}(X_i)(\frac{1}{\tau^*_T}\hat{\varepsilon}_{Yi}-\frac{\tau^*_Y}{\tau^{*2}_T}\hat{\varepsilon}_{Ti})e^*_i; \quad s_n^2=\sum_{i=1}^n \mathbb{V}(H_i).$$

Based on CCT's Lemma SA.1, we know that

$$\sum_{i=1}^n \mathbb{E}|H_i(Xi)|^{2+\delta}=O_p(\frac{1}{(nh)^{1+\delta}}),$$
$$s_n^2=O_p(\frac{1}{nh}),$$

which verifies the Liapunov's condition given that $nh \to \infty$. Similar arguments can be applied to other three parts. □.

# APPENDIX C. ADDITIONAL MATERIAL FOR CHAPTER 3

*Proof of Proposition 3.1.* Under the minimum probabilistic requirements specified by equations (3.3b) - (3.3d), the vector $((\mathbf{P}^W)^T, (\mathbf{P}^E)^T)^T$ has a convex geometry because it is defined by a set of linear restrictions. Its geometry is closed because the linear restrictions are in the form of "$\geq$", "$=$" and "$\leq$". As a result, the geometry of the subvector $\mathbf{P}^E$ is also closed and convex. □

*Proof of proposition 3.2.* The expected outcome conditional on the treatment $Z = z$ is

$$\begin{aligned}\mathbb{E}[Y|Z=z] &= \sum_y y P_{Y|Z}(y|z) = \sum_y y \frac{P_{Y,Z}(y,z)}{P_Z(z)} \\ &= \sum_y y \frac{\sum_f \mathbb{1}[f(z)=y] P_F(f) P_{Z|F}(z|f)}{P_Z(z)} \\ &= \sum_y \sum_f y\, \mathbb{P}[Y(z)=y, F=f] \frac{P_{Z|F}(z|f)}{P_Z(z)}.\end{aligned}$$

The expected outcome if treatment $z$ is received is

$$\mathbb{E}[Y(z)] = \sum_y y\, \mathbb{P}[Y(z)=y] = \sum_y \sum_f y\, \mathbb{P}[Y(z)=y, F=f].$$

So $\mathbb{E}[Y|Z=z]$ is a weighted average of $Y(z)$, with the weights being proportional to $\frac{P_{Z|F}(z|f)}{P_Z(z)}$. □

*Proof of Proposition 3.3.* Similar to Proposition 3.1, one can show that $H^p[\mathbf{P}^{FWE}]$ is convex. If $H^i[\mathbf{P}^{FWE}]$ is convex, then their intersection $H[\mathbf{P}^{FWE}]$ will also be convex. The mapping through linear transformation matrix $\boldsymbol{B}$ preserves the convexity. □