Essays in Causal Inference

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Claudia Luise Charlotte Noack

Frühjahrs-/Sommersemester 2021

Abteilungssprecher	Prof. Volker Nocke, Ph.D.
Referent	Prof. Christoph Rothe, Ph.D.
Koreferent	Prof. Yoshiyasu Rai, Ph.D.

Tag der Verteidigung	20. Juli 2021

Acknowledgments

First and foremost, I would like to express my deep gratitude to my advisor Christoph Rothe for his outstanding guidance and invaluable support throughout the last four years. Through providing me with the great opportunity to work together, he taught me the passion and dedication to deeply analyze and discuss research questions in econometrics. By creating an encouraging and intellectually stimulating research environment, he enabled me to ask any questions, which he always answered patiently. His honest and constructive comments helped me to develop and improve not only this thesis but also my abilities as a researcher more generally. I am more than grateful for all this.

I would like to thank Carsten Trenkler for giving me helpful feedback about my research and for always being very supportive in all other aspects related to research. I am very grateful to Yoshiyasu Rai for his invaluable comments and suggestions about my paper and seminar presentations. I am also extremely thankful to Tim Armstrong. I have benefited a lot from his comments and insightful discussions with him, which helped me to see things in a different light.

I am grateful to my fellow student Tomasz for going together through all the marvelous and tough moments of the Ph.D. program. I enjoyed working on the third chapter with him as I learned that not only life but also research is much more delightful when you experience moments of desperation and optimism not alone but together with a friend.

I want to thank all former and current members of the Mannheim econometrics group with whom I enjoyed everyday lunch, including great discussions about econometrics, research, and life in general; in particular, I thank Alex, Catherine, Giovanni, and Jonas. Jonas took especially care of my well-being and regularly provided me with delicious food.

This thesis is built on inspirations for both research in economics in general but also solutions to particular problems, which I gained from many more people. I am grateful to all of them: Professor Jochen Streb brought me to the idea and encouraged me to pursue a Ph.D. in economics. I benefited greatly from presentations of and discussions about my research at seminars and conferences, specifically from numerous comments and suggestions, which I have received from the econometrics group at Yale.

For facilitating the administrative part of both my Ph.D. and especially my stays abroad, I am thankful to the administration of the CDSE and the Economics Department

of the University of Mannheim. This work was further supported by the University of Mannheim's Graduate School of Economic and Social Sciences, the European Research Council (ERC) through grant SH1-77202 with Christoph Rothe as principal investigator, and the state of Baden-Württemberg through computational resources (bwHPC).

Further, I would like to thank all my fellow CDSE students for making my Ph.D. studies a wonderful experience. I am especially grateful to my fellow students in Mannheim and abroad, who became my friends. In particular, I thank Matthias for being always available for intensive discussions and refreshing distractions from work; Nils for accompanying me throughout my entire time in Mannheim; Frauke for her constant support and for continuously challenging me to broaden my views and opinions; Elisa for inspiring me to find my way; Jasmin for constantly encouraging me; Sylvia for becoming like my own sister; and numerous other friends I am thinking of.

Last but not least, I would like to thank my entire family. I am thankful to my godparents Almuth, Ebi, Ulrich, and their families for always supporting me. I am grateful to my brother Andreas for listening to all my doubts and complaints and taking them away. I am indebted to my parents for their tireless and unconditional support throughout my entire life. Thank you so much! Finally, I am deeply grateful to have Philipp and his family by my side. Philipp has constantly encouraged me to find and to go my own way. Thank you, Philipp, for always putting a smile on my face, and thank you above all for simply being the person you are.

Contents

Acknowledgments **i**

List of Figures **vii**

List of Tables **ix**

Preface **1**

1 Sensitivity Analysis of the Monotonicity Assumption in the LATE framework **5**

1.1 Introduction 5
1.2 Setup 9
1.2.1 Model of the Local Average Treatment Effect 9
1.2.2 Illustration of the Sensitivity Analysis 10
1.3 Sensitivity Parameters 13
1.4 Partial Identification of Distribution Functions 14
1.4.1 Preliminaries 14
1.4.2 Preliminary Bounds 15
1.4.3 Identification Result 17
1.5 Sensitivity Analysis 18
1.5.1 Sensitivity Region 18
1.5.2 Robust Region 20
1.6 Extensions 21
1.6.1 Treatment Effects for other Populations 21
1.6.2 Binary Outcome Variable 23
1.7 Estimation and Inference 24
1.7.1 Estimation 24
1.7.2 Goal of Inference 25
1.7.3 Inference for a Continuous Outcome Variable 26
1.7.4 Inference for a Binary Outcome Variable 27

1.8 Simulations . . . 28
1.8.1 Setup . . . 28
1.8.2 Simulation Results . . . 28
1.9 Empirical Application . . . 29
1.9.1 Sensitivity Analysis for Binary Outcome Variable . . . 29
1.9.2 Sensitivity Analysis for Continuous Outcome Variable . . . 31
1.10 Conclusion . . . 32

Appendices **33**
1.A Additional Materials for the Sensitivity Analysis . . . 33
1.B Proofs of Main Results . . . 37
1.C Additional Materials for Estimation and Inference . . . 47
1.D Proofs of Additional Results . . . 53
1.E Further Illustrations . . . 56

2 Bias-Aware Inference in Fuzzy Regression Discontinuity Designs **59**
2.1 Introduction . . . 59
2.2 Setup and Preliminaries . . . 62
2.2.1 Fuzzy RD Designs . . . 62
2.2.2 Honest Confidence Sets . . . 63
2.2.3 Smoothness Conditions . . . 63
2.2.4 Discrete Settings . . . 64
2.2.5 Local Linear Estimation . . . 65
2.3 Existing Methods for RD Inference . . . 65
2.3.1 SRD Inference . . . 65
2.3.2 Delta Method FRD Inference . . . 66
2.4 Bias-Aware Fuzzy RD Confidence Sets . . . 68
2.5 Theoretical Properties . . . 70
2.5.1 Assumptions . . . 70
2.5.2 Honesty . . . 71
2.5.3 Shape . . . 71
2.5.4 Comparison with Bias-Aware Delta Method CIs . . . 72
2.6 Implementation Details . . . 73
2.6.1 Standard Errors . . . 73
2.6.2 Bandwidth Choice . . . 74
2.6.3 Computation . . . 75
2.6.4 Choosing Smoothness Bounds . . . 76

2.7 Simulations . . . 79
2.7.1 Setup . . . 79
2.7.2 Simulations Results . . . 80
2.8 Empirical Application . . . 81
2.9 Conclusions . . . 86

Appendices **86**
2.A Proofs of Main Results . . . 89
2.B More general Bandwidth Choices . . . 102
2.C Properties of Rule-of-Thumb Smoothness Bounds . . . 103
2.D Extension to Fuzzy Regression Kink Designs . . . 106
2.E Additional Materials for the Empirical Application . . . 111
2.F Additional Materials for the Simulations . . . 112

3 Flexible Covariates Adjustments in Regression Discontinuity Designs 121
3.1 Introduction . . . 121
3.2 Setup . . . 124
3.2.1 Model and Parameter of Interest . . . 124
3.2.2 Standard RD Estimator . . . 125
3.3 Covariate adjustments . . . 125
3.3.1 Covariate-Adjusted Outcome Variable . . . 125
3.3.2 Optimal Adjustment Function . . . 126
3.3.3 Estimator . . . 127
3.4 Theoretical results . . . 128
3.4.1 Assumptions . . . 128
3.4.2 Main Asymptotic Results . . . 130
3.4.3 Standard Error . . . 132
3.5 Implementation Details . . . 133
3.5.1 Bandwidth Choice . . . 133
3.5.2 Confidence Intervals . . . 133
3.5.3 Different Bandwidths . . . 134
3.6 Examples of Covariate Adjustments . . . 135
3.6.1 Linear Adjustments . . . 136
3.6.2 Non-linear Parametric Adjustments . . . 136
3.6.3 Nonparametric Adjustments . . . 137
3.6.4 Adjustments Based on Machine Learning Methods . . . 138
3.7 Simulations . . . 139

3.7.1 Setup 139
3.7.2 Simulations Results 140
3.8 Conclusions 142

Appendices **145**
3.A Further Sufficient Conditions for Main Assumptions 145
3.B Relation to the Literature 146
3.C Proofs of Main Results 148
3.D Variance Calculations 156
3.E Additional Simulation Results 156

References **161**

Curriculum Vitæ **171**

Eidesstattliche Erklärung **173**

List of Figures

1.1 Illustration of sensitivity and robust region I. 12
1.2 Illustration of sensitivity and robust region II. 21
1.3 Application - Confidence sets for the sensitivity and robust region I. . . . 30
1.4 Application - Confidence Sets for the Sensitivity and Robust Region II . . . 31
A.1.1 Derivation of the compliers outcome distributions 57
A.1.2 Illustration of sensitivity region. 58

2.1 Examples of elements of $\mathcal{F}_H(B_Y)$ for various values of B_Y 78
2.2 Simulations - Conditional expected outcome 79
2.3 Simulations - Simulated coverage rates 83
2.4 Application - CEF of outcome an treatment by birth cohort 87
A.2.1 Properties of rule-of-thumb smoothness bounds 104
A.2.2 Application - Fits of polynomial specifications underlying ROT1 and ROT2. 113
A.2.3 Application - Examples of elements of $\mathcal{F}_H(B_Y)$ based on the full sample I. 114
A.2.4 Application - Examples of elements of $\mathcal{F}_H(B_Y)$ based on full sample II. . 115
A.2.5 Application - Examples of elements of $\mathcal{F}_H(B_Y)$ based on restricted sample 116
A.2.6 Application - Examples of elements of $\mathcal{F}_H(B_T)$ based on full sample . . . 117
A.2.7 Application - Examples of elements of $\mathcal{F}_H(B_T)$ based on restricted sample 118

3.1 Normalized difference of RD estimates with local linear adjustments. . . . 143
3.2 Normalized difference of RD estimates with post-lasso regression adjustments. 144

List of Tables

1.1 Simulated coverage rates of the sensitivity and robust region for a positive treatment effect. 28

2.1 Simulated coverage rates for various types of confidence sets 82
2.2 Application - Results for Oreopoulos (2006). 85
A.2.1 Simulations - Further results for simulated coverage rates 119

3.1 Simulation Results 140
A.3.1 Full simulation results for different numbers of relevant covariates 158
A.3.2 Simulation results for different signal-to-noise ratios 159

Preface

Knowledge-based economic policy decisions often rely on empirical conclusions drawn from observational studies about the causal effect of a treatment on a set of outcomes. If units can self-select into treatment, a simple comparison of the outcomes of treated and untreated units does not identify any causal effect as both groups can be inherently different. Several empirical strategies to identify causal effects have been developed for specific setups and are widely applied in empirical economic research today. While being theoretical in nature, this thesis aims to provide practical and easy-to-implement tools to improve the reliability and accuracy of empirical estimates in these methods. It consists of three self-contained chapters. Chapter 1 considers the local average treatment effect framework, Chapters 2 and 3 the regression discontinuity design.

In the canonical setting of the local average treatment effect (LATE), units are incentivized to take up a treatment by a randomly assigned instrument; for instance, a group of people is randomly assigned to participate in a job training program or a health treatment. Under additional assumptions, the treatment effect of those units whose treatment status is indeed affected by the instrument, compliers, is identified (Imbens and Angrist, 1994). One of the key additional assumptions, however, is monotonicity restricting the effect of the instrument on the treatment status to be monotone across all units. In many empirical economics applications, the validity of this assumption might be questionable.

In Chapter 1, I develop a method to assess the sensitivity of LATE estimates to potential violations of the monotonicity assumption. I parameterize the degree to which monotonicity is violated using two sensitivity parameters: the first one determines the share of defiers in the population, and the second one measures differences in the distributions of outcomes between compliers and defiers. I derive sharp bounds on the compliers' outcome distributions in the first-order stochastic dominance sense for each value of these two sensitivity parameters. I identify the robust region, which is the set of all values of sensitivity parameters for which a given empirical conclusion, e.g., the LATE is positive, is valid. Researchers can assess the credibility of their conclusion by verifying that all plausible sensitivity parameters lie in the robust region so that their estimates gain credibility.

In regression discontinuity designs (RDs), treatment is assigned if a specific covariate, the running variable, exceeds a known cutoff value; for instance, unemployment benefits, access to credits, or a health treatment can be based on an administrative or health score exceeding a specific cutoff value. Under mild assumptions, units that are close to the cutoff are as good as randomly assigned so that the causal effect for units at the margin of being assigned is identified in a credible and transparent way. In sharp RD designs, units fully comply with their treatment assignment, whereas in fuzzy designs, units may only partially comply (Hahn et al., 2001).

The statistical challenge in RD designs is that units that are directly at the cutoff are in general not observed, so that their expected outcome has to be inferred from units that are close to the cutoff. Imposing strong functional form assumptions between the expected outcome and the running variable, e.g., a linear or polynomial relation, would identify the expected outcome of units at the cutoff. However, conclusions drawn on these estimates might be misleading if the imposed functional form does not accurately approximate the true conditional expectation function. Therefore, empirical economic research often relies on nonparametric methods, that only impose that the relation of outcome and running variable is sufficiently smooth in a neighborhood of the cutoff; for instance,, by assuming that the second derivative of the true function is bounded. In particular, local linear regression methods are used, where a linear regression is fitted only to observations that are in a small neighborhood of the cutoff. The choice of the size of the neighborhood is key here. A smaller neighborhood implies that the linear function approximates the regression function more accurately so that the potential smoothing bias is reduced; a larger neighborhood implies that more observations are used so that the variance of the estimator is reduced. This thesis considers two econometrics issues in RD designs: the construction of confidence sets in fuzzy designs that adequately take the smoothing bias into account; and the use of additional pretreatment covariates to efficiently reduce the variance of RD estimators.

Chapter 2 is joint work with Christoph Rothe. We argue that confidence sets for the parameter of interest in fuzzy RD designs that are commonly applied in the literature generally fail to be valid under a wide range of empirically relevant conditions such as setups with discrete running variables, donut designs, and weak identification. We propose new confidence sets that are bias-aware in the sense that they take possible smoothing bias explicitly into account. Their construction shares similarities with that of Anderson-Rubin confidence sets in exactly identified instrumental variable models and thereby avoids issues with "delta method" approximations that underlie most commonly used existing inference methods for fuzzy RD analysis. Our confidence sets compare

favorably in terms of both theoretical and practical performance to existing procedures in canonical settings with strong identification and a continuous running variable.

Chapter 3 is joint work with Tomasz Olma and Christoph Rothe. We propose a novel class of covariate-adjusted RD estimators that can have a smaller variance than estimators used in the literature. Our procedure accommodates a wide range of covariate adjustments under mild conditions. We consider classic parametric and nonparametric, as well as machine learning methods so that suitable estimators can be chosen for any given type of covariates. We allow for discrete and continuous covariates in low- and high-dimensional settings. The proposed estimators are easily applicable because the tuning parameters can be selected, and confidence intervals can be constructed following standard methods used in the literature. We characterize the covariate adjustments that lead to the smallest variance in this class of RD estimators.

Chapter 1

Sensitivity Analysis of the Monotonicity Assumption in the LATE Framework

1.1. INTRODUCTION

The local average treatment effect framework (LATE), introduced in Imbens and Angrist (1994), is one of the most popular econometric frameworks for instrumental variable analysis in setups of heterogeneous treatment effects. We consider settings of a binary instrumental variable and a binary treatment variable. The Wald estimand then equals the treatment effect of *compliers*, individuals for which the instrument influences the treatment status, given the well-known classical LATE assumptions: monotonicity, independence, and relevance.

Monotonicity states that the effect of the instrument on the treatment decision is monotone across all units. In the canonical example, in which the instrument encourages units to take up the treatment, monotonicity rules out the existence of *defiers*, i.e., units that receive the treatment only if the instrument discourages them. Researchers might question the validity of this assumption in empirical applications. In these settings, the local treatment effect estimates might be biased and might lead the researchers to draw incorrect conclusions about the true treatment effect.

As an example of a setup in which monotonicity could plausibly be violated, consider the study of Angrist and Evans (1998), who analyze the effect of having a third child on the labor market outcomes of mothers. As the decision to have a third child is endogenous, the authors use a dummy for whether the first two children are of the same sex as an instrument. The underlying reasoning is that some parents would only decide to have a third child if their first two children were of the same sex; these parents are compliers. The monotonicity assumption seems questionable in this setting as parents, who have a strong preference for one specific sex, might act as a defier in this setup. Consider, for example parents who want to have at least two boys and their first child is a boy. Contrary to the incentivization through instrument, they have two children if their second child is a boy, and three children if their second child is a girl. As the monotonicity assumption might be questionable in this example, one can question the validity of empirical conclusions

drawn from the classical LATE analysis.[1]

In this chapter, we provide a framework to evaluate the sensitivity of treatment effect estimates to a potential violation of the monotonicity assumption. As noted in Angrist et al. (1996), a violation of the monotonicity assumption always has two dimensions: The first dimension is the heterogeneous effect of the instrumental variable on the treatment variable, the presence of defiers. The second dimension is the heterogeneous effect of the treatment variable on the outcome variable, the outcome heterogeneity between defiers and compliers. We derive the extent to which monotonicity is violated by parameterizing these two dimensions.

We parameterize the existence of defiers by their population size and the outcome heterogeneity by the Kolmogorov-Smirnov norm, which bounds the difference of the cumulative distribution functions of compliers and defiers. For each of these two sensitivity parameters, we identify sharp bounds of the outcome distribution of compliers in a first-order stochastic dominance sense. These bounds also imply sharp bounds on various treatment effects, e.g., the average treatment effect or quantile treatment effects of compliers.

Our analysis precedes in tow steps. In a first step, we identify the *sensitivity region*. The sensitivity region defines the set of sensitivity parameters for which a data generating process exists, that is consistent with our model assumptions and implies both the observed probabilities and the sensitivity parameters. Since sensitivity parameters lying in the complement of the sensitivity region are not compatible with our model, we do not analyze them further. For the derivation of the sensitivity region, we also derive sharp bounds of the population size of defiers.

In a second step, we identify the *robust region*, which is the set of sensitivity parameters that imply treatment effects that are consistent with a particular empirical conclusion; for instance, the treatment effect of compliers has a specific sign or a particular order of magnitude.[2] Parameters lying in the complement of the robust region, the *nonrobust region*, imply treatment effects that are not, or may not be, consistent with the given empirical conclusion. The robust region and the nonrobust region are separated from each other by the *breakdown frontier*, following the terminology of Masten and Poirier (2020). For each population size of defiers, the breakdown frontier identifies the weakest assumption about outcome heterogeneity, which is necessary to be imposed to imply treatment effects being consistent with the particular empirical conclusion under consideration.

[1]The other LATE assumptions seem to be plausible here. As the sex of a child is determined by nature and as only the number of and not the sex of the child arguably influences the labor market outcome of mothers, the independence assumption seems to be satisfied. The relevance assumption is testable.

[2]See Masten and Poirier (2020) for a detailed exposition of this approach.

This approach is useful in the following aspects. First, by evaluating the size of the sensitivity region, one can determine the plausibility of the model. If this set is empty, the model is refuted, which implies that even if one would allow for an arbitrary violation of the monotonicity assumption, the independence assumption has to be violated. Second, researchers can analyze the sensitivity of their estimates with respect to the degree to which the monotonicity assumption is violated by varying the sensitivity parameters within the sensitivity region. Third, by evaluating the plausibility of the parameters within the robust region, researchers can assess the sign or the order of magnitude of the treatment effect. While being transparent about the imposed assumptions, they might still arrive at a particular empirical conclusion of interest in a credible way. Fourth, one can assess to which degree monotonicity has to be violated to overturn a particular empirical conclusion. Within our framework, researchers can use their economic insights about the analyzed situation to judge the severity of a violation monotonicity.

While the main focus of this chapter lies on the treatment effects of compliers, we also show how this framework can be exploited to analyze treatment effects of defiers. Under further support assumptions of the outcome variable, treatment effects of even the entire population are partially identified, which complements known results in the literature (see Kitagawa, 2021; Balke and Pearl, 1997; Machado et al., 2019). As the explicit expressions of the sensitivity and robust regions are rather complicated and difficult to interpret, we also provide simplified analytical expressions of these regions in the case of a binary outcome.

To construct confidence sets for both the sensitivity and the robust region, we show that both regions are determined through mappings of some underlying parameters. These mappings are not Hadamard-differentiable, and inference methods relying on standard Delta-method arguments are therefore not applicable. We show how to construct smooth mappings that bound the parameters of interest. This construction leads to mappings for which standard Delta-method arguments are applicable, and we use the nonparametric bootstrap to construct valid confidence sets for the parameters of interest. With a binary outcome variable, the mappings resulting in the sensitivity and robust region are considerably simpler. Therefore, we can use a generalized Delta-method to show asymptotic distributional results and apply a bootstrap procedure to construct asymptotically valid confidence sets.

We show in a Monte Carlo study that our proposed inference method has good finite sample properties. We further apply our method to the setup studied by Angrist and Evans (1998) introduced above. We show that relatively strong assumptions on either the population size of the defiers or the outcome heterogeneity have to be imposed to

preserve the sign of the estimated treatment effect. This result demonstrates that the monotonicity assumption is key in the local treatment effect framework.

The remainder of this chapter is structured as follows: A literature review follows, and Section 1.2 illustrates the setup in a simplified setting. Section 1.3 introduces the sensitivity parameters and Section 1.4 derives sharp bounds on the distribution functions of compliers. The main sensitivity analysis is presented in Section 1.5. Section 1.6 also discusses extensions and Section 1.7 derives estimation and inference results. Section 1.8 contains a simulation study and Section 1.9 an empirical example. Section 1.10 concludes. All proofs and additional materials are deferred to the appendix.

Literature. This chapter relates to several strands of the literature. First, this chapter contributes to the growing strand of the literature, which considers sensitivity analysis in various applications. These applications include, among many others, violations of parametric assumptions, violations of moment conditions, and multiple examples within the treatment effect literature (see, among others, Armstrong and Kolesár, 2021b; Mukhin, 2018; Christensen and Connault, 2019; Kitamura et al., 2013; Bonhomme and Weidner, 2018, 2019; Andrews et al., 2017, 2020a; Andrews and Shapiro, 2020; Andrews et al., 2020b; Roth and Rambachan, 2019; Conley et al., 2012; Imbens, 2003; Chen et al., 2011). This paper is very closely related to Masten and Poirier (2020, 2021), who generalize ideas of breakdown points developed in Horowitz and Manski (1995); Imbens (2003); Kline and Santos (2013); Stoye (2005, 2010). These papers consider several assumptions in the treatment effect literature, but not the monotonicity assumption.

Second, it is related to the local average treatment effect framework literature, which is formally introduced in Imbens and Angrist (1994) and further in Vytlacil (2002). Several papers consider violations of the monotonicity assumption through different types of assumptions. Balke and Pearl (1997); Machado et al. (2019); Huber et al. (2017); Manski (1990); Huber and Mellace (2015); Huber (2015) consider a binary and Kitagawa (2021) a continuous outcome variable and partially identify the average treatment effect. Small et al. (2017); Manski and Pepper (2000); Dahl et al. (2017); Huber et al. (2017) propose alternative assumptions on the data generating process, which are strictly weaker than monotonicity and obtain bounds on various treatment effects.

De Chaisemartin (2017) shows that in the presence of defiers, under certain assumptions, the Wald estimand still identifies a convex combination of causal treatment effects of only a subpopulation of compliers. In a policy context, the treatment effect of compliers might be of particular interest because the treatment status of compliers is most likely to change with a small policy change. However, the same reasoning does not apply to the subpopulation of compliers. Klein (2010) evaluates the sensitivity of the treatment

effect of compliers to random departures from monotonicity. Fiorini et al. (2014) give examples of analyzing the sensitivity of the monotonicity, and Huber (2014) considers a violation of monotonicity in a specific example. They do not provide sharp identification results of the treatment effect of compliers in the presence of defiers, nor do they derive the robust region. A violation of the monotonicity assumption with a non-binary instrumental variable is considered, and alternative assumptions and testing procedures are proposed in Mogstad et al. (2019); Frandsen et al. (2019); Norris et al. (2020). This chapter contributes to this literature by presenting an effective tool to analyze the severity of a potential violation of the monotonicity assumption. It thus gives applied researchers a new tool to evaluate the robustness of their estimates to a violation of the monotonicity assumption, and their estimates may thereby gain credibility.

Our proposed inference procedure builds on seminal work about Delta-methods for non-differentiable mappings by Shapiro (1991); Fang and Santos (2018); Dümbgen (1993); Hong and Li (2018), and it further exploits ideas of smoothing population parameters by Masten and Poirier (2020); Chernozhukov et al. (2010); Haile and Tamer (2003).

1.2. SETUP

1.2.1. **Model of the Local Average Treatment Effect.** We observe the distribution of the random variables (Y, D, Z), where Y is the outcome of interest; D is the actual treatment status, with $D = 1$ if the person is treated and $D = 0$ otherwise; and Z is the instrument, with $Z = 1$ if the person is assigned to treatment and $Z = 0$ otherwise. We assume that each unit has potential outcomes Y_0 in the absence and Y_1 in the presence of treatment, and potential treatment status D_1 when assigned to treatment and D_0 when not assigned to treatment. The observed and potential outcomes are related by $Y = DY_1 + (1-D)Y_0$, and observed and potential treatment status by $D = ZD_1 + (1-Z)D_0$.

Based on the effect of the instrument on the treatment status, we distinguish four different groups: compliers that are only treated if they are assigned to treatment (CO); defiers that are only treated if they are not assigned to treatment (DF); always takers that are independently of the instrument always treated (AT), and never takers that are never treated (NT). We denote the population sizes of the respective group by π_{AT}, π_{NT}, π_{CO}, and π_{DF}. We denote by Y_d^T the potential outcome variable of group $T \in \{AT, NT, CO, DF\}$ under treatment status d. To simplify the notation, we write Y_d^{dT} for the potential outcome variable of always takers if $d = 1$ and otherwise of never takers, and similarly π_{dT} for the respective population size. We denote the outcome distribution

of a variable Y by F_Y, its density function, if it exists, by f_Y, and its support by $\mathbb{Y}$.[3]

The key parameters of interest in this analysis are treatment effects of compliers. We denote the average treatment effect of compliers by[4]

$$\Delta_{CO} = \mathbb{E}[Y_1 - Y_0 | D_0 = 0,\ D_1 = 1].$$

Throughout the chapter, we assume that $\mathbb{P}(D = 1|Z = 1) \geq \mathbb{P}(D = 1|Z = 0)$ without loss of generality, and we impose the following identifying assumptions.

Assumption 1.1. *The instrument satisfies* $(Y_1, Y_0, D_1, D_0) \perp Z$ *(Independence), and* $\mathbb{P}(D = 1|Z = 1) > \mathbb{P}(D = 1|Z = 0)$ *(Relevance).*

We refer to Angrist et al. (1996) for an extensive discussion of these assumptions.

1.2.2. **Illustration of the Sensitivity Analysis.** In this section, we illustrate the sensitivity analysis in a very simplified framework, where we introduce the sensitivity parameters, the sensitivity, and the robust region. We do not consider any sharp identification results in this illustration, but we do in our main sensitivity analysis in Section 1.3-1.5.

1.2.2.1. *Sensitivity Parameter Space.* In the presence of defiers, the average treatment effect of compliers is not point identified. Angrist et al. (1996) show that the Wald estimand, $\beta^{IV} = \text{Cov}(Y, Z)/\text{Cov}(D, Z)$, equals a weighted difference of the average treatment effect of compliers and defiers:

$$\beta^{IV} = \frac{1}{\pi_{\text{CO}} - \pi_{\text{DF}}} \left(\pi_{\text{CO}} \Delta_{CO} - \pi_{\text{DF}} \Delta_{DF}\right). \tag{1.1}$$

Clearly, if either $\pi_{\text{DF}} = 0$, implying the absence of defiers, or $\Delta_{CO} = \Delta_{DF}$, implying that compliers and defiers have the same average treatment effect, the treatment effect Δ_{CO} is still point identified. In general, however, three parameters of Equation (1.1) are not identified: the population size of defiers π_{DF}, the treatment effect of compliers Δ_{CO} and of defiers Δ_{DF}. To bound the average treatment effect of compliers, we introduce two sensitivity parameters. The first one determines the population size of defiers, and the second one outcome heterogeneity between compliers and defiers. These two parameters measure the degree to which monotonicity is violated and represent the two dimensions of heterogeneity: (i) heterogeneous effects of the instrument on the treatment status and (ii) heterogeneous effects of the treatment on the outcome.

The heterogeneous impact of the instrument on the treatment status, is parameterized,

[3]Throughout the chapter, we implicitly assume that all necessary moments of all random variables for the parameter of interest exist; for instance, if we consider the local average treatment effect, we assume Y_d^T has first moments for all $d \in \{0, 1\}$ and $T \in \{C, DF, AT, NT\}$.

[4]Similarly, the average treatment effect of defiers is denoted by $\Delta_{DF} = \mathbb{E}[Y_1 - Y_0 | D_0 = 1,\ D_1 = 0]$.

in the most simplest ways, by the population size of defiers

$$\pi_{\text{DF}} = \mathbb{P}(D_0 = 1 \text{ and } D_1 = 0). \tag{1.2}$$

A larger sensitivity parameter π_{DF} implies a more severe violation of monotonicity. It is clear that, for a given population size of defiers, π_{DF}, the population sizes of the other groups are point identified. In our analysis, these population sizes are, therefore, functions of the sensitivity parameter π_{DF}, but we leave this dependence implicit.[5]

We parameterize the second dimension of heterogeneity by the sensitivity parameter δ_a which equals the absolute differences in treatment effects of both groups

$$\delta_a = |\Delta_{CO} - \Delta_{DF}|.$$

A larger sensitivity parameter δ_a implies a more severe violation of monotonicity.

1.2.2.2. *Sensitivity Region and Robust Region.* The *sensitivity region* is the set of sensitivity parameters which do not violate our model assumptions. For instance, a sensitivity parameter $\pi_{\text{DF}} \geq 0.5$ would violate our model assumptions as the relevance assumption implies that $\pi_{\text{CO}} > \pi_{\text{DF}}$. Therefore, such a sensitivity parameter does not lie within our sensitivity region, which is identified without imposing any additional assumptions. In this illustrative example, we simplify the derivation and say that the sensitivity region is trivially given by

$$\text{SR}_a = [0, 0.5) \times \mathbb{R}_+.$$

In our main sensitivity analysis, this set, however, is nontrivial and can even be empty. In this case, the model is rejected, implying that even though the monotonicity assumption may be violated, the independence assumption has to be violated as well.

Even though the treatment effect of compliers is generally not identified if $\pi_{\text{DF}} > 0$, using (1.1), it is partially identified for any given pair of sensitivity parameters $(\pi_{\text{DF}}, \delta_a)$ by

$$\Delta_{CO} \in \left[\beta^{IV} - \frac{\pi_{\text{DF}}}{\pi_{\text{CO}} - \pi_{\text{DF}}}\delta_a,\ \beta^{IV} + \frac{\pi_{\text{DF}}}{\pi_{\text{CO}} - \pi_{\text{DF}}}\delta_a\right].$$

In a typical sensitivity analysis, researchers now consider different values of the sensitivity parameters to evaluate the identified sets of the parameter of interest and to evaluate the robustness of the LATE estimates to a potential violation of monotonicity. However, in many empirical applications, the interest does not lie in the precise treatment effect but in its sign or in its order of magnitude. It is, therefore, natural to start with the empirical conclusion of interest and to ask which sensitivity parameters imply treatment effects

[5]It follows from the definitions of the groups and our assumptions that $\pi_{\text{AT}} = \mathbb{P}(D = 1|Z = 0) - \pi_{\text{DF}}$, $\pi_{\text{NT}} = \mathbb{P}(D = 0|Z = 1) - \pi_{\text{DF}}$ and $\pi_{\text{CO}} = \mathbb{P}(D = 1|Z = 1) - \mathbb{P}(D = 1|Z = 0) + \pi_{\text{DF}}$.

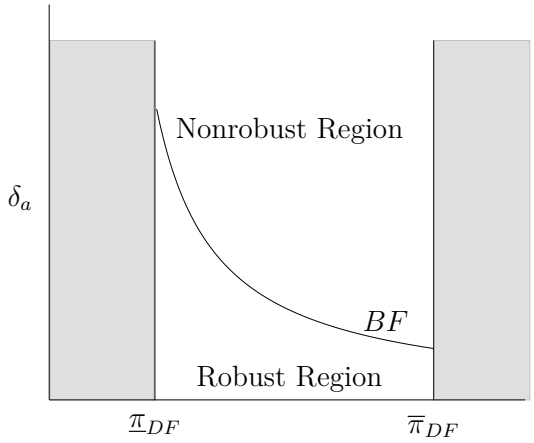


Figure 1.1: Illustration of Sensitivity and Robust Region. Non-shaded area represents sensitivity region. $[\underline{\pi}_{DF}, \overline{\pi}_{DF}]$ represent some bounds on the population size of defiers.

that are consistent with this conclusion. This approach is formalized by the breakdown frontier (see, e.g., Kline and Santos, 2013; Masten and Poirier, 2020).

We now consider the empirical conclusion that $\Delta CO \geq \mu$, and we assume that $\beta^{IV} \geq \mu$.[6] Under our model assumptions and for a given value of the population size of defiers π_{DF}, the breakdown point determines the largest value of outcome heterogeneity δ_a that implies treatment effect that are consistent with our empirical conclusion of interest. Specifically, for any $\pi_{\text{DF}} \in [0, 0.5]$, the breakdown point is given by

$$BP_a(\pi_{\text{DF}}) = \frac{\pi_{\text{CO}} - \pi_{\text{DF}}}{\pi_{\text{DF}}}(\beta^{IV} - \mu).$$

The breakdown frontier (BF) is the set of all breakdown points and the robust region (RR) is the set of all sensitivity parameters that are consistent with the empirical conclusion of interest. They are respectively given by

$$BF_a = \{(\pi_{\text{DF}}, BP_a(\pi_{\text{DF}})) \in \text{SR}_a\} \quad \text{and} \quad RR_a = \{(\pi_{\text{DF}}, \delta_a) \in \text{SR}_a : \delta_a \leq BP_a(\pi_{\text{DF}})\}.$$

The nonrobust region is the complement of the robust region within the sensitivity region. It contains sensitivity parameters that may or may not be consistent with the empirical conclusion. Due to the functional form of the breakdown frontier, the nonrobust region is a convex set in this example. An illustrative example of this setup is shown in Figure 1.1.

In this simple example, neither the sensitivity region nor the robust regions are sharp. For example, if the outcome is binary than the difference between compliers and defiers treatment effects is bounded by $[-2, 2]$ and not by $\mathbb{R}_+$. Similarly, the robust region might

[6] If $\beta^{IV} \leq \mu$, the robust region for the conclusion that $\Delta CO \geq \mu$ is empty.

also be substantially reduced by taking into account the actually observed outcomes.[7] This reasoning means that even though a parameter pair may lie within the sensitivity region, it might not imply a well-defined data generating process that is consistent with the model assumptions and the observed probabilities. Similarly, even though a parameter pair may lie within the nonrobust region, it might be robust. Empirical conclusions that can be drawn from this analysis might, therefore, not be very informative. Consequently, we improve upon this framework in the remainder of this chapter.

1.3. SENSITIVITY PARAMETERS

Since treatment effects of compliers are generally not identified in the presence of defiers, we introduce two sensitivity parameters in this section that are interpretable and imply bounds on the outcome distributions of compliers so that the parameter of interest is partially identified. They allow us to consider a trade-off between the strength of the imposed assumption and the size of the identified set.

To derive the sensitivity parameters, we consider the following function :

$$G_d(y) = \frac{\text{Cov}(\mathbf{1}\{Y \leq y\}, \mathbf{1}\{D = d\})}{\text{Cov}(Z, \mathbf{1}\{D = d\})},$$

for $d \in \{0, 1\}$. In the absence of defiers, $G_d(y)$ is the cumulative distribution function of compliers under treatment status d. In the presence of defiers, it holds analogously to the Wald estimand (1.1) that

$$G_d(y) = \frac{1}{\pi_{\text{CO}} - \pi_{\text{DF}}} \left(\pi_{\text{CO}} F_{Y_d^{CO}}(y) - \pi_{\text{DF}} F_{Y_d^{DF}}(y) \right). \tag{1.3}$$

The outcome distributions of compliers are thus identified up to the population size of defiers and the heterogeneity between the outcome distributions of compliers and defiers.

We introduce two sensitivity parameters to parameterize these two dimensions. First, the presence of defiers is parameterized by the population size of defiers π_{DF} (1.2). Second, outcome heterogeneity is represented by δ, which bounds the maximal difference between cumulative distribution functions of the outcome of compliers and defiers by the Kolmogorov-Smirnov (KS) norm

$$\max_{d \in \{0,1\}} \sup_{y \in \mathbb{Y}} \{|F_{Y_d^{CO}}(y) - F_{Y_d^{DF}}(y)|\} = \delta,$$

where $\delta \in [0, 1]$. Without a restriction on δ, the outcome distributions can be arbitrarily

[7] To give a more concrete example, assume that all treated units have a realized outcome of 1 and all nontreated units have a realized outcome of 0. Then it is clear, that the treatment effect of compliers is point identified to be one.

different. If $\delta = 0$, the outcome distributions are restricted the most as both distribution functions coincide. We say that a larger value of the parameter δ implies a more severe violation of monotonicity.

There are clearly many different possibilities for how heterogeneity between distribution functions can be specified. In this chapter, we choose the Kolmogorov-Smirnov norm, as it leads to tractable analytical solutions of the bounds on the compliers outcome distribution. More importantly, this parameterization is simple enough to be interpretable in an empirical conclusion. A similar parameterization is chosen in Kline and Santos (2013) in a different context.[8]

1.4. PARTIAL IDENTIFICATION OF DISTRIBUTION FUNCTIONS

Since our main sensitivity analysis exploits bounds on parameters defined by the distribution function $F_{Y_d^{CO}}$ for $d \in \{0, 1\}$, we bound this distribution function for a fixed given sensitivity parameter pair $(\pi_{\text{DF}}, \delta)$ in this section. We illustrate the derivation of the bounds in the subsequent sections, and the main result is stated in Section 1.4.3.

1.4.1. **Preliminaries.**

1.4.1.1. *Identification Strategy.* Our goal is to obtain sharp lower and upper bounds of the distribution function $F_{Y_d^{CO}}$ in a first-order stochastic dominance sense. That is, we derive analytical characterizations of the distribution functions $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ that are feasible candidates for $F_{Y_d^{CO}}$, in the sense that they are compatible with the imposed sensitivity parameters, our assumptions, and the population distributions of observable probabilities. They are further such that $\underline{F}_{Y_d^{CO}}(y) \leq F_{Y_d^{CO}}(y) \leq \overline{F}_{Y_d^{CO}}(y)$, for all $y \in \mathbb{Y}$. The identification strategy for deriving such sharp bounds $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ is based on the premise that any candidate distribution function of $F_{Y_d^{CO}}$ then also implies distribution functions of $F_{Y_d^{dT}}$ and $F_{Y_d^{DF}}$. Our candidate function $F_{Y_d^{CO}}$ is therefore feasible, only if the implied functions of $F_{Y_d^{dT}}$ and $F_{Y_d^{DF}}$ are indeed distribution functions.

The explicit analytical characterization of these sharp bounds illustrates the effect of the sensitivity parameters on the bounds, and more importantly, it implies sharp bounds on a variety of treatment effects of interest, e.g., the average treatment effect of compliers (Stoye, 2010, Lemma 1).[9]

[8] Since the parameterization of δ is weak on the tails of the distributions, the bounds on the tails are likely to be uninformative. Imposing a *weighted* KS assumption, that penalizes deviations at the tails of the two distributions more, would overcome this issue but would also lead to less tractable results.

[9] The explicit characterization also allows the inference procedure to be based on $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$.

1.4.1.2. *Notation.* We here collect the notation used in the following subsections. Let $d, s \in \{0,1\}$ and $y \in \mathbb{Y}$. Let the differences in population sizes of compliers and defiers be denoted by $\pi_\Delta = \pi_{\text{CO}} - \pi_{\text{DF}}$. Let $Q_{ds}(y) \equiv \mathbb{P}(Y \leq y, D = d | Z = s)$ be the observed joint distribution of Y and D. We further let, for $\mathscr{B}$ denoting the Borel σ-algebra,

$$\widetilde{G}_d^+(y) = \sup_{B \in \mathscr{B}} \{\mathbb{P}(Y \in B, Y \leq y, D = d | Z = d) - \mathbb{P}(Y \in B, Y \leq y, D = d | Z = 1 - d)\}.$$

and $G_d^+ = \frac{1}{\pi_{\text{CO}}} \widetilde{G}_d^+(y)$. Our sensitivity analysis is based on the following observed underlying parameters

$$\theta = \left(Q_{11}, Q_{10}, Q_{01}, Q_{00}, \widetilde{G}_1^+, \widetilde{G}_0^+\right). \tag{1.4}$$

1.4.2. **Preliminary Bounds.** To illustrate the identification argument, we first derive preliminaries bounds on the distribution function $F_{Y_d^{CO}}$, which are not necessarily sharp in general. Based on the law of total probability and our assumptions, the probability function Q_{dd} is a weighted average of the distribution functions $F_{Y_d^{CO}}$ and $F_{Y_d^{dT}}$, specifically $Q_{dd}(y) = \pi_{\text{CO}} F_{Y_d^{CO}}(y) + \pi_{\text{d}} F_{Y_d^{dT}}(y)$. Any feasible distribution function of $F_{Y_d^{CO}}$ has to imply a function $F_{Y_d^{dT}}$ that is a distribution function. Exploiting this argument and using our sensitivity parameter π_{DF}, it follows that

$$\frac{1}{\pi_{\text{CO}}} Q_{dd}(y) \leq F_{Y_d^{CO}}(y) \leq \frac{1}{\pi_{\text{CO}}} \left(Q_{dd}(y) - \pi_{\text{d}}\right). \tag{1.5}$$

These bounds correspond to the extreme scenarios where compliers have the highest or the lowest outcomes compared to always and never takers.Using the same argument for defiers and the definition of $G_d(y)$ in (1.3), it further follows that

$$\frac{\pi_\Delta}{\pi_{\text{CO}}} G_d(y) \leq F_{Y_d^{CO}}(y) \leq \frac{1}{\pi_{\text{CO}}} \left(\pi_\Delta G_d(y) + \pi_{\text{DF}}\right). \tag{1.6}$$

We now consider the second sensitivity parameter δ. Based on the definition of $G_d(y)$ in (1.3), we conclude that any feasible candidate of $F_{Y_d^{CO}}$ also has to satisfy that

$$G_d(y) - \frac{\pi_{\text{DF}}}{\pi_\Delta} \delta \leq F_{Y_d^{CO}}(y) \leq G_d(y) + \frac{\pi_{\text{DF}}}{\pi_\Delta} \delta. \tag{1.7}$$

Since the function G_d is not necessarily increasing in y for all $y \in \mathbb{Y}$, bounds on the distribution function $F_{Y_d^{CO}}$ based on (1.6) and (1.7) have to take this into account. We therefore directly consider bounds on $F_{Y_d^{CO}}$ that employ this information. To be precise, for the lower bound, we consider equation (1.6) and (1.7), where we replace G_d by its smallest, nondecreasing upper envelope; vice versa, for the upper bound, where we replace

G_d by its greatest, nondecreasing lower envelope.[10] Following this reasoning and taking (1.5)-(1.7) into account, the lower bound is given by

$$\underline{H}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) = \max\{0, \frac{1}{\pi_{\text{CO}}}(Q_{\text{dd}}(y) - \pi_d), \ \frac{\pi_\Delta}{\pi_{\text{CO}}} \sup_{\tilde{y} \leq y} G_d(\tilde{y}), \ \sup_{\tilde{y} \leq y} G_d(\tilde{y}) - \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta\}, \tag{1.8}$$

and the upper bound by

$$\overline{H}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) = \min\{1, \frac{1}{\pi_{\text{CO}}}Q_{dd}(y), \frac{\pi_\Delta}{\pi_{\text{CO}}}(\inf_{\tilde{y} \geq y} G_d(\tilde{y}) + \pi_{\text{DF}}), \inf_{\tilde{y} \geq y} G_d(\tilde{y}) + \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta\}. \tag{1.9}$$

Any value outside of these bounds is clearly incompatible with the distribution of (Y, D, Z) and our assumptions. To illustrate the effect of our sensitivity parameters, we consider the width of these bounds for any fixed $y \in \mathbb{Y}$ as a function of $(\pi_{\text{DF}}, \delta)$, that is[11]

$$\overline{H}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) - \underline{H}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta).$$

The width is weakly increasing in the sensitivity parameter δ, which implies that a larger violation of monotonicity leads to a larger identified set. However, the effect of the sensitivity parameter π_{DF} on this width can be both negative and positive depending on the specific underlying parameters θ. For example, we note that $F_{Y_d^{CO}}$ is point identified either if $\pi_{\text{DF}} = 0$ or $\pi_d = 0$, which denotes the absence of always or never takers. Heuristically speaking, the parameter π_{DF}, therefore, trades off the identification power gained from the non-existence of defiers and the non-existence of always or never takers.

The functions $\underline{H}_{Y_d^{CO}}$ and $\overline{H}_{Y_d^{CO}}$ clearly bound $F_{Y_d^{CO}}$ in a first-order stochastic dominance sense. However, since they do not imply that the implied functions of $F_{Y_d^{dT}}$ and $F_{Y_d^{DF}}$ are nondecreasing, they are not necessarily a feasible candidate of $F_{Y_d^{CO}}$. To give an intuition for this result and for the sake of argument, we now assume that all outcome variables are continuously distributed. We consider $\underline{H}_{Y_d^{CO}}$, and we assume that the bound on the outcome heterogeneity δ determines the bound, i.e., $\underline{H}_{Y_d^{CO}}(y) = G_d(y) - \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta$. This bound does not necessarily imply that the always takers have a positive density. Specifically, the density of the lower bound is $g_d(y) = (q_{dd} - q_{d(1-d)}(y))/\pi_\Delta$, whereas to guarantee that the density function $f_{Y_d^{dT}}$ does not take any negative value, any feasible candidate of $f_{Y_d^{CO}}$ has to satisfy that

$$f_{Y_d^{CO}}(y) \leq \frac{q_{dd}(y)}{\pi_{\text{CO}}} \tag{1.10}$$

[10] We give an illustration of this derivation in Appendix 1.E.1.

[11] This comparison is helpful as the qualitative size of the width of the bounds on the distribution functions is related to the width of the identified set of many parameters of interest, e.g., the LATE.

for all $y \in \mathbb{Y}$.[12] A similar restriction as (1.10) can be derived for defiers such that any feasible candidate of the density $f_{Y_d^{CO}}(y)$ has to also satisfy that, for all $y \in \mathbb{Y}$,

$$f_{Y_d^{CO}}(y) \geq \frac{\pi_\Delta}{\pi_{\text{CO}}} \max\{g_d(y), 0\}. \tag{1.11}$$

Based on this argument, we construct our final bounds, $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$. Specifically, the distribution function $\underline{F}_{Y_d^{CO}}$ is dominated by $\underline{H}_{Y_d^{CO}}$ in a first-order stochastic dominance sense, and the distribution function $\overline{F}_{Y_d^{CO}}$ dominates $\overline{H}_{Y_d^{CO}}$ in a first-order stochastic dominance sense, and they both carefully take into account the reasoning of (1.11) and (1.10). In Appendix 1.B.1, we show that these distribution functions both bound the distribution function $F_{Y_d^{CO}}$ and are feasible candidates.

1.4.3. **Identification Result.** We first provide the analytical expressions of the bounds in the following. The lower bound of the distribution functions $F_{Y_d^{CO}}$ is given by

$$\begin{aligned}\underline{F}_{Y_d^{CO}}(y,\pi_{\text{DF}},\delta) = \frac{1}{\pi_{\text{CO}}} Q_{\text{dd}}(y) \\ - \frac{1}{\pi_{\text{CO}}} \inf_{\tilde{y} \geq y} \left(Q_{dd}(\tilde{y}) - \left(\pi_\Delta G_d^+(\tilde{y}) - \inf_{\widehat{y} \leq \tilde{y}} \left(\pi_\Delta G_d^+(\widehat{y}) - \pi_{\text{CO}} \underline{H}_{Y_d^{CO}}(\widehat{y}, \pi_{\text{DF}}, \delta) \right) \right) \right),\end{aligned} \tag{1.12}$$

and similarly the upper bound by

$$\begin{aligned}\overline{F}_{Y_d^{CO}}(y,\pi_{\text{DF}},\delta) = \frac{\pi_\Delta}{\pi_{\text{CO}}} G_d^+(y) \\ - \frac{1}{\pi_{\text{CO}}} \sup_{\tilde{y} \geq y} \left(\pi_\Delta G_d^+(\tilde{y}) - \left(Q_{dd}(\tilde{y}) - \sup_{\widehat{y} \leq \tilde{y}} \left(Q_{dd}(\widehat{y}) - \pi_{\text{CO}} \overline{H}_{Y_d^{CO}}(\widehat{y}, \pi_{\text{DF}}, \delta) \right) \right) \right).\end{aligned} \tag{1.13}$$

Based on the derivation above, Theorem 1.1 summarizes the result.

Theorem 1.1. *Suppose that Assumption 1.1 holds, and the data generating process is compatible with the sensitivity parameters* $(\pi_{\text{DF}}, \delta)$*. Then, it holds that*

$$\underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) \leq F_{Y_d^{CO}}(y) \leq \overline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta),$$

for $d \in \{0, 1\}$ *and for all* $y \in \mathbb{Y}$*. Moreover, there exist DGPs which are consistent with the above assumptions such that the outcome distribution of compliers equals either* $\underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$*,* $\overline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$*, or any convex combination of these bounds.*

Theorem 1.1 shows not only that the proposed bounds are valid but also that without imposing further assumptions, the bounds cannot be tightened in a first-order stochastic

[12] To be precise, one can assume that $q_{d(1-d)}(y) = 0$ and as $\pi_\Delta = \pi_{\text{CO}} - \pi_{\text{DF}} \leq \pi_{\text{CO}}$ the claim follows.

dominance sense.[13]

Remark 1.1. Theorem 1.1 does clearly not imply that all distribution functions that are bounded by the distribution functions $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ are feasible candidates of the distribution function of $F_{Y_d^{CO}}$. The reason for that is that these functions do not necessarily imply nondecreasing distribution functions of the other groups. Since we are not interested in the distributions functions themselves but in parameters defined through the bounds, this result is sufficient to derive sharp bounds on the sensitivity and robust region for empirical conclusions about these parameters.

Remark 1.2. In empirical applications, the parameter of interest is often not only the average treatment effect but also, e.g., quantile and distribution treatment effects. As Theorem 1.1 identifies the entire outcome distributions of compliers, these treatment effects are identified as well and are sharp for many relevant parameters. We present them in Appendix 1.A.2.

Remark 1.3. Researchers also often have access to pre-intervention covariates. In Appendix 1.A.3, we show how these covariates can be exploited to reduce the size of the identified set of the distribution function $F_{Y_d^{CO}}$. These covariates can then be used to tighten the sensitivity and to enlarge the robust regions.

1.5. SENSITIVITY ANALYSIS

We present our main sensitivity analysis in this section.

1.5.1. **Sensitivity Region.** We derive the sensitivity region, which is the set of sensitivity parameter pairs for which a feasible candidate of the distribution function $F_{Y_d^{CO}}$ exists. Sensitivity parameters that ly in the complement of this set refute the model, and we, therefore, do not consider them further.[14]

1.5.1.1. *Population Size of Defiers.* We show that the population size of defiers is partially identified. We denote an upper bound by

$$\overline{\pi}_{DF} = \min\{\mathbb{P}(D = 1|Z = 0),\ \mathbb{P}(D = 0|Z = 1)\}. \tag{1.14}$$

The first element of the minimum represents the sum of the population size of always takers and defiers, whereas the second one of never takers and defiers. The population size of defiers is clearly smaller than both of these quantities.

[13] As the derived bounds are rather complicated, we propose simpler bounds for each of our sensitivity parameters in Appendix 1.A.1. These bounds are possibly conservative. We explain how to evaluate in an empirical setting whether they are close to the sharp bounds derived in this section.

[14] Masten and Poirier (2021) call the complement of the sensitivity region the falsification region.

The lower bound on the population size of defiers is denoted by

$$\underline{\pi}_{DF} = \max_{s\in\{0,1\}} \{\sup_{B\in\mathscr{B}} \{\mathbb{P}(Y \in B, D = s|Z = 1 - s) - \mathbb{P}(Y \in B, D = s|Z = s)\}\}. \quad (1.15)$$

The supremum is taken over the differences in the population sizes of defiers and compliers, which bounds the population size of defiers from below. This lower bound is similar to bounds presented in Kitagawa (2015) and Balke and Pearl (1997). The following proposition shows that these bounds are indeed sharp.[15]

Proposition 1.1. *Suppose Assumption 1.1 holds. Then the population size of defiers* π_{DF} *is sharply bounded by* $[\underline{\pi}_{DF}, \overline{\pi}_{DF}]$.

If the lower bound on population size of defiers is greater than zero, $\underline{\pi}_{DF} > 0$, at least one of the classical LATE assumptions, including monotonicity, is violated (see, e.g., Kitagawa, 2015). However, if the above inequalities contradict, i.e., $\underline{\pi}_{DF} > \overline{\pi}_{DF}$, the sensitivity region is empty. This implies that even if one allows for a violation of monotonicity, our model assumptions must be violated as well.

1.5.1.2. *Outcome Heterogeneity.* We now consider the sensitivity parameter δ. Based on Theorem 1.1, we can bound the sensitivity parameter δ from below and from above for a given value of the sensitivity parameter π_{DF}.

A given pair fo sensitivity parameters $(\pi_{\text{DF}}, \delta)$ is refuted if the implied lower and upper bounds, $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$, intersect, so that there does not exists a feasible candidate of the distribution function $F_{Y_d^{CO}}$ which is compatible with these sensitivity parameters. The domain of the sensitivity parameter δ is bounded from below by

$$\underline{\delta}(\pi_{\text{DF}}) = \min_{d\in\{0,1\}} \inf\{\delta : \inf_y \underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) - \overline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) \geq 0\}. \quad (1.16)$$

The feasible set of the sensitivity parameter δ is further bounded from above. The bounds $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ imply bounds on the distribution function of $F_{Y_d^{DF}}$, where the largest value of the Kolmogorov-Smirnov norm between the distributions of $F_{Y_d^{CO}}$ and $F_{Y_d^{DF}}$ is achieved when $\delta = 1$. It follows that there does not exists a feasible candidate function of $F_{Y_d^{CO}}$ such that the implied outcome heterogeneity parameter exceeds this value. We denote the upper bounds by

$$\overline{\delta}(\pi_{\text{DF}}) = \max_{d\in\{0,1\}} \sup_{y\in\mathbb{Y}} \Big\{ |\overline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, 1) - \overline{F}_{Y_d^{DF}}(y, \pi_{\text{DF}}, 1)|, \\ |\underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, 1) - \underline{F}_{Y_d^{DF}}(y, \pi_{\text{DF}}, 1)| \Big\}. \quad (1.17)$$

[15] Huber et al. (2017) also present bounds on the population size of defiers. They note that their bounds are not sharp.

By the reasoning of Theorem 1.1, these bounds are sharp, and any convex combination of these bounds is feasible as well. It follows that our sensitivity region is given by

$$SR = \{(\pi_{\text{DF}}, \delta) : \pi \in [\underline{\pi}_{DF}, \overline{\pi}_{DF}] \text{ and } \underline{\delta}(\pi_{\text{DF}}) \leq \delta \leq \overline{\delta}(\pi_{\text{DF}})\}. \tag{1.18}$$

1.5.2. **Robust Region.** We now derive the robust region for the empirical conclusion that $\Delta_{CO} \geq \mu$.[16] To simplify the presentation, we assume in the following that the sensitivity region is nonempty and that $\Delta_{CO}(\underline{\pi}_{DF}, \underline{\delta}(\underline{\pi}_{DF})) \geq \mu$.[17]

By first-order stochastic dominance of the distribution functions $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$, we can construct sharp bounds on many treatment effect parameters, that depend on these bounds (see Lemma 1 in Stoye, 2010). Specifically, let

$$\underline{\Delta}_{CO}(\pi_{\text{DF}}, \delta) = \int_{\mathbb{Y}} y \, d\overline{F}_{Y_1^{CO}}(y, \pi_{\text{DF}}, \delta) - \int_{\mathbb{Y}} y \, d\underline{F}_{Y_0^{CO}}(y, \pi_{\text{DF}}, \delta) \tag{1.19}$$

$$\overline{\Delta}_{CO}(\pi_{\text{DF}}, \delta) = \int_{\mathbb{Y}} y \, d\underline{F}_{Y_1^{CO}}(y, \pi_{\text{DF}}, \delta) - \int_{\mathbb{Y}} y \, d\overline{F}_{Y_0^{CO}}(y, \pi_{\text{DF}}, \delta). \tag{1.20}$$

Corollary 1.1. *Suppose that Assumption 1.1 holds, and the data generating process is compatible with the sensitivity parameters* $(\pi_{\text{DF}}, \delta)$*. Then, the average treatment effect of compliers,* Δ_{CO}*, is sharply bounded by* $[\underline{\Delta}_{CO}(\pi_{\text{DF}}, \delta), \overline{\Delta}_{CO}(\pi_{\text{DF}}, \delta)]$*.*

For a given sensitivity parameter π_{CO}, we now consider the breakdown point given by

$$BP(\pi_{\text{DF}}) = \sup\{\delta : (\pi_{\text{DF}}, \delta) \in \text{SR and } \underline{\Delta}_{CO}(\pi_{\text{DF}}, \delta) \geq \mu\}.$$

For a given sensitivity parameter π_{DF}, it identifies the weakest assumption on outcome heterogeneity between compliers and defiers such that the empirical conclusion holds. The breakdown point, as a function of the sensitivity parameter π_{DF}, is not necessarily decreasing in the population size of defiers as the bounds on the outcome distribution of compliers can become tighter if the value of π_{DF} increases (see the discussion in Section 1.4.2). The breakdown frontier of the average treatment effect is the boundary of the robust region and given by the set of all breakdown points

$$BF = \{(\pi_{\text{DF}}, \delta) \in \text{SR} : \delta = BP(\pi_{\text{DF}})\}. \tag{1.21}$$

The robust region of the empirical conclusion that $\Delta_{CO} \geq \mu$ is characterized by

$$RR = \{(\pi_{\text{DF}}, \delta) \in \text{SR} : \delta \leq BP(\pi_{\text{DF}})\}. \tag{1.22}$$

[16] In Appendix 1.A.2, we also consider other treatment effects than the average treatment effect of compliers. Sensitivity and robust regions for empirical conclusions about these parameters can then also be derived based on the reasoning of this section.

[17] If $\Delta_{CO}(\underline{\pi}_{DF}, \underline{\delta}(\underline{\pi}_{DF})) < \mu$, the robust region is empty.

The nonrobust region, that is the complement of the robust region within the sensitivity region, contains pairs of sensitivity parameters which only may not imply treatment effects being consistent with the empirical conclusion. Figure 1.2 illustrates one example of the sensitivity and robust region.[18]

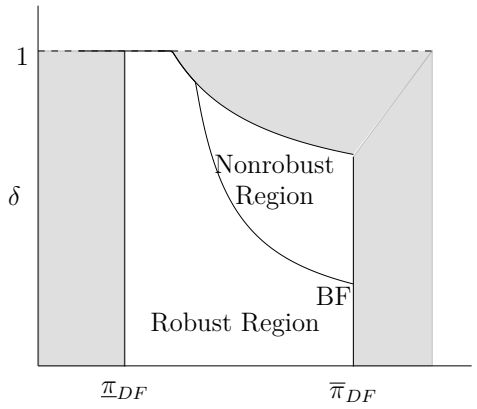


Figure 1.2: Sensitivity and Robust Region. Non-shaded region represents sensitivity region.

1.6. EXTENSIONS

In this section, we show how our framework can be exploited to draw empirical conclusions about other population parameters, and how it simplifies if the outcome variable is binary.

1.6.1. **Treatment Effects for other Populations.** To show how empirical questions about treatment effects of the entire population can be analyzed, we exploit that the proof of Theorem 1.1 presents sharp bounds on all groups in a first-order stochastic dominance sense. For $d \in \{0,1\}$, let the lower bound be denoted by

$$\underline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta) = \pi_{\text{CO}} \cdot \underline{F}_{Y_0^{CO}}(y, \pi_{\text{DF}}, \delta) + Q_{d(1-d)}(y),$$

and the upper bound by

$$\overline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta) = \pi_d + \pi_{\text{CO}} \cdot \overline{F}_{Y_0^{CO}}(y, \pi_{\text{DF}}, \delta) + Q_{d(1-d)}(y).$$

Proposition 1.2. *Suppose the instrument satisfies Assumption 1.1, and the data generating process is compatible with the sensitivity parameters* $(\pi_{\text{DF}}, \delta)$*. Then, it holds that*

$$\underline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta) \leq F_{Y_d}(y, \pi_{\text{DF}}, \delta) \leq \overline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta)$$

for $d \in \{0,1\}$ *and for all* $y \in \mathbb{Y}$*. Moreover, there exist data generating processes which are consistent with the above assumptions such that the potential outcome distributions*

[18] We refer to a discussion on how these sets can be used in an empirical setting to Section 1.2 and 1.9

equal either $\overline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta)$, $\underline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta)$, *or any convex combination of these bounds.*

As the data do not contain any information about the distribution functions $F_{Y_0^{AT}}$ and $F_{Y_1^{NT}}$, the bounds $\underline{F}_{Y_d}$ and $\overline{F}_{Y_d}$ are such that their respective probability mass is shifted to the extreme of the support $\mathbb{Y}$. To interpret these bounds, for any given $y \in \mathbb{Y}$, we consider the difference

$$\overline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta) - \underline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta) = \pi_d.$$

The size of the bounds decreases with the population size of defiers, π_{DF}, as the population size of always and never takers π_{d} decreases with π_{DF}. As if π_{d} decreases, the observed probabilities represent more of the population of interest and correspondingly less of the population mass has to be set to the extreme of the support of the outcome variable. However, the sensitivity parameter δ does not influence the distribution of the outcome of the entire population, as it only influences how the observed outcomes are distributed between the groups.

This reasoning aligns with results of Kitagawa (2021), who showed that imposing the monotonicity assumptions (e.g., $\pi_{\text{DF}} = 0$) does not imply a smaller identified set of the average treatment effect of the entire population. The bounds evaluated at $\pi_{\text{DF}} = \underline{\pi}_{DF}$ are equivalent to the bounds derived in Kitagawa (2021) and for the special case of a binary outcome variable bounds derived in Balke and Pearl (1997); Machado et al. (2019); De Chaisemartin (2017).

Based on the bounds presented in Proposition 1.2, we can now derive a sensitivity analysis similar to the one presented in Section 1.5. However, to derive informative results about the average treatment effect of the entire population, we would have to impose that the outcome is bounded as otherwise the average treatment effect is not identified.

The sensitivity analysis of this chapter is based on the premise that the treatment effect of compliers is the object of interest. However, if the parameter of interest is the treatment effect of the entire population, one might then be willing to impose assumptions not only on outcome heterogeneity between compliers and defiers but also between other groups. To be precise, we can replace the sensitivity parameter δ by δ_p such that

$$\max_d \sup_y \{|F_{Y_d^T}(y) - F_{Y_d^{T'}}(y)|\} \leq \delta_p \quad \forall\, T, T' \in \{AT, NT, CO, DF\},$$

where $\delta_p \in [0, 1]$. Using similar arguments as in the proof of Theorem 1.1, one can then derive sharp bounds on the outcome distribution functions of the entire population and then conduct a sensitivity analysis similar to the one described in Section 1.5. Empirical conclusions drawn on this parameterization might be substantially more informative.

1.6.2. **Binary Outcome Variable.** In many empirical applications, the outcome of interest is binary. The results of Section 1.4 and 1.5 are still valid in this case, but we show in this section that the bounds substantially simplify so that they are easier applicable. Let $P_d^T = \mathbb{P}(Y_d^T = 1)$ denote the probability that the random variable Y_d^T equals one, and let the conditional joint probability of the outcome and the treatment status be given by $P_{ds} = \mathbb{P}(Y = 1, D = d|Z = s)$. We denote the underlying parameters by $\theta_b = (P_{11}, P_{10}, P_{01}, P_{00}, P_0, P_1) \in [0,1]^6$.

Following the same arguments as above, the sensitivity and robust region depend on the marginal outcome distributions of the compliers. The presence of defiers is also bounded by π_{DF}, and the parameter of outcome heterogeneity simplifies to

$$\delta_b = \max_{d\in\{0,1\}} |P_d^{CO} - P_d^{DF}|.$$

The outcome probabilities of compliers are bounded from below by

$$\underline{P}_d^{CO}(\pi_{\text{DF}}, \delta) = \max\left\{0, \frac{P_{dd} - \pi_{\text{d}}}{\pi_{\text{CO}}}, \frac{P_{dd} - P_{d(1-d)}}{\pi_{\text{CO}}}, \frac{P_{dd} - P_{d(1-d)} - \pi_{\text{DF}}\delta_b}{\pi_\Delta}\right\}, \tag{1.23}$$

and from above by

$$\overline{P}_d^{CO}(\pi_{\text{DF}}, \delta) = \min\left\{1, \frac{P_{dd}}{\pi_{\text{CO}}}, \frac{P_{dd} - P_{d(1-d)} + \pi_{\text{DF}}}{\pi_{\text{CO}}}, \frac{P_{dd} - P_{d(1-d)} + \pi_{\text{DF}}\delta_b}{\pi_\Delta}\right\}. \tag{1.24}$$

Corollary 1.2. *Assumption 1.1 holds, and the data generating process is compatible with the sensitivity parameters* $(\pi_{\text{DF}}, \delta)$*. The outcome probabilities of compliers are sharply bounded by* $\underline{P}_d^{CO} \leq P_d^{CO} \leq \overline{P}_d^{CO}$ *and they may attain any value inbetween. Thus, they are sharp.*

The interpretation of the width of these bounds follows the same reasoning as in Section 1.4.2. The lower bound of the population size of defiers simplifies to

$$\underline{\pi}_{DF} = \max_{d\in\{0,1\}} \left\{\sum_{y=0}^{1} \max\{0, \mathbb{P}(Y = y, D = d|Z = 1-d) - P(Y = y, D = d|Z = d)\}\right\}.$$

The upper bound on π_{DF} cannot be simplified further and is given by (1.14). The lower bound on outcome heterogeneity is given by

$$\underline{\delta}_b(\pi_{\text{DF}}) = \frac{\underline{\pi}_{DF}}{\pi_{\text{DF}}}.$$

The lower bound on the sensitivity parameter δ decreases with the population size of defiers. The upper bound on the sensitivity parameter δ is given by the maximal difference

between the outcome probabilities of compliers and defiers

$$\overline{\delta}_b(\pi_{\text{DF}}) = \max_{d\in\{0,1\}} \max\{|\underline{P}_d^{CO}(\pi_{\text{DF}},1) - \underline{P}_d^{DF}(\pi_{\text{DF}},1)|, |\overline{P}_d^{CO}(\pi_{\text{DF}},1) - \overline{P}_d^{DF}(\pi_{\text{DF}},1)|\}.$$

The sensitivity parameter space is given by

$$\text{SR}_b = \{(\pi_{\text{DF}},\delta_b) \in [\underline{\pi}_{DF},\overline{\pi}_{DF}]\times[0,1] : \underline{\delta}_b(\pi_{\text{DF}}) \leq \delta_b \leq \overline{\delta}_b(\pi_{\text{DF}})\},$$

and the robust region for the claim $\Delta_{CO} \geq \mu$ is given by [19]

$$RR_b = \{(\pi_{\text{DF}},\delta_b) \in \text{SR}_b : \underline{P}_1^{CO}(\pi_{\text{DF}},\delta_b) - \overline{P}_0^{CO}(\pi_{\text{DF}},\delta_b) \geq \mu\}.$$

Using the simple algebra structure of the bounds of the outcome probabilities, a closed-form expression for both the robust and the sensitivity region can be derived. As this expression is rather lengthy without providing much intuition, we state it in Appendix 1.B.5.

1.7. ESTIMATION AND INFERENCE

Even though the contribution of this chapter is the derivation of the sensitivity and robust region for a particular empirical conclusion, we consider some methods for estimation and inference of these two regions. While the technical details are deferred to Appendix 1.C, in this section, we sketch the main issues of conducting inference in this setting and our proposed solutions. To simplify the exposition, we consider the setting of a continuous and a binary outcome variable, but our method is not restricted to these distributions.

Throughout this section, we assume that we have access to the data $\{(Y_i^z, D_i^z)\}_{i=1}^{n_z}$ for $z \in \{0,1\}$ that are independent and identically distributed according to the distribution of (Y,D) conditionally on $Z=z$ with support $\mathbb{Y}\times\{0,1\}$. We denote this distribution by (Y^z, D^z) and we let $n = n_0 + n_1$, where n_0/n converges to a nonzero constant as $n \to \infty$.[20]

1.7.1. **Estimation.** To construct estimators of the sensitivity and robust region for a particular empirical conclusion, we note that the identification argument of these regions are constructive. It follows from Section 1.5 that the boundaries of both regions are identified by the following mapping,[21]

$$\phi(\theta,\pi_{\text{DF}}) = (\underline{\pi}_{DF},\ -\overline{\pi}_{DF},\ \underline{\delta}(\pi_{\text{DF}}),\ -\overline{\delta}(\pi_{\text{DF}}),\ BP(\pi_{\text{DF}})), \tag{1.25}$$

which is evaluated at the sensitivity parameter $\pi_{\text{DF}} \in [0, 0.5)$ and the underlying parameters θ, that is defined in (1.4). Estimating the sensitivity and robust region is then

[19] We assume again that $\underline{P}_1^{CO}(\underline{\pi}_{DF},\underline{\delta}_b) - \overline{P}_0^{CO}(\underline{\pi}_{DF},\underline{\delta}_b) \geq \mu$

[20] We discuss this assumption in Assumption A.1.3.

[21] The signs of the components of the mapping $\phi(\theta,\pi_{\text{DF}})$ simplify the subsequent analysis.

equivalent to estimating this mapping. To do so, we consider estimates of the underlying parameters θ that are simply obtained by replacing unknown population quantities by their corresponding nonparametric sample counterparts and by standard nonparametric kernel methods. We denote the estimates of θ by $\widehat{\theta}$. Point estimates of the mapping $\phi(\theta, \pi_{\text{DF}})$ can then be derived by simple plug-in methods. We defer a detailed description to Appendix 1.C.4.

1.7.2. **Goal of Inference.** We propose to construct confidence sets for the sensitivity and robust region such that the confidence set for the sensitivity region is an outer confidence set and for the robust region is an inner confidence set.[22] These confidence sets should therefore jointly satisfy with probability approaching the confidence level, $1-\alpha$, that (i) any sensitivity parameter pair of the sensitivity region lies within the confidence set for the sensitivity region and (ii) not any single parameter pair of the nonrobust region lies within the confidence set for the robust region.[23] Let $\widehat{\text{SR}}_L$ and $\widehat{RR}_L$ denote two sets of the sensitivity parameters. They satisfy the described condition if

$$\lim_{n\to\infty} \mathbb{P}(\text{SR} \subseteq \widehat{\text{SR}}_L \text{ and } \widehat{RR}_L(\text{SR}) \subseteq RR(\text{SR})) \geq 1-\alpha. \tag{1.26}$$

Based on the definition of the mapping $\phi(\theta, \pi_{\text{DF}})$, it therefore suffices to construct a lower confidence band for each component of the estimator $\phi(\widehat{\theta}, \pi_{\text{DF}})$ as a function of π_{DF} that are jointly valid.[24] That is, we need to find a function that is componentwise a uniformly lower bound $\phi_L(\widehat{\theta}, \pi_{\text{DF}})$ in π_{DF} of $\phi(\theta, \pi_{\text{DF}})$ so that[25]

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{1\leq l\leq 5} \inf_{\pi_{\text{DF}}\in[0,0.5)} e_l^\top(\phi_L(\widehat{\theta}, \pi_{\text{DF}}) - \phi(\theta, \pi_{\text{DF}})) \leq 0\right) \geq 1-\alpha, \tag{1.27}$$

where e_l is the l-th unit vector.[26]

[22] Considering inner confidence set for the robust region follows from Masten and Poirier (2020).

[23] To give one more interpretation of the confidence sets and using the language of hypothesis testing, a sensitivity parameter pair, $(\pi_{\text{DF}}, \delta)$, does not lie in the sensitivity region only if we can reject such a hypothesis with confidence level $1-\alpha$. Contrary, $(\pi_{\text{DF}}, \delta)$ lies in the robust region, only if we can reject that it is nonrobust with confidence level $1-\alpha$. The confidence sets are constructed so that the hypothesis tests are valid uniformly in the sensitivity parameter space.

[24] Throughout this section, we consider confidence sets that are uniformly valid in the sensitivity parameter space, but not necessarily in the distribution of the underlying parameters θ

[25] We verify this equivalence in Appendix 1.B.7.3.

[26] Conservative confidence sets for only the average treatment effect of compliers for specific values of $(\pi_{\text{DF}}, \delta)$ directly follow from the presented procedure. To obtain nonconservative confidence sets, one can follow the literature on partially identified parameters (see, e.g., Imbens and Manski, 2004)

1.7.3. **Inference for a Continuous Outcome Variable.** We analyze the distribution of $\phi(\widehat{\theta}, \pi_{\text{DF}})$ in order to construct confidence sets for the mapping $\phi(\theta, \pi_{\text{DF}})$.[27] Under regularity assumptions presented in Appendix 1.C.2, the estimators of the underlying parameters $\widehat{\theta}$ converge in $\sqrt{n}$ to a tight Gaussian process (see Proposition A.1.5 in Appendix 1.D). Since the mapping ϕ is not Hadamard-differentiable, as it depends on minimum, maximum, supremum, and infimum of random functions, standard Delta-method arguments do not apply in this setup (see Fang and Santos, 2018). We propose a method to construct confidence sets that are asymptotically conservative but valid in the sense of (1.26). It is based on ideas of population smoothing that have been suggested by, e.g., Haile and Tamer (2003); Chernozhukov et al. (2010); Masten and Poirier (2020).

In contrast to considering the mapping ϕ, which identifies the sensitivity and robust region, we construct a smooth mapping, ϕ_κ, which yields valid bounds of both regions. The smoothed mapping ϕ_κ is indexed by a fixed smoothing parameter $\kappa \in \mathbb{N}$. The mapping ϕ_κ is differentiable such that the standard functional Delta-method can be applied to ϕ_κ and we can study its asymptotic distribution by standard methods. The mapping ϕ_κ is further such that it yields an outer set of the sensitivity region and an inner set of the robust region. This reasoning implies that confidence sets of the smooth mappings ϕ_κ, which are valid in the sense of (1.26), are also valid for the mapping ϕ.

In finite samples, the choice of the smoothing parameter κ comprises the trade-off of constructing conservative confidence sets and better finite sample approximations of the underlying distributions. Suppose the smoothing parameter κ is small. In that case, the smoothed sensitivity and robust region are very similar to the original regions. However, the finite-sample distribution of $\phi_\kappa(\widehat{\theta}\pi_{\text{DF}})$ might not be well-approximated by its asymptotic distribution. Vice versa, suppose the smoothing parameter κ is large. The finite-sample distribution of $\phi_\kappa(\widehat{\theta}\pi_{\text{DF}})$ might be well-approximated by its asymptotic distribution. However, the smoothed sensitivity and robust region are conservative to the original regions.

In Appendix 1.C.7.1, we show how the smoothed mappings can be constructed. It then follows that plug-in estimators of the smoothed mappings converge in $\sqrt{n}$ to a Gaussian process by standard functional Delta-method arguments. The covariance structure of this process is, in general, rather complicated and tedious to estimate. We, therefore, apply the nonparametric bootstrap to simulate its distribution. Consistency of this bootstrap procedure follows from arguments of Fang and Santos (2018). In Appendix 1.C, we show how to construct the confidence sets based on the described procedure and that they

[27] We want to emphasize that this procedure is valid for a fixed distribution. In particular, we do not consider settings of weak instruments or data generating processes which are such that the robust region becomes empty.

achieve the outlined goal (1.26).

1.7.4. **Inference for a Binary Outcome Variable.** Following the discusssion about a bianry outcome model in Section 1.6.2, the mapping yielding the sensitivity and the robust region for a particular conclusion for a binary outcome variable is given by[28]

$$\phi_b(\theta_b, \pi_{\text{DF}}) = (\underline{\pi}_{\text{DF,b}}, -\overline{\pi}_{\text{DF,b}}, -\overline{\delta}_b(\pi_{\text{DF}}), BP_b(\pi_{\text{DF}})),$$

The interpretation of ϕ_b follows the one for a continuously distributed outcome variable, and in principle, we could apply the same inference procedure as described above. However, the mapping $\phi_b(\theta_b, \pi_{\text{DF}})$ is substantially simpler than the mapping $\phi(\theta, \pi_{\text{DF}})$ so that, in this section, we can apply more classical inference procedure to obtain confidence sets in the sense of (1.26); in particular, we follow ideas of Masten and Poirier (2020) and the literature about moment inequalities (see, e.g., Andrews and Soares, 2010).

Under standard sampling assumptions, it follows that the estimators of the underlying parameters are jointly $\sqrt{n}$ normally distributed (see Appendix 1.C.3). The mapping $\phi_b(\theta_b, \pi_{\text{DF}})$ is clearly not Hadamard-differentiable, as it consist of minimum and maximum of random functions. Standard Delta-method arguments are therefore not applicable here as well. Valid confidence sets could be obtained by projection arguments, which, however, are known to be conservative in general.

We show instead that the mapping ϕ_b is Hadamard directionally differentiable in the direction of θ when evaluated at finitely many $\{\pi^k_{\text{DF}}\}_{k=1}^K$. Using generalized Delta-method arguments, the estimator of the mapping ϕ_b converges to some tight random process, which is a continuous transformation of a Gaussian process, indexed at the finite set $\{\pi^k_{\text{DF}}\}_{k=1}^K$. As this limiting distribution is rather complicated, we do not construct our inference procedure directly on its limiting distribution, but one can choose various modified bootstrap methods to simulate this distribution, e.g., subsampling or numerical-Delta-method (see Dümbgen, 1993; Hong and Li, 2018). In this chapter, we follow a bootstrap method which relies on ideas based on the moment inequality literature (see, e.g., Andrews and Soares, 2010; Bugni, 2010) and we explain the procedure in detail in Appendix 1.C.3. Based on this bootstrap procedure, we can construct valid lower confidence sets for ϕ_b indexed at the finite set of sensitivity parameters $\{\pi^k_{\text{DF}}\}_{k=1}^K$. Using these confidence sets and exploiting the functional form of ϕ_b, we then obtain lower confidence sets for the estimator of the mapping ϕ_b, which are uniformly valid in π_{DF}. We state these arguments precisely in Appendix 1.C.3 and show that these confidence sets are asymptotically valid in the sense of our goal of inference (1.26).

[28] where its precise definition follows from Section 1.6.2 and Appendix 1.C.3.

Table 1.1: Simulated coverage rates of the sensitivity and robust region for a positive treatment effect.

π_{CO}	Δ_{CO}	Δ_{TE}	$\eta = 0.2$	$\eta = 0.5$	$\eta = 1$	$\eta = 1.5$	$\eta = 2$
0.35	0.3	0	99.1	97.8	95.1	91.3	90.8
	0.3	-0.3	96.9	94.3	91.2	92.5	91.6
	0.1	0	99.3	98.5	96.0	92.9	91.2
	0.1	-0.3	99.3	98.5	95.6	93.1	91.3
0.25	0.3	0	98.9	98.1	95.2	92.4	91.2
	0.3	-0.3	99.1	98.0	94.3	92.9	91.4
	0.1	0	99.3	98.6	96.0	93.5	91.2
	0.1	-0.3	99.0	97.7	94.3	92.9	90.9

The data generating process and the expressions follow the description of the text. Results are based on 10,000 Monte Carlo draws.

1.8. SIMULATIONS

1.8.1. **Setup.** We study the finite sample performance of the proposed estimators of the sensitivity and robust regions through a Monte Carlo study. We consider different data generating processes with varying degrees of violations of monotonicity, implying different sizes and shapes of both the sensitivity and robust regions. Specifically, we consider the following population sizes $(\pi_{\text{CO}}, \pi_{\text{DF}}) \in \{(0.35, 0.05), (0.25, 0.15)\}$, where $\pi_{\text{DF}} = \pi_{\text{AT}} = 0.3$. We set $\mathbb{P}(Z = 1) = 0.5$ and we generate the outcome by

$$Y_1^{CO} \sim \mathcal{B}(1, 0.5 + \Delta_{CO}) \quad Y_1^{DF} \sim \mathcal{B}(1, 0.5 + \Delta_{DF}) \quad Y_1^{AT}, Y_0^{NT}, Y_0^{DF}, Y_0^{CO} \sim \mathcal{B}(1, 0.5),$$

where $\Delta_{CO} \in \{0.2, 0.1\}$, and $\mathcal{B}(1, p)$ denotes the Bernoulli distribution with parameter p. The sensitivity region is nonempty as the data generating process satisfies our model assumptions. We consider the empirical conclusion of a positive treatment effect of compliers, so that the robust region is nonempty in each of the data generating processes as the Wald estimand is positive. The bootstrap procedure requires to choose the tuning parameter η, which is explained in Appendix 1.C.3. We consider different values of η given by $\{0.2, 0.5, 1, 1.5, 2\}/\sqrt{N}$. The results are based on 10,000 Monte Carlo draws.

1.8.2. **Simulation Results.** Table 1.1 shows the results of the simulated coverage rates at which the confidence sets cover the population sensitivity and nonrobust region for the different data generating processes and choices of tuning parameters. Our considered choice of tuning parameters implies that the simulated coverage of our confidence sets is close to the nominal one in most data generating processes. These results illustrate that the confidence method performs reasonably well in finite samples.

To illustrate our proposed framework, we apply this sensitivity analysis to data from Angrist and Evans (1998), who analyze the effect of having a third child on the labor market outcomes of mothers. It is shown that even small violations of the monotonicity assumption may have a large impact on the robustness of the estimated treatment effects such that even the sign of the treatment effects may be indeterminate. The same-sex instrument in Angrist and Evans (1998) arguably satisfies Assumption 1.1: The independence assumption seems to be plausible by the following reasoning: The sex of a child is determined by nature, and only the number of and not the sex of the child arguably influences the labor market outcome. The relevance assumption is testable. However, monotonicity might be violated. We apply the proposed sensitivity analysis to evaluate the robustness of the estimated treatment effects to a potential violation of monotonicity in this setting. For simplicity, we focus on two outcome variables: the labor market participation of mothers and their annual wage.[29] The binary decision to treat represents the extensive margin and the continuous outcome variable a mix of extensive and intensive variables. We use the same data as Angrist and Evans (1998).[30] The sample size is 211,983. The point estimated difference of the population sizes of compliers and defiers is given by 0.06.

1.9.1. **Sensitivity Analysis for Binary Outcome Variable .** We consider the labor market participation of mothers as the outcome variable. The Wald estimate is given by -0.13. Figure 1.3 illustrates the 95% confidence set for the sensitivity and the robust region for the claim that the treatment effect of compliers is negative. The formal definition of these confidence sets is given in Section 1.7.2. In this example, a (conservative) 95% confidence set for π_{DF} is given by $[0, 0.37]$. Following the literature, one can therefore not conclude that monotonicity is violated in this example (see for a comparison, e.g., Small et al., 2017). The sensitivity parameter pairs below the red line represent the robust region, which is the estimated set of sensitivity parameters implying a negative treatment effect. This figure shows that concerns about the validity of the monotonicity assumption have to be taken seriously. Since $BP(0.37)$ is almost zero, the hypothesis that the treatment effect is negative cannot be rejected without imposing any assumptions on the data generating process, If the population size of defiers increases, the breakdown frontier is relatively steeply declining, and thus the robust region is rather small. This implies that relatively strong assumptions on the outcome distributions of compliers and defiers have

[29]The annual wage is a continuously distributed running variable with a point mass at zero.

[30]Data are taken from the website Joshua D. Angrist website `www.economics.mit.edu/faculty/angrist` from 1980. The sample is restricted to women at the age of 20-36, having at least two children, being white, and having their first child at the age of 19-25.

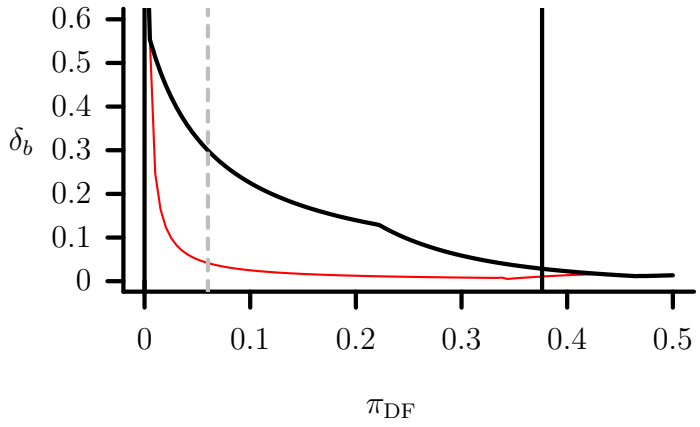


Figure 1.3: Confidence sets for the sensitivity and robust region for a negative treatment effect of compliers. The confidence level is 95 %. The treatment effect of compliers is the effect of having a third child on the labor market participation of mothers complying with the same-sex instrument. The black lines bound the sensitivity region, and the red line indicates the boundary of the robust region. The population size of defiers is on the horizontal axis, and outcome heterogeneity between compliers and defiers on the vertical axis.

to be imposed to conclude that the treatment effect is negative in the presence of defiers. In contrast, if the population size of defiers is small, it is not necessary to impose strong assumptions about heterogeneity in the outcome variables to imply a negative effect.

This example shows that without imposing any assumptions on the data generating process, only non-informative conclusions can be drawn in this example, which is the case as the population size of defiers is not much restricted and is arguably implausible high. [31] One, therefore, might be willing to impose further assumptions to arrive at more interesting results, and we show how one could plausibly proceed. These assumptions should only serve as an example, and obviously, they have to be always adapted to the analyzed situation. We adopt the approach of De Chaisemartin (2017). One of the most essential inherently unknown quantities of interest is the population size of defiers. Imposing a smaller upper bound of this quantity based on economic reasoning allows us to derive sharper results. Based on a survey conducted in the US, De Chaisemartin (2017) states that it seems reasonable that 5% of defiers is a conservative upper bound of the population size of defiers in this setting. If one is willing to impose this assumption,

[31] To interpret these numbers, we note that the upper bound is a rather conservative estimate. If roughly 37% of the population were a defier, then approximately 43% of the population would have been a complier. This reasoning implies that roughly 90% of the population would base their decision to have a third child on the sex composition of the first two children.

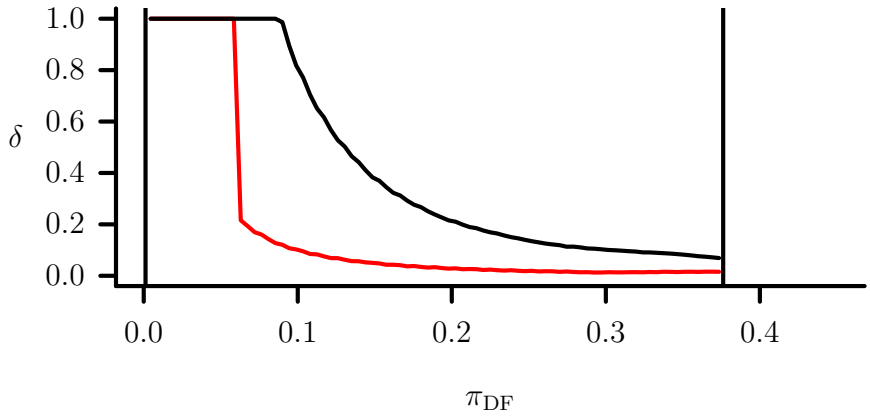


Figure 1.4: Confidence Sets for the Sensitivity and Robust Region for a Negative Treatment Effect of Compliers. The confidence level is 95 %. The compliers treatment effect is the effect of getting a third child on the annual log wage of mothers complying with the same-sex instrument. The black lines bound the sensitivity region, and the red line indicates the boundary of the robust region.

one would still have to assume that the differences in the Kolmogorov-Smirnov norm are less than 0.05, which is a quite strong assumption. Therefore, we would conclude that the treatment effect is not robust to a potential violation in this specific example.

1.9.2. **Sensitivity Analysis for Continuous Outcome Variable.** We now consider the annual log income of the mother. This variable has a point mass at zero, representing all women who do not work but is otherwise continuously distributed. The Wald estimate is given by -1.23. Figure 1.4 shows the corresponding 95% confidence sets for the robust and sensitivity region. If the monotonicity assumption were not violated, this estimate would imply that women who get a third child have an annual log wage reduced by 1.23.

Figure 1.4 shows 95 % confidence sets for both the sensitivity and robust region. The same line of interpretation applies as in the case of a binary outcome variable. One can see that without imposing any assumption about the population size of defiers, the empirical conclusion of a negative treatment effect is not robust to a potential violation of monotonicity. However, applying the same reasoning as above and imposing a maximal population size of 5% as an upper bound of the population size of defiers, one can see that the empirical conclusion is now robust to a potential violation of monotonicity.

To conclude, this sensitivity analysis is of interest, as one can identify the sign and the order of magnitude of the treatment effects by imposing further assumptions. These imposed assumptions are substantially weaker than the monotonicity assumption. The

estimates, therefore, gain credibility.

1.10. CONCLUSION

The local average treatment effect framework is popular to evaluate heterogeneous treatment effects in settings of endogenous treatment decisions and instrumental variables. In some empirical settings, one might doubt the validity of one of its key identifying assumptions, the monotonicity assumption. Conducting a sensitivity analysis of the estimates in these settings improves the reliability of the results. This chapter, therefore, proposes a new framework, which allows researchers to assess the robustness of the treatment effect estimates to a potential violation of monotonicity. It parameterizes a violation of monotonicity by two parameters, the presence of defiers and heterogeneity of defiers and compliers. The former parameter is represented by the population size of defiers and the latter by the Kolmogorov-Smirnov norm bounding the outcome distributions of both groups. Based on these two parameters, we derive sharp identified sets for the average treatment effect of compliers and for any other group under further mild support assumption on the outcome variable. These identification results allow us to identify the sensitivity parameters that imply conclusions of treatment effect being consistent with the empirical conclusion. The empirical example of Angrist and Evans (1998) same-sex instruments underlines the importance of the validity of the monotonicity assumptions as small violations of monotonicity may already lead to uninformative results.

Appendix to Chapter 1

1.A. Additional materials for the sensitivity analysis

In this section, we collect additional materials on identification of the sensitivity region. In Section 1.A.1, we present simplified bounds, and we consider additional treatment effects in Section 1.A.2. We explain how covariates can be used to tighten the bounds in Section 1.A.3 and we give further results on a binary outcome variable in Section 1.A.4.

1.A.1. **Simple Bounds II.** The bounds presented in Theorem 1.1 are rather tedious, and conducting inference for the sensitivity and robust regions based on these bounds is complicated as it depends on many tuning parameters to choose. In this section, we, therefore, present simpler bounds, which might be more easily applicable in an empirical context, and confidence sets of these regions might be more reliable. The proposed simple bounds are especially suited for settings in which the empirical researcher does not have evidence for the existence of defiers and expects that the number of defiers is, if any, smaller than of defiers. In this case, the simplified bounds on the distribution function $F_{Y_d^{CO}}$ are similar to the sharp bounds of Theorem 1.1.

The main reason for the rather complicated expression of the sharp bounds $\overline{F}_{Y_d^{CO}}$ and $\underline{F}_{Y_d^{CO}}$ is heuristically that these bounds exploit all the information contained about the defiers included in the function $G_d(y)$. However, if we forgo the aim of constructing sharp bounds, we can simply look at the functions

$$\underline{F}^S_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) = \max\left\{0,\ \frac{1}{\pi_{\text{CO}}}(Q_{\text{dd}}(y) - \pi_d),\ \frac{\pi_\Delta}{\pi_{\text{CO}}}G_d(y),\ G_d(y) - \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta\right\},$$

and

$$\overline{F}^S_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) = \min\left\{1, \frac{1}{\pi_{\text{CO}}}Q_{dd}(y),\ \frac{\pi_\Delta}{\pi_{\text{CO}}}(G_d(y) + \pi_{\text{DF}}),\ G_d(y) + \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta\right\}.$$

It follows from the reasoning of the main text that these functions satisfy that any value outside of these bounds is incompatible with the distribution of (Y, D, Z), and our model assumptions.[32] They are therefore always valid bounds of the distribution function $F_{Y_d^{CO}}$.

Our simplified sensitivity analysis is then based on the simplified bounds, and we fur-

[32] This reasoning directly follows from the discussion of Section 1.4.

ther ignore the lower bound on the sensitivity parameters $\underline{\pi}_{DF}$ and $\underline{\delta}$ and set both to zero. These bounds are still valid for both the sensitivity and robust region. Estimation and conducting inference within this sensitivity analysis is substantially more straightforward.

In an empirical setting, it remains the question under which conditions these are actually "good" bounds in the sense that they are close to the sharp bounds and do not lead to a substantial loss in information. The most important difference between the sharp and these simple bounds is that we do not exploit the information about the defiers, which is contained in the function G_d for $d \in \{0,1\}$ and the information that the distribution function is not allowed to increase too much. So if G_d is indeed decreasing, we expect that the sharp bounds are substantially more informative than the simple bounds. To assess how conservative these bounds might be, a researcher could also test whether $G_d(y)$ is non-decreasing for all $y \in \mathbb{Y}$ (Kitagawa, 2015).

1.A.2. **Additional Treatment Effects.** The sharp bounds on the distribution function, $F_{Y_d^{CO}}$, in a first-order stochastic dominance sense, allows us to consider various other treatment effects as well. In this section, we consider quantile treatment effects and define the τ-th quantile effect of the compliers by $\Delta_{CO}(\tau)$ and we consider empirical conclusions of the form $\Delta_{CO}(\tau) \geq \mu$.

We define the lower and upper bounds of the quantile functions by the respectively left and right inverse of the bounds of the outcome distributions

$$\underline{Q}_{Y_d^{CO}}(\tau, \pi_{\text{DF}}, \delta) = \inf\{y \in \mathbb{Y} : \overline{F}_{Y_d^{DF}}(y, \pi_{\text{DF}}, \delta) \geq \tau\}$$
$$\overline{Q}_{Y_d^{CO}}(\tau, \pi_{\text{DF}}, \delta) = \sup\{y \in \mathbb{Y} : \underline{F}_{Y_d^{DF}}(y, \pi_{\text{DF}}, \delta) \leq \tau\}.$$

The quantile treatment effect of a quantile τ is then given by

$$\begin{aligned}&[\underline{\Delta}_{CO}(\tau, \pi_{\text{DF}}, \delta), \overline{\Delta}_{CO}(\tau, \pi_{\text{DF}}, \delta)] \\ &\quad = [\underline{Q}_{Y_1^{CO}}(\tau, \pi_{\text{DF}}, \delta) - \overline{Q}_{Y_0^{CO}}(\tau, \pi_{\text{DF}}, \delta), \overline{Q}_{Y_1^{CO}}(\tau, \pi_{\text{DF}}, \delta) - \underline{Q}_{Y_0^{CO}}(\tau, \pi_{\text{DF}}, \delta)].\end{aligned}$$

It follows from the reasoning of Lemma 1 in Stoye (2010) that these bounds are indeed sharp as well, and there exist feasible candidate distribution functions of $F_{Y_d^{CO}}$, which also imply any value between these bounds.

The sensitivity region is defined independently of the particular empirical conclusion under consideration and is therefore given by the expression of the main text (1.18). It follows that the breakdown point for the conclusion that $\Delta_{CO}(\tau) \geq \mu$ is given by

$$BP_\tau(\pi_{\text{DF}}) = \sup\{\delta : (\pi_{\text{DF}}, \delta) \in \text{SR and } \underline{\Delta}_{CO}(\tau, \pi_{\text{DF}}, \delta) \geq \mu\}.$$

The breakdown frontier of the quantile treatment effect is given by

$$BF_\tau = \{(\pi_{\text{DF}}, \delta) \in : \delta = BP_\tau(\pi_{\text{DF}})\}.$$

and the robust region by

$$RR_\tau = \{(\pi_{\text{DF}}, \delta) \in \text{SR} : \delta \leq BP_\tau(\pi_{\text{DF}})\}.$$

1.A.3. **Additional Covariates.** Additional covariates, which are measured prior to treatment assignment, can be used to tighten the bounds on the identified set of treatment effects of compliers and can thus lead to greater sets of the robust region; the arguments are similar to those in Lee (2009). It further holds that conditioning on pretreatment covariates can imply that the identified set of the sensitivity parameters can be reduced such that the analysis becomes more informative. We, therefore, assume that the covariates are discrete and given by $\mathcal{X} = \{x_1, \ldots, x_K\}$, which splits the population into non-overlapping groups. We further impose the following assumptions.

Assumption A.1.2. *(i) Conditional independence assignment:* $(Y_1, Y_0, D) \perp Z|X = x$*, (ii) Conditional relevance:* $\mathbb{P}(D = 1|Z = 1, X = x) > \mathbb{P}(D = 1|Z = 0, X = x)$*, (iii) Common support:* $0 < \mathbb{P}(Z = 1|X = x) < 1$*.*

We construct the sensitivity parameters such that for all $x \in \{x_1, \ldots, x_K\}$

$$\pi_{\text{DF}}(x) \leq \pi_{\text{DF}}.$$

This parameterization implies that the population size of defiers is bounded from above for each value of the covariates. We note that this parameterization implies without further assumptions conservative bounds as long as $\pi_{\text{DF}}(x) \neq \pi_{\text{DF}}$ for some values of x.[33]

By similar reasoning the heterogeneity in the outcome distribution is restricted by

$$|F_{Y_d^{CO}|X=x}(y|X = x) - F_{Y_d^{DF}|X=x}(y|X = x)| \leq \delta.$$

Based on the pre-intervention covariates one can calculate for each k lower and upper bounds on the population size of defiers $\underline{\pi}_{DF}(x_k)$ and $\overline{\pi}_{DF}(x_k)$, respectively. The bounds on the sensitivity parameters can then be calculated based on the definition of the sensitivity parameters by $\underline{\pi}_{DF} = \min_k(\underline{\pi}_{DF}(x_k))$ and $\overline{\pi}_{DF} = \max_k(\overline{\pi}_{DF}(x_k))$. Let

[33]We consider two alternative parameterization: First, one could argue that $\pi_{\text{DF}}(x) = \pi_{\text{DF}}$ for all $x \in \mathcal{X}$. The implied bounds would be sharp, but this assumption is very restrictive. Second, we could consider a setting of $\pi_{\text{DF}}(x) = \pi^x_{\text{DF}}$. In this parameterization, however, the parameter space might be very large and therefore difficult to interpret. The parameterization chosen in the text is plausible and interpretable.

$\pi^k_{\text{DF}} = \max\{\underline{\pi}^k_{DF}, \pi^k_{\text{DF}}\}$ and denote the lower bound by

$$\underline{F}^x_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta) = \frac{1}{\pi_{\text{DF}}} \sum_{k=1}^{K} \mathbb{P}(X = x_k, \pi^k_{\text{DF}})\ \underline{F}^x_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta | X = x_k)$$

and the upper bound by

$$\overline{F}^x_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta) = \frac{1}{\pi_{\text{DF}}} \sum_{k=1}^{K} \mathbb{P}(X = x_k, \pi^k_{\text{DF}})\ \overline{F}^x_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta | X = x_k).$$

Proposition A.1.3. *Suppose that Assumption A.1.2 holds, and the data generating process is compatible with the sensitivity parameters* $(\pi_{\text{DF}}, \delta)$*. Then, for* $d \in \{0, 1\}$

$$\underline{F}_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta) \leq F_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta, \theta_x) \leq \overline{F}_{Y_d}(y, \pi_{\text{DF}}, \delta).$$

Moreover, there exist DGPs which are consistent with the model assumptions such that the outcome distribution of compliers equals either $\overline{F}_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta)$*,* $\underline{F}_{Y^{CO}_d}(y, \pi_{\text{DF}}, \delta)$*, or any convex combination of these bounds, if for all* $x \in \mathcal{X}$ *it holds that* $\pi_{\text{DF}}(x) = \pi_{\text{DF}}$ *and for each* $x \in \mathcal{X}$ *and for* $d \in \{0, 1\}$*, it holds that*

$$\sup_{y \in \mathbb{Y}} |F_{Y^{CO}_d | X=x}(y | X = x) - F_{Y^{DF}_d | X=x}(y | X = x)| = \delta.$$

The derivation of the sensitivity and robust region follows the same line of arguments as in Section 1.5.

1.A.4. **Form of Sensitivity and Robust Region for Binary Outcome Variable.** Since our inference procedure exploits the shape of the sensitivity and robust region for a particular empirical conclusion about a binary outcome variable, we discuss this form in this section. We note again that they are determined by the following parameters.

$$\phi_b(\theta_b, \pi_{\text{DF}}) = (\underline{\pi}_{DF},\ -\overline{\pi}_{DF},\ -\overline{\delta}(\pi_{\text{DF}}),\ BP(\pi_{\text{DF}})),$$

We discuss the components in turn. The sensitivity region is determined based on four parameters: the lower and upper bound on the population size of defiers and the lower and upper bound on the sensitivity parameter of outcome heterogeneity. Due to their simple form, we do not have to discuss the lower and upper bound on the population size of defiers further, neither the lower bound on the outcome heterogeneity. However, the upper bound on the sensitivity parameter of outcome heterogeneity is given by

$$\overline{\delta}_b(\pi_{\text{DF}}) = \max_{d \in \{0,1\}} \max\{|\underline{P}^{CO}_d(\pi_{\text{DF}}, 1) - \underline{P}^{DF}_d(\pi_{\text{DF}}, 1)|, |\overline{P}^{CO}_d(\pi_{\text{DF}}, 1) - \overline{P}^{DF}_d(\pi_{\text{DF}}, 1)|\}.$$

Following the discussion about how the bounds are constructed, e.g., in Appendix 1.E.1, it follows that the upper bound on outcome heterogeneity has to be decreasing in the population size of defiers as the distribution of both compliers and defiers become more similar. We now consider the breakdown point as a function of the population size of defiers. We note that it can be rewritten as

$$BP(\pi_{\text{DF}}) = \frac{1}{\pi_{\text{DF}}} \max\{BP_0(\pi_{\text{DF}}), BP_1(\pi_{\text{DF}}), BP_2(\pi_{\text{DF}})\},$$

where $BP_0(\pi_{\text{DF}})$ is decreasing, $BP_1(\pi_{\text{DF}})$ and $BP_2(\pi_{\text{DF}})$ are potentially increasing so that

$$\begin{aligned}
&BP_0(\pi_{\text{DF}}) \\
&= \max\Bigg\{P_{11} - P_{10} - (\mu + \frac{P_{00} - P_{01} + \pi_{\text{DF}}}{\pi_{\text{CO}}})\pi_\Delta, -((\mu - \frac{P_{11} - P_{10}}{\pi_{\text{CO}}})\pi_\Delta + P_{00} - P_{01}) \\
&\qquad - (\mu \cdot \pi_\Delta + P_{00} - P_{01}), P_{11} - P_{10} - (\mu + 1)\pi_\Delta, \frac{1}{2}(P_{11} - P_{10} - P_{00} + P_{01} - \mu \cdot \pi_\Delta), 0\Bigg\} \\
&BP_1(\pi_{\text{DF}}) = \max\left\{0, -((\mu - \frac{P_{11} - \pi_{\text{AT}}}{\pi_{\text{CO}}})\pi_\Delta + P_{00} - P_{01})\right\} \\
&BP_2(\pi_{\text{DF}}) = \max\left\{0, P_{11} - P_{10} - (\mu + \frac{P_{00}}{\pi_{\text{CO}}})\pi_\Delta\right\}.
\end{aligned}$$

We therefore denote by

$$\tilde{\phi}_b(\theta_b, \pi_{\text{DF}}) = (\underline{\pi}_{DF},\ -\overline{\pi}_{DF},\ -\overline{\delta}(\pi_{\text{DF}}),\ BP_0(\pi_{\text{DF}}), BP_1(\pi_{\text{DF}}), BP_2(\pi_{\text{DF}})). \tag{A.1.28}$$

The mapping $\tilde{\phi}_b$ is either nondecreasing or nonincreasing in each of its component. We exploit this shape to construct confidence sets for the sensitivity and robust region that are uniformly valid in π_{DF}.

1.B. PROOFS OF MAIN RESULTS

In this section, we prove the main results of this paper.

1.B.1. **Proof of Theorem 1.1.** As we consider a fix sensitivity parameter pair $(\pi_{\text{DF}}, \delta)$, we omit the dependence of all functions on the sensitivity parameter in this section, for instance, we write $\underline{F}_{Y_d^{CO}}(y)$ instead of $\underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$.

To determine whether a given distribution function is indeed a feasible candidate for the distribution function $F_{Y_d^{CO}}$, we construct random variables $\widetilde{W} \equiv (\widetilde{Y}_0, \widetilde{Y}_1, \widetilde{D}_0, \widetilde{D}_1, \widetilde{Z})$, which do not only imply the candidate distribution function as outcome distribution of compliers, but also they imply the observed distribution function (Y, D, Z), are consistent with our model assumptions, and the imposed sensitivity parameters.

Based on the instrument's independence assumption, the definition of our groups, and

our sensitivity parameters, this joint distribution is not restricted beyond the population sizes, it suffices to consider the marginal outcome distributions of each group for $d \in \{0,1\}$ [34]

$$F_{\widetilde{Y}_d^{CO}},\ F_{\widetilde{Y}_d^{DF}},\ F_{\widetilde{Y}_d^{AT}},\ F_{\widetilde{Y}_d^{NT}}$$

to construct the joint distribution function $\widetilde{W}$. Since the data are also noninformative about the distribution functions $F_{Y_0^{AT}}$ and $F_{Y_1^{NT}}$ these distributions are left unrestricted as well. As the outcome distributions of the groups are defined independently of the random variable $\widetilde{Z}$, they satisfy by their construction the independence assumption, and the relevance assumption is satisfied by the imposed sensitivity parameters. It therefore follows that $F_{\widetilde{Y}_d^{CO}}$ is a feasible candidate of $F_{Y_d^{CO}}$ if we can construct outcome distributions $F_{\widetilde{Y}_d^{DF}}$ and $F_{\widetilde{Y}_d^{dT}}$ which are compatible with the observed distribution functions and the sensitivity parameters for $d \in \{0,1\}$.

We argue in the main text that any feasible candidate of the distribution function $F_{Y_d^{CO}}$ has to satisfy at least that

$$\underline{H}_{Y_d^{CO}}(y) \leq F_{Y_d^{CO}}(y) \leq \overline{H}_{Y_d^{CO}}(y) \tag{A.1.29}$$

The proof now proceeds in two parts. In part I, we exploit which information can be obtained from the observed probabilities about the compliers outcome distribution to show which additional restriction, besides (A.1.29), any feasible candidate of $F_{Y_d^{CO}}$ has to satisfy. In part II, we then verify that the proposed bounds $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ are feasible candidates of the distribution function $F_{Y_d^{CO}}$. We show that these bounds satisfy that any value outside of these bounds is incompatible with the distribution of (Y, D, Z), and our assumptions. We denote by $\underline{y}$ and $\overline{y}$ the respectively left and right limits of $\mathbb{Y}$, which might equal $\pm\infty$.

Let further $G_d^{\sup}(y) = \sup_{\widehat{y} \leq y} G_d(\widehat{y})$ be the smallest envelope function that is nondecreasing and satisfies $G_d(y) \leq G_d^{\sup}(y)$ for all $y \in \mathbb{R}$; similarly, let $G_d^{\inf}(y) = \inf_{\widehat{y} \geq y} G_d(\widehat{y})$ be the greatest envelope function that is nondecreasing and satisfies $G_d^{\inf}(y) \leq G_d(y)$ for all $y \in \mathbb{R}$.

Part I. Using (1.5) and that distribution functions are nondecreasing, any feasible candidate of $F_{Y_d^{CO}}$ has to satisfy that, for any $y, y' \in \mathbb{Y}$ and $y' \leq y$,

$$F_{Y_d^{CO}}(y) - F_{Y_d^{CO}}(y') \leq \frac{Q_{dd}(y) - Q_{dd}(y')}{\pi_{\text{CO}}}. \tag{A.1.30}$$

[34] In the discussion of the sensitivity region, we obviously analyze the joint distributions.

Using the same reasoning and (1.6), it follows that

$$F_{Y_d^{CO}}(y) - F_{Y_d^{CO}}(y') \geq \frac{\pi_\Delta}{\pi_{\text{CO}}} \left(G_d(y) - G_d(y')\right).$$

for any arbitrary y and y'. As $G_d(y)$ is not necessarily nondecreasing, we can similarly conclude that it has to hold that

$$\mathbb{P}(Y_d^{CO} \in B) \geq \frac{\pi_\Delta}{\pi_{\text{CO}}} \left(\mathbb{P}(Y \in B, D = d | Z = d) - \mathbb{P}(Y \in B, D = d | Z = 1 - d)\right)$$

for any $B \in \mathcal{B}$ and therefore

$$F_{Y_d^{CO}}(y) - F_{Y_d^{CO}}(y') \geq \frac{\pi_\Delta}{\pi_{\text{CO}}} \left(G_d^+(y) - G_d^+(y')\right). \tag{A.1.31}$$

Any feasible candidate of $F_{Y_d^{CO}}$ has to further satisfy the conditions

$$(viii) \ \lim_{y \to \underline{y}} F_{Y_d^{CO}}(y) = 0 \quad \text{and} \quad \lim_{y \to \overline{y}} F_{Y_d^{CO}}(y) = 1. \tag{A.1.32}$$

The distribution functions $F_{Y_d^{DF}}$, and $F_{Y_d^{dT}}$ fulfill then these limit conditions based on (1.5) and (1.6), as it holds that $\lim_{y \to \underline{y}} G_d(y) = \lim_{y \to \underline{y}} Q_{ds}(y) = 0$ and $\lim_{y \to \overline{y}} G_d(y) = 1$ and $\lim_{y \to \overline{y}} Q_{dd}(y) = \pi_d + \pi_{\text{CO}}$ and $\lim_{y \to \overline{y}} Q_{d(1-d)}(y) = \pi_d + \pi_{\text{DF}}$ for any $d, s \in \{0, 1\}$.

Any real-valued function, which is defined on $\mathbb{Y}$ and right-continuous, which left-limits exists and which satisfy equations (A.1.29) – (A.1.32) implies by construction potential outcome distributions for all four groups, which are consistent with the imposed model assumption, the sensitivity parameter constraints, and the observed probability functions. It is thus a feasible candidate of $F_{Y_d^{CO}}$. It is clear that the simple additive structure of all imposed conditions implies that if there are two different such feasible candidate functions, any convex combinations of these functions satisfy these conditions as well.

Part II. We show in the following both that the proposed bounds $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ satisfy the constraints in (A.1.29) – (A.1.32) and that any function which takes values outside of these bounds contradicts one of these conditions and is therefore incompatible with the distribution of (Y, D, Z), our assumption and the sensitivity parameters. As the considered sensitivity parameters lie within the sensitivity region by assumption, bounds on the outcome distribution of compliers exist and are therefore non-intersecting by construction. The condition in (A.1.29) is therefore satisfied if our bounds $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ satisfy

$$\underline{H}_{Y_d^{CO}}(y) \leq \underline{F}_{Y_d^{CO}}(y) \qquad \text{and} \qquad \overline{F}_{Y_d^{CO}}(y) \leq \overline{H}_{Y_d^{CO}}(y) \tag{A.1.33}$$

for all $y \in \mathbb{Y}$. Additionally, both bounds preserve the existence of limits and continuity.

Part II - Lower Bound. We consider first

$$\underline{H}_1(y) = \frac{1}{\pi_{\text{CO}}} \left(\pi_\Delta G_d^+(y) - \inf_{\widetilde{y} \leq y} \left(\pi_\Delta G_d^+(\widetilde{y}) - \pi_{\text{CO}} \underline{H}(\widetilde{y}) \right) \right). \tag{A.1.34}$$

It clearly holds that $\underline{H}_1(y) \geq \underline{H}_{Y_d^{CO}}(y)$. Consider again any $y, y' \in \mathbb{Y}$ such that $y' \leq y$. Based on this reasoning $\underline{H}_1(y)$ satisfies constraint (A.1.31) as

$$\begin{aligned}
&\underline{H}_1(y) - \underline{H}_1(y') \\
&= \frac{1}{\pi_{\text{CO}}} \left(\pi_\Delta G_d^+(y) - \pi_\Delta G_d^+(y') - \inf_{\widetilde{y} \leq y} \left(\pi_\Delta G_d^+(\widetilde{y}) - \underline{H}(\widetilde{y}) \right) - \inf_{\widetilde{y} \leq y'} \left(\pi_\Delta G_d^+(\widetilde{y}') - \underline{H}_{Y_d^{CO}}(\widetilde{y}') \right) \right) \\
&\geq \frac{1}{\pi_{\text{CO}}} \left(\pi_\Delta G_d^+(y) - \pi_\Delta G_d^+(y') \right).
\end{aligned}$$

Any function F such that $F(y) \leq \underline{H}_1(y)$ either violates (A.1.31) or (A.1.33) for some $y \in \mathbb{Y}$. We conclude that any feasible candidate function of $F_{Y_d^{CO}}$ has to satisfy

$$\underline{H}_1(y) \leq F_{Y_d^{CO}}(y) \tag{A.1.35}$$

We now consider our final lower bound

$$\underline{F}_{Y_d^{CO}}(y) = \frac{1}{\pi_{\text{CO}}} \left(Q_{dd}(y) - \inf_{\widetilde{y} \geq y} \left(Q_{dd}(\widetilde{y}) - \underline{H}_1(\widetilde{y}) \right) \right).$$

It is clear that $\underline{F}_{Y_d^{CO}}(y) \geq \underline{H}_1(y)$ and that $Q_{dd}(y) - Q_{dd}(y') \geq \underline{F}_{Y_d^{CO}}(y) - \underline{F}_{Y_d^{CO}}(y')$. As it further holds that, for any $y, y' \in \mathbb{Y}$ and $y' \leq y$,

$$Q_{dd}(y) - Q_{dd}(y') \geq G_d^+(y) - G_d^+(y')$$

$\underline{F}_{Y_d^{CO}}(y)$ satiates (A.1.31), (A.1.30), and it holds that $\underline{F}_{Y_d^{CO}}(y) \geq \underline{H}_1(y)$. Clearly, any function F such that $F(y) \leq \underline{H}_1(y)$ is incompatible with the distribution of (Y, D, Z), the sensitivity parameters and our assumptions.

We now show that $\underline{F}_{Y_d^{CO}}(y)$ satisfies (A.1.32) . By construction, $\underline{F}_{Y_d^{CO}} \in [0, 1]$. We therefore show that $\lim_{y \to \underline{y}} \underline{F}_{Y_d^{CO}}(y) \leq 0$ and $\lim_{y \to \overline{y}} \underline{F}_{Y_d^{CO}}(y) \geq 1$. It holds that

$$\lim_{y \to \underline{y}} \underline{F}_{Y_d^{CO}}(y) = \frac{1}{\pi_{\text{CO}}} \inf_{\widetilde{y} \in \mathbb{R}} \left(Q_{dd}(\widetilde{y}) - (\pi_\Delta G_d^+(\widetilde{y}) - \inf_{\widehat{y} \leq \widetilde{y}} \left(\pi_\Delta G_d^+(\widehat{y}) + \pi_{\text{CO}} \underline{H}_{Y_d^{CO}}(\widehat{y}) \right)) \right)$$

The equality follows as $\lim_{y \to \underline{y}} Q_{dd}(y) = 0$. We note that for all $y, y' \in \mathbb{Y}$ and $y' \leq y$

$Q_{dd}(y) - Q_{dd}(y') \geq \pi_\Delta \left(G_d^+(y) - G_d^+(y') \right)$. It follows that

$$\begin{aligned}
&\lim_{y \to \underline{y}} \underline{F}_{Y_d^{CO}}(y) \\
&\leq \frac{1}{\pi_{\text{CO}}} \inf_{\widehat{y} \in \mathbb{Y}} \big(\max\{ \underbrace{0}_{(1a)}, \underbrace{Q_{dd}(\widehat{y}) - \pi_d}_{(2a)}, \underbrace{\pi_\Delta G_d^{\sup}(\widehat{y})}_{(3a)}, \underbrace{\pi_{\text{CO}} G_d^{\sup}(\widehat{y}) - \pi_{\text{CO}} \frac{\pi_{\text{DF}}}{\pi_\Delta} \delta}_{(4a)} \} - Q_{dd}(\widehat{y}) \big).
\end{aligned}$$

We now show that each of the expressions (1a)–(4a) evaluated at any $\widehat{y} \in \mathbb{Y}$ is bounded by $Q_{dd}(\widehat{y})$ so that it holds that $\lim_{y \to \underline{y}} \pi_{\text{CO}} \underline{F}_{Y_d^{CO}}(y) \leq 0$. It is obvious that expressions $(1a)$ and $(2a)$ satisfy this reasoning. Considering $(3a)$, we note that it holds that $\pi_\Delta G_d^{\sup}(\widehat{y}) \leq Q_{dd}(\widehat{y})$. We turn to (4a). It holds that

$$G_d^{\sup}(\widehat{y}) - \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta - \frac{Q_{dd}(\widehat{y})}{\pi_{\text{CO}}} \leq F_{Y_d^{CO}}(\widehat{y}) + \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta - \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta - \frac{Q_{dd}(\widehat{y})}{\pi_{\text{CO}}} \leq F_{Y_d^{CO}}(\widehat{y}) - \frac{Q_{dd}(\widehat{y})}{\pi_{\text{CO}}} \leq 0$$

We consider the right limit. It holds that $\underline{F}_{Y_d^{CO}} \geq \underline{H}_{Y_d^{CO}}$ and therefore

$$\lim_{y \to \overline{y}} \underline{F}_{Y_d^{CO}}(y) \geq \lim_{y \to \overline{y}} \max\{0, G_d^{\sup}(y) - \frac{\pi_{\text{DF}}}{\pi_\Delta}\delta, \frac{\pi_\Delta}{\pi_{\text{CO}}} G_d^{\sup}(y), \frac{Q_{dd}(y) - \pi_d}{\pi_{\text{CO}}}\} \geq 1.$$

The second inequality follows as $\lim_{y \to \overline{y}} Q_{dd}(y) - \pi_d = \pi_{\text{CO}} + \pi_d - \pi_d = \pi_{\text{CO}}$. This reasoning concludes the proof of the lower bound.

Part II - Upper Bound. A similar reasoning applies to the upper bound. To briefly sketch this reasoning, let

$$\overline{H}_1(y) = Q_{dd}(y) - \sup_{\widehat{y} \leq y} \left(Q_{dd}(\widehat{y}) - \pi_{\text{CO}} \overline{H}_{Y_d^{CO}}(\widehat{y}) \right).$$

It is clear that $\overline{H}_1(y) \geq \overline{H}_{Y_d^{CO}}(y)$ and that $Q_{dd}(y) - Q_{dd}(y') \geq \overline{H}_1(y) - \overline{H}_1(y')$. It holds that $\overline{H}_1(y')$ satisfies (A.1.30). Clearly, any function F such that $F(y) \leq \underline{H}_1(y)$ is incompatible with the distribution of (Y, D, Z), the sensitivity parameters and our assumptions. It therefore follows that any function which is a feasible candidate of the distribution function $F_{Y_d^{CO}}$ has to satisfy

$$F_{Y_d^{CO}}(y) \leq \overline{H}_1(y) \tag{A.1.36}$$

We now consider our proposed bound.

$$\overline{F}_{Y_d^{CO}}(y) = \pi_\Delta G_d^+(y) - \sup_{\widetilde{y} \geq y} \left(\pi_\Delta G_d^+(\widetilde{y}) - \overline{H}_1(y) \right).$$

It follows from the same reasoning as above that $\overline{F}_{Y_d^{CO}}(y)$ satisfies (A.1.36), (A.1.30) and (A.1.31). Clearly, any function F such that $F(y) \geq \overline{F}_{Y_d^{CO}}(y)$ is incompatible with the

distribution of (Y, D, Z), the sensitivity parameters and our assumptions.

We conclude by showing that $\overline{F}_{Y_d^{CO}}(y)$ satisfies (A.1.32). It holds that

$$\lim_{y\to\underline{y}} \overline{F}_{Y_d^{CO}}(y) \leq \lim_{y\to\underline{y}} \min\{1, G_d^{\inf}(y) + \frac{\pi_{\mathrm{DF}}}{\pi_\Delta}\delta, \frac{\pi_{\mathrm{CO}}}{\pi_\Delta} G_d^{\inf}(y) + \frac{\pi_{\mathrm{CO}}}{\pi_{\mathrm{DF}}}, \frac{Q_{dd}(y)}{\pi_{\mathrm{CO}}}\} \leq 0,$$

where the second inequality follows by $\lim_{y\to\underline{y}} \frac{Q_{dd}(y)}{\pi_{\mathrm{CO}}} = 0$. We now consider

$$\begin{aligned}
&\lim_{y\to\overline{y}} \overline{F}_{Y_d^{CO}}(y) \\
&= \frac{\pi_{\mathrm{CO}} + \pi_d}{\pi_{\mathrm{CO}}} - \sup_{\widehat{y}\in\mathbb{Y}} \big(\frac{Q_{dd}(\widehat{y})}{\pi_{\mathrm{CO}}} - \min\{\underbrace{1}_{(1b)}, \underbrace{\frac{Q_{dd}(\widehat{y})}{\pi_{\mathrm{CO}}}}_{(2b)}, \underbrace{\frac{\pi_\Delta}{\pi_{\mathrm{CO}}} G_d^{\inf}(\widehat{y}) + \frac{\pi_{\mathrm{DF}}}{\pi_{\mathrm{CO}}}}_{(3b)}, \underbrace{G_d^{\inf}(\widehat{y}) + \frac{\pi_{\mathrm{DF}}}{\pi_\Delta}\delta}_{(4b)}\}\big).
\end{aligned}$$

We show that (1b)–(4b) are bounded from below by $\frac{Q_{dd}(\widehat{y})}{\pi_{\mathrm{CO}}} - \frac{\pi_d}{\pi_{\mathrm{CO}}}$ such that

$$\lim_{y\to\overline{y}} \overline{F}_{Y_d^{CO}}(y) \geq \frac{\pi_{\mathrm{CO}} + \pi_d}{\pi_{\mathrm{CO}}} - \frac{\pi_d}{\pi_{\mathrm{CO}}} = 1.$$

It is clear that (1b)-(2b) satisfies this restriction. Concerning (3b), we note that

$$\frac{1}{\pi_{\mathrm{CO}}} \left(\pi_\Delta G_d^{\inf}(\widehat{y}) + \pi_{\mathrm{DF}}\right) \geq \frac{1}{\pi_{\mathrm{CO}}} (Q_{dd}(\widehat{y}) + \pi_{\mathrm{DF}} - \pi_{\mathrm{d}} - \pi_{\mathrm{DF}}) = \frac{Q_{dd}(\widehat{y}) - \pi_{\mathrm{d}}}{\pi_{\mathrm{CO}}}.$$

Concerning (4b), we note that

$$G_d^{\inf}(\widehat{y}) + \frac{\pi_{\mathrm{DF}}}{\pi_\Delta}\delta \geq F_{Y_d^{CO}}(\widehat{y}) - \frac{\pi_{\mathrm{DF}}}{\pi_\Delta}\delta + \frac{\pi_{\mathrm{DF}}}{\pi_\Delta}\delta \geq \frac{Q_{dd}(\widehat{y})}{\pi_{\mathrm{CO}}} - \frac{\pi_d}{\pi_{\mathrm{CO}}}$$

This completes this proof. □

1.B.2. **Proof of Proposition 1.1.** We show that the population size of compliers is sharply bounded by $\underline{\pi}_{CO} \leq \pi_{\mathrm{CO}} \leq \overline{\pi}_{CO}$, where for $B \in \mathscr{B}$ and $d, z \in \{0, 1\}$

$$\begin{aligned}
\overline{\pi}_{CO} &= \min\{\mathbb{P}(D = 1|Z = 1), \mathbb{P}(D = 0|Z = 0)\} \\
\underline{\pi}_{CO} &= \max_{d\in\{0,1\}} \{\sup_{B\in\mathscr{B}} \{\mathbb{P}(Y \in B, D = d|Z = d) - \mathbb{P}(Y \in B, D = d|Z = 1 - d)\}\}.
\end{aligned}$$

The proposition follows from this statement as $\pi_{DF} = \pi_{CO} - \mathbb{P}(D = 1|Z = 1) + \mathbb{P}(D = 1|Z = 0)$. Let $\mathbb{P}(Y_d^t \in B)$ denotes the unobserved probability distribution of the potential outcome of group t with treatment status d.[35]

$\underline{\pi}_{CO}$ is a valid lower bound of the population size of compliers as it follows from the

[35]In principle, Proposition 1.1 is a Corollary of Theorem 1.1. Considering the sharp lower bound on the population size of defiers, one could simply use the bounds to solve for the minimal size of defiers for which there exists one value of outcome heterogeneity δ such that the bounds are non-intersecting. However, this exercise is tedious, and we propose a simpler and direct proof for this claim in this section.

definition of groups that

$$\overline{\pi}_{CO} = \min\{\pi_{AT} + \pi_{CO}, \pi_{NT} + \pi_{CO}\} \geq \pi_{CO}.$$

Similarly, $\underline{\pi}_{CO}$ bounds the population size of compliers from below as $\underline{\pi}_{CO}$ equals

$$\begin{aligned}
\max \Bigg\{ & \sup_{B \in \mathscr{B}} \{\mathbb{P}(Y_1 \in B, AT) + \mathbb{P}(Y_1 \in B, CO) - \mathbb{P}(Y_1 \in B, AT) - \mathbb{P}(Y_1 \in B, DF)\}, \\
& \sup_{B \in \mathscr{B}} \{\mathbb{P}(Y_0 \in B, NT) + \mathbb{P}(Y_0 \in B, CO) - \mathbb{P}(Y_0 \in B, NT) - \mathbb{P}(Y_0 \in B, DF)\} \Bigg\} \\
\leq \max\{ & \sup_{B \in \mathscr{B}} \{\mathbb{P}(Y_1 \in B, CO)\}, \sup_{B \in \mathscr{B}} \{\mathbb{P}(Y_0 \in B, CO)\}\} = \pi_{CO}.
\end{aligned}$$

The inequality follows from the independence assumption and the definition of the groups. It is therefore clear that the population size of compliers lies within the bounds. It remains to show that these bounds are sharp. To show this, we consider any fix $\widetilde{\pi}_{CO} \in [\underline{\pi}_{CO}, \overline{\pi}_{CO}]$. Let $B \in \mathscr{B}$ and $\mathscr{B}_B = \{A \cap B | A \in \mathscr{B}\}$. Using the discussion of the proof of Theorem 1.1 how to verify that a candidate distribution is a feasible distribution, we consider the following marginal outcome distributions of the groups.[36]

$$\begin{aligned}
\mathbb{P}(\widetilde{Y}_d \in B, T = CO) &= \mathbb{P}(Y \in B, D = d | Z = d) - \mathbb{P}(\widetilde{Y}_d \in B, T = dT), \\
\mathbb{P}(\widetilde{Y}_d \in B, T = DF) &= \mathbb{P}(Y \in B, D = d | Z = 1 - d) - \mathbb{P}(\widetilde{Y}_d \in B, T = \ dT), \\
\mathbb{P}(\widetilde{Y}_d \in B, T = dT) &= L_1 \cdot L_2,
\end{aligned}$$

where

$$\begin{aligned}
L_1 &= \frac{\mathbb{P}(D = d | Z = d) - \widetilde{\pi}_{CO}}{\mathbb{P}(D = d | Z = d) - \sup_{C \in \mathscr{B}}(\mathbb{P}(Y \in C, D = d | Z = d) - \mathbb{P}(Y \in C, D = d | Z = 1 - d))}, \\
L_2 &= \mathbb{P}(Y \in B, D = d | Z = d) \\
&\quad - \sup_{C \in \mathscr{B}_B} (\mathbb{P}(Y \in C, D = d | Z = d) - \mathbb{P}(Y \in C, D = d | Z = 1 - d)).
\end{aligned}$$

The outcome distribution of group dT is the product of two terms. The term L_1 guarantees that the probability distributions integrate to the corresponding population size. The term L_2 guarantees that the outcome probabilities of compliers and defiers are non-negative. The outcome distributions of the other groups are respectively defined.

By construction, the proposed outcome probability distributions imply the observed outcome probability distributions.[37] We show now that the implied probability distri-

[36] Otherwise $\mathbb{P}(\widetilde{Y}_d \in B, T = dT)$ is defined to be zero if $\mathbb{P}(D = d | Z = d) = \sup_{C \in \mathscr{B}}(\mathbb{P}(Y \in C, D = d | Z = d) - \mathbb{P}(Y \in C, D = d | Z = 1 - d))$. The other probability distributions stay the same.

[37] This means that $\forall B \in \mathscr{B}$ and $\forall d, z \in \{0, 1\}$ $\mathbb{P}(Y \in B, D = d | Z = z) = \mathbb{P}(\widetilde{Y} \in B, \widetilde{D} = d | \widetilde{Z} = z)$.

butions are indeed distributions, which satisfy $\forall\ T \in \{CO, DF, AT, NT\}$, $d \in \{0,1\}$, and $B, B' \in \mathscr{B}$, where $B' \subseteq B$: (i) $\mathbb{P}(\widetilde{Y}_d^T \in \mathbb{Y}) = 1$; (ii) $\mathbb{P}(\widetilde{Y}_d^T \in B) \geq 0$; and (iii) $\mathbb{P}(\widetilde{Y}_d^T \in B) \geq \mathbb{P}(\widetilde{Y}_d^T \in B')$. We consider any $B \in \mathcal{B}$ and any $d \in \{0,1\}$ in the following.

We first consider condition (i). It clearly holds that $\mathbb{P}(\widetilde{Y}_d \in \mathbb{Y}, T = CO) = \widetilde{\pi}_{CO}$, and

$$\begin{aligned}
&\mathbb{P}(\widetilde{Y}_d \in \mathbb{Y}, T = dT) = \mathbb{P}(D = d|Z = d) - \widetilde{\pi}_{CO} = \widetilde{\pi}_{dT} \\
&\mathbb{P}(\widetilde{Y}_d \in \mathbb{Y}, T = DF) = \mathbb{P}(D = d|Z = 1-d) - \mathbb{P}(D = d|Z = d) + \widetilde{\pi}_{CO} = \widetilde{\pi}_{DF}.
\end{aligned}$$

We turn to condition (ii). First note that it follows from the bounds on the population size of compliers $\underline{\pi}_{CO}$ that $0 \leq L_1 \leq 1$. Second, we note that

$$L_2 \geq \mathbb{P}(Y \in B, D = d|Z = d) - \sup_{C \in \mathscr{B}_B} (\mathbb{P}(Y \in C, D = d|Z = d)) = 0.$$

This reasoning implies that $\mathbb{P}(\widetilde{Y}_d \in B, T = dT) \geq 0$. Further, note that

$$\begin{aligned}
&\mathbb{P}(\widetilde{Y}_d \in B, T = DF) = \mathbb{P}(Y \in B, D = d|Z = 1-d) - \mathbb{P}(\widetilde{Y}_d \in B, T = dT) \\
&= \mathbb{P}(Y \in B, D = d|Z = 1-d) - L_1 \mathbb{P}(Y \in B, D = d|Z = d) \\
&\quad + L_1 \sup_{C \in \mathscr{B}_B} (\mathbb{P}(Y \in C, D = d|Z = d) - \mathbb{P}(Y \in C, D = d|Z = 1-d)) \geq 0,
\end{aligned}$$

by basic arguments about sets. A similar reasoning applies to the compliers.

We consider condition (iii). Let $B' \subseteq B$. We note that

$$\begin{aligned}
&\mathbb{P}(\widetilde{Y}_d \in B, T = dT) - \mathbb{P}(\widetilde{Y}_d \in B', T = dT) \\
&\geq \mathbb{P}(Y \in B \backslash B', D = d|Z = d) \\
&\quad - \sup_{C \in \mathscr{B}_{B \backslash B'}} (\mathbb{P}(Y \in C, D = d|Z = d) - \mathbb{P}(Y \in C, D = d|Z = 1-d)) \geq 0.
\end{aligned}$$

Using a simple arguments, it further holds that $\mathbb{P}(\tilde{Y}_d \in B, T = DF) \geq \mathbb{P}(\tilde{Y}_d \in B', T = DF)$ as

$$\begin{aligned}
&\mathbb{P}(\tilde{Y}_d \in B, T = dT) - \mathbb{P}(\tilde{Y}_d \in B, T = dT) \\
&= \mathbb{P}(Y \in B, D = d|Z = d) - \mathbb{P}(Y \in B', D = d|Z = d) \\
&\quad - \sup_{C \in \mathscr{B}_{B \backslash B'}} (\mathbb{P}(Y \in C, D = d|Z = d) - \mathbb{P}(Y \in C, D = d|Z = 1-d)) \\
&\leq \mathbb{P}(Y \in B, D = d|Z = 1-d) - \mathbb{P}(Y \in B', D = d|Z = 1-d).
\end{aligned}$$

A similar reasoning applies to the compliers, which completes this proof. □

1.B.3. **Proof of Proposition 1.2.** It holds by our assumptions that, for $d \in \{0,1\}$,

$$F_{Y_d}(y) = \pi_{1-d} F_{Y_d^{(1-d)T}}(y) + \pi_{\text{CO}} F_{Y_d^{CO}}(y) + \pi_{\text{DF}} F_{Y_d^{DF}}(y) + \pi_d F_{Y_d^{dT}}(y),$$

where π_{1-d} is the population size of always takers if $d = 0$ and otherwise of never takers and $F_{Y_d^{(1-d)T}}$ respectively. In the absence of treatment, the data generating process does not reveal anything about the distribution of the always takers, and neither in the presence of treatment of the never takers. The proof of Theorem 1.1 implies sharp bounds on the remaining six potential outcome distributions. Using (1.5) and (1.6), it follows that

$$\begin{aligned}
&F_{Y_d}(y) \\
&= \pi_{1-d} F_{Y_d^{(1-d)T}}(y) + \pi_{\text{CO}} F_{Y_d^{CO}}(y) - \pi_{\Delta} G_d(y) + \pi_{\text{CO}} F_{Y_d^{CO}}(y) + Q_{dd}(y) - \pi_{\text{CO}} F_{Y_d^{CO}}(y). \\
&= \pi_{1-d} F_{Y_d^{(1-d)T}}(y) + \pi_{\text{CO}} F_{Y_d^{CO}}(y) - \pi_{\Delta} G_d(y) + Q_{dd}(y) \\
&= \pi_{1-d} F_{Y_d^{(1-d)T}}(y) + \pi_{\text{CO}} F_{Y_d^{CO}}(y) + Q_{d(1-d)}(y)
\end{aligned}$$

Sharp bounds in a first-order stochastic dominance sense of $F_{Y_d}(y)$ are therefore obtained by Theorem 1.1 by taking the distribution functions $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ and setting $F_{Y_d^{(1-d)T}}(y)$ to its most extreme values, respectively. The statement follows from this reasoning. □

1.B.4. **Proof of Corollary 1.1.** The statement follows directly from first-order stochastic dominance of the distribution functions $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ by Theorem 1.1 and Lemma 1 in Stoye (2010). □

1.B.5. **Proof of Corollary 1.2.** The statement directly follows from Theorem 1.1 by noting how the bounds simplify for a binary variable. □

1.B.6. **Proof of Proposition A.1.3.** By the same arguments of the proof of Theorem 1.1, one can show that these bounds are sharp conditionally on the covariates given the respective assumptions. □

1.B.7. **Verification of Expressions used throughout the Paper.** In this section, we verify a few expressions, which we have used through the text. We emphasize that they rely on textbook arguments, but we show them for completeness.

1.B.7.1. *Verification of Equation* (1.3). For completeness, we want here also to verify one of the main equations used in this analysis, Equation (1.3). Similar arguments can be found in Imbens and Angrist (1994), and in many textbooks.

Lemma A.1.1. *Let Assumptions 1.1 hold. Then Equation* (1.3) *is satisfied.*

Proof: We note that

$$G_1(y) = \frac{\text{Cov}(\mathbf{1}\{Y \leq y\}, Z)}{\text{Cov}(Z, D)} = \frac{\mathbb{E}[\mathbf{1}\{Y \leq y\}D|Z = 1] - \mathbb{E}[\mathbf{1}\{Y \leq y\}D|Z = 0]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]}.$$

It then follows from the independence assumption and the definition of the groups that

$$\begin{aligned} G_1(y) &= \frac{\mathbb{E}[D\mathbf{1}\{Y^{CO} \leq y\}]\pi_{\text{CO}} + \mathbb{E}[D\mathbf{1}\{Y^{AT} \leq y\}]\pi_{\text{AT}}}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]} \\ &\quad - \frac{\mathbb{E}[D\mathbf{1}\{Y^{AT} \leq y\}]\pi_{\text{AT}} + \mathbb{E}[D\mathbf{1}\{Y^{DF} \leq y\}]\pi_{\text{DF}}}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]} \\ &= \frac{\pi_{\text{CO}}}{\pi_{\text{CO}} - \pi_{\text{DF}}} F_{Y_1^{CO}}(y) - \frac{\pi_{\text{DF}}}{\pi_{\text{CO}} - \pi_{\text{DF}}} F_{Y_1^{DF}}(y) \end{aligned}$$

The denominator is positive by the relevance assumption. $G_0(y)$ follows similarly. □

1.B.7.2. *Verification of Properties of the function $G_d^+(y)$.*

Lemma A.1.2. *Suppose $Q_{ds}(y)$ is continuously differentiable in $y \in \mathbb{R}$ for $d, s \in \{0, 1\}$. Then,*

$$G_d^+(y) = \int_{\mathbb{Y}} \mathbf{1}\{z \leq y\} \max\{0, g_d(z)\} dz. \tag{A.1.37}$$

Proof: We note that

$$\begin{aligned} G_d^+(y) &= \frac{1}{\pi_\Delta} \sup_{B \in \mathscr{B}} \{\mathbb{P}(Y \in B, Y \leq y, D = d|Z = d) - \mathbb{P}(Y \in B, Y \leq y, D = d|Z = 1 - d)\} \\ &= \frac{1}{\pi_\Delta} \sup_{B \in \mathscr{B}} \{\int_{\mathbb{Y}} \mathbf{1}\{z \in B\}\mathbf{1}\{z \leq y\} q_{dd}(z) dz - \int_{\mathbb{Y}} \mathbf{1}\{z \in B\}\mathbf{1}\{z \leq y\} q_{d(1-d)}(z) dz\} \\ &= \int_{\mathbb{Y}} \mathbf{1}\{z \leq y\} \max\{0, g_d(z)\} dz. \end{aligned}$$

The first inequality follows from the definition of probabilities and our definition of $q_{ds}(z)$. The second equality follows by continuity of $q_{d(1-s)}$. □

1.B.7.3. *Outer and Inner Set for Sensitivity and Robust Region.* We first verify that our expression (1.26) follows from expression (1.27). Let it holds that $\phi_L(\pi_{\text{DF}}; \theta,) \leq \phi(\pi_{\text{DF}}; \theta)$ for each component and for all $\pi_{\text{DF}} \in [0, 0.5)$. We denote the $l-th$ unit vector by e_l. We then note that by the definition of $\phi_L(\pi_{\text{DF}}; \theta)$ and SR of Section 1.5 that

$$\begin{aligned} SR &= \big\{(\pi_{\text{DF}}, \delta) : e_1^\top \phi(\theta, \pi_{\text{DF}}) \leq \pi_{\text{DF}} \leq -e_2^\top \phi(\theta, \pi_{\text{DF}}) \\ &\qquad e_3^\top \phi(\theta, \pi_{\text{DF}}) \leq \delta \leq -e_4^\top \phi(\theta, \pi_{\text{DF}})\big\} \\ &\subseteq \big\{(\pi_{\text{DF}}, \delta) : e_1^\top \phi_L(\theta, \pi_{\text{DF}}) \leq \pi_{\text{DF}} \leq -e_2^\top \phi_L(\theta, \pi_{\text{DF}}) \\ &\qquad e_3^\top \phi_L(\theta, \pi_{\text{DF}}) \leq \delta \leq -e_4^\top \phi_L(\theta, \pi_{\text{DF}})\big\} = SR_L. \end{aligned}$$

By a similar argument we note that

$$\begin{aligned} RR_L(\mathrm{SR}_L) &= \{(\pi_{\mathrm{DF}}, \delta) \in \mathrm{SR}_L : \delta \leq e_5^\top \phi_L(\theta, \pi_{\mathrm{DF}})\} \\ &\supseteq \{(\pi_{\mathrm{DF}}, \delta) \in \mathrm{SR}_L : \delta \leq e_5^\top \phi(\theta, \pi_{\mathrm{DF}})\} \quad = RR(\mathrm{SR}_L). \end{aligned}$$

As we have shown above that $SR \subseteq SR_L$ it follows that $RR_L(\mathrm{SR}) \supseteq RR(\mathrm{SR})$. □

1.C. ADDITIONAL MATERIALS FOR ESTIMATION AND INFERENCE

In this section, we present more details on the estimation and inference methods proposed in the main text. We first consider a binary and then a continuous outcome variable. For both cases, we provide more details about estimating the sensitivity and robust regions, we discuss the imposed assumptions and then proceed by showing asymptotic results. Many of the following results are based on applications and ideas from other papers and we therefore only sketch most of them.

1.C.1. **Estimation for a binary outcome variable.** We consider a binary outcome variable, where the mapping of interest is $\phi_b(\theta_b, \pi_{\mathrm{DF}})$. As shown in Section 1.6.2, the underlying parameters θ_b are given by $(P_{11}, P_{10}, P_{01}, P_{00}, P_0, P_1)$. We estimate the probabilities by their sample counterparts, i.e. $P_{ds} = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{1}\{Y_i^s = 1, D_i^s = d\}$ and $P_s = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{1}\{D_i^s = 1\}$. We then estimate $\phi_b(\widehat{\theta}_b, \pi_{\mathrm{DF}})$ by simple plug-in estimates, where the precise formulas are given in Section 1.6.2 and Appendix 1.A.4.

1.C.2. **Assumptions.** We consider the following sampling process.

Assumption A.1.3. *For $z \in \{0,1\}$, $\{(Y_i^z, D_i^z)\}_{i=1}^{n_z}$ are identically and independently distributed according to the distribution of (Y^z, D^z) which is drawn conditional on $Z = z$ with support $\mathbb{Y} \times \{0,1\}$. It holds that n_0/n converges to a nonzero constant as $n \to \infty$.*

By Assumption 1.1, the instrument is independent of all potential outcomes, so that the distribution of the instrument does not contain any further information and we can assume that the sampling is conditionally on the instrument (see, e.g., Kitagawa, 2015).

1.C.3. **Inference for a binary outcome variable.** In this section, we present more details on how to construct confidence sets for a binary outcome variable. Based on the derivation in Section 1.6.2 and in Appendix 1.A.4, it suffices to construct a lower confidence band for $\tilde{\phi}_b(\hat{\theta}_b, \pi_{\mathrm{DF}})$ given in (A.1.28). To unify the notation, let us denote the i-th component of this mapping by $\phi_{b,i}(\theta_b, \pi_{\mathrm{DF}})$ for $i \in \{1, \cdots, 6\}$. We note that each of these components can be written as

$$\phi_{b,i}(\theta_b, \pi_{\mathrm{DF}}) = \max\{\psi_{i,j}(\theta_b, \pi_{\mathrm{DF}})\}_{j=1}^{J(i)},$$

where $\psi_{i,j}(\theta_b, \pi_{\text{DF}})$ are Hadamard-differentiable functions of $(\theta_b, \pi_{\text{DF}})$ by the relevance assumption. The mappings $\phi_{b,i}(\theta_b, \pi_{\text{DF}})$ are not Hadamard-differentiable on $(\theta_b, \pi_{\text{DF}})$, but they are Hadamard-directionally-differentiable in the direction of θ_b when evaluated at any finite set of $\{\pi_{\text{DF}}^k\}_{k=1}^K$, where $\pi_{\text{DF}}^k \in [0, 0.5]$ and K is some finite number.

Following ideas of Fang and Santos (2018) and Masten and Poirier (2020), we consider a bootstrap method to construct confidence sets $\tilde{\phi}_{b,i}(\theta_b, \pi_{\text{DF}}^k)$ which are uniformly valid across k and i. Specifically, the directional derivative of $\tilde{\phi}_{b,i}(\theta_b, \pi_{\text{DF}})$ in the direction of θ_b evaluated at some π_{DF} is given by

$$\widetilde{\phi}'_{i,b,\theta_b}(h, \pi_{\text{DF}}) = \max_{j:\psi_{1,j}(\theta_b,\pi_{\text{DF}}) \geq \max_{s \leq J(i)}\{\psi_{1,s}(\theta_b,\pi_{\text{DF}})\}} h_j,$$

for all $h \in \mathbb{R}^{J(i)}$.[38] Following Fang and Santos (2018), we consider as an estimator of this directional derivative,

$$\widehat{\widetilde{\phi}}'_{i,b,\theta_b}(h, \pi_{\text{DF}}) = \max_{j:\psi_{1,j}(\theta_b,\pi_{\text{DF}}) \geq \max_{s \leq J(i)}\{\psi_{1,s}(\theta_b,\pi_{\text{DF}})+\kappa\}} h_j,$$

where $\kappa > 0$ and $\kappa \to 0$ and $\kappa\sqrt{n} \to \infty$ as $n \to \infty$.

We first get estimates of θ_b and $\phi_b(\theta_b, \pi_{\text{DF}}^k)$ from the original sample for all $k \in \{1, \ldots, K\}$. We then generate B bootstrap samples $\{(Y_i^{b,z}, D_i^{b,z})\}_{i=1}^{n_z}$, $b = 1, \ldots, B$ by drawing n_z observations with replacements from the original data $\{Y_i^z, D_i^z\}_{i=1}^{n_z}$ for $z \in \{0, 1\}$ and we calculate $\widehat{\phi}'_{b,\theta}$ for each bootstrap iteration. We take

$$\widehat{\text{cv}}_{1-\alpha} = \inf(z : \mathbb{P}(\max_{k \in 1,\ldots,K} \tilde{\phi}'_{b,\theta}((\sqrt{n}(\widehat{\theta}_b^\star - \widehat{\theta}_b); \pi_{\text{DF}}^k) - z) \leq 0) \geq 1 - \alpha'),$$

where $\alpha' < \alpha$ but arbitrarily close to α.[39] We then consider as lower confidence set $\tilde{\phi}_b(\widehat{\theta}_b, \pi_{\text{DF}}^k) - \widehat{\text{cv}}_{1-\alpha}/\sqrt{n}$ for all $k \in \{1, \ldots, K\}$. These lower confidence sets are uniformly valid for the mapping $\tilde{\phi}_b$ when evaluated at $\{\pi_{\text{DF}}^k\}_{k=1}^K$.

To obtain a lower confidence band of $\tilde{\phi}_b$, which is valid uniformly in π_{DF}, we exploit the functional form of $\tilde{\phi}_b$ similarly to Masten and Poirier (2020). The lower bound for intermediates points, that are not within the set $\{\pi_{\text{DF}}^k\}_{k=1}^K$, is interpolated based on the left and right nearest neighbor of the point of evaluation. The respectively lowest confidence set is taken. By monotonicity of $\tilde{\phi}_b$, this lower confidence set is then also valid uniformly valid in π_{DF}.

To construct a valid confidence set for ϕ_b, we then consider a simple projection argu-

[38] See Definition 2.1 in Fang and Santos (2018) for a definition of Hadamard-directional differentiable mappings.

[39] To simplify the notation, we just consider a fix critical value $\text{cv}_{1-\alpha}$ here. In principle, it might be different for each component and for each point of evaluation $\{\pi_{\text{DF}}^k\}_{k=1}^K$ and indeed it would be more efficient to do this.

ment of $\tilde{\phi}_b$ by taking the maximum of the last three components of $\tilde{\phi}_b$ into account. We construct our confidence set for our sensitivity and our robust region $\widehat{RR}_{b,L}$ and $\widehat{\text{SR}}_{b,L}$ based on our constructed lower confidence set.

Proposition A.1.4. *Suppose that Assumption 1.1 and A.1.3 hold and the variance of each component of θ_b is bounded away from zero. It then holds that,*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{RR}_{b,L} \subseteq RR_b,\ \text{SR}_b \subseteq \widehat{\text{SR}}_{b,L}) \geq 1-\alpha.$$

We impose the variance condition to ensure that the underlying parameters converge to a non-degenerated distribution.

1.C.4. **Estimation for a Continuous Outcome Variable.** In this section, we give further details on the construction of the estimators for a continuous outcome variable. We first estimate the underlying parameters θ. We estimate the conditional joint densities by standard nonparametric kernel density estimator

$$\widehat{q}_{dz}(y) = \frac{1}{n_z h} \sum_{i=1}^{n_z} K_h(Y_i^z - y) \cdot \mathbf{1}\{D_i^z = d\},$$

where $K_h(\cdot) = K(\cdot/h)/h$ and $K(\cdot)$ denotes a density function and $h > 0$ a bandwidth. We show in Lemma A.1.2 that our estimator of $G_d^+(y) = \int_{\mathbb{Y}} \mathbf{1}\{\tilde{y} \leq y\} \max\{0, g_d(\tilde{y})\} d\tilde{y}$ under our assumptions. We therefore define

$$\widehat{G}_d^+(y) = \int_{\mathbb{Y}} \mathbf{1}\{\tilde{y} \leq y\} \max\{0, \widehat{g}_d(\tilde{y})\} d\tilde{y},$$

where $\widehat{g}_d(y) = (\widehat{q}_{dd}(y) - \widehat{q}_{d(1-d)}(y))/\widehat{\pi}_\Delta$. The conditional probability functions are further estimated by $\widehat{Q}_{dz}(y) = \int_{\mathbb{Y}} \mathbf{1}\{\tilde{y} \leq y\} \widehat{q}_{dz}(\tilde{y}) d\tilde{y}$. Based on these estimators, the parameters of θ are estimated and we estimate $\phi_b(\widehat{\theta}_b, \pi_{\text{DF}})$ by simple plug-in methods, where infimum, supremum and integrals are numerically evaluated.

1.C.5. **Assumptions for a Continuous Outcome Variable.** We first impose the following regularity assumptions.

Assumption A.1.4. *(i) $\mathbb{Y}_d$ is given by $[\underline{y}_d, \overline{y}_d]$ for $\infty < \underline{y}_d < \overline{y}_d < \infty$ for $d \in \{0,1\}$. (ii) $\forall d, z \in \{0,1\}$, the functions $q_{dz}(y)$ are bounded and bounded away from zero, absolutely continuous and two times continuously differentiable with uniformly bounded derivatives. (iii) For $d \in \{0,1\}$, the functions $q_{dd}(y)$ and $q_{d(1-d)}(y)$ cross at a finite number of times.*

Assumption (i) assumes compact support of the outcome variable as it simplifies the following analysis. Assumption (ii) imposes smoothness conditions on the joint densities, which are standard in the nonparametric literature. Assumption (iii) is imposed

for simplicity and substantially simplifies the analysis of the estimator of the function $G_d^+(y)$.[40]

Assumption A.1.5. (*i*) *The kernel is a second order kernel function, being symmetric around zero, integrates to one, twice continuously differentiable, of bounded variation and zero-valued off, say* $[-0.5, 0.5]$. (*ii*) *The bandwidth satisfies: (a)* $nh^4 \to 0$, *(b)* $nh^2 \to \infty$, *(c)* $nh/\log(n) \to \infty$.

Assumption A.1.5 (i) imposes conditions on the choice of kernel which can be satisfied by construction and Assumption A.1.5 (ii) imposes conditions on the bandwidth.

1.C.6. **Asymptotic Results for a Continuous Outcome Variable.** We first note that we have the following result.

Proposition A.1.5. *Suppose Assumptions A.1.3–A.1.5 hold. It then follows that*

$$\sqrt{n}(\widehat{\theta}(y) - \theta(y)) \to \mathcal{Z}_1(y),$$

where $\mathcal{Z}_1(y)$ *is a tight mean-zero Gaussian process in* $\ell^\infty(\mathbb{R}, \mathbb{R}^6)$.[41]

As explained in the main text, we cannot directly base our inference procedure on the mapping $\phi(\theta, \pi_{\text{DF}})$, as this mapping is non-smooth and standard asymptotic theory cannot be applied. We, therefore, consider a smoothed version of this mapping in this section. To be more precise, we consider the definition of Masten and Poirier (2020), which we cite here for completeness.

Definition 1.1 (Definition 1, Masten and Poirier (2020))**.** Let $(\Theta, \|\cdot\|_\Theta)$ and $(\mathcal{H}, \|\cdot\|_\mathcal{H})$ be Banach spaces. Let $\leq$ be a partial order on $\mathcal{H}$. Let $h : \Theta \to \mathcal{H}$ be a function. Consider a function $H_\kappa : \Theta \to \mathcal{H}$, where $\kappa \in \mathbb{R}_+^{\dim(\kappa)}$ is a vector of smoothing parameters. Then H_κ denotes a *smooth lower approximation* (SLA) of H if

1. Lower envelope: $H_\kappa(\theta) \leq H(\theta)$ for all $\theta \in \Theta$ and $\kappa \in \mathbb{R}_+^{\dim(\kappa)}$.
2. Approximating: For each $\theta \in \Theta$, $H_\kappa(\theta) \to H(\theta)$ for $\kappa \to \infty$ (pointwise).
3. Smoothing: H_κ is Hadamard-differentiable.

[40] This assumption is satisfied if the weighted densities $\pi_{\text{DF}} f_{Y_d^{DF}}$ and $\pi_{\text{CO}} f_{Y_d^{CO}}$ intersect only finitely many times. Without this assumption, our proposed estimator of $\widehat{G}_d^+(y)$ is a biased estimator of $G_d^+(y)$. Following the arguments of Anderson et al. (2012), one can construct a debiased estimator of $G_d^+(\overline{y})$, which converges in $\sqrt{n}$ to a mean-zero normal distribution. Based on similar arguments, one could now construct a debiased estimator of $G_d^+(y)$. As this is a rather tedious exercise and not the purpose of this chapter, we impose this stronger assumption.

[41] Let A be some arbitrary set and B a Banach space. Then $\ell^\infty(A, B)$ denotes the set of all mappings $f : A \to B$, which satisfy that $\sup_{a \in A} ||f(a)||_B \leq \infty$.

This definition of a *smooth upper approximation* (SUA) is analogues. We now assume that ϕ_κ is a SLA of ϕ componentwise, and we show in the subsequent sections how we can obtain such a smooth mapping. Let $\widehat{\theta}^\star$ denotes a draw from the nonparametric bootstrap. We then choose the critical value such that

$$\widehat{\mathrm{cv}}_{1-\alpha} = \inf\{z \in \mathbb{R} : \\ \mathbb{P}\left((\sup_{\pi_{\mathrm{DF}} \in [0,0.5], l \leq 5} \sqrt{n} e_l^\top (\phi_\kappa(\widehat{\theta}^\star, \pi_{\mathrm{DF}}) - \phi_\kappa(\widehat{\theta}, \pi_{\mathrm{DF}})) \leq z | \{\{Y_i^z, D_i^z\}_{i=1}^{n_z}\}_{z=0}^1 \right) \geq 1-\alpha\}$$

We can also allow that z is a known function of π_{DF} and l. By doing so, we can exploit the trade-off by constructing the confidence set for the sensitivity and robust region. We construct our function $\phi_{\kappa,L}(\theta, \pi_{\mathrm{DF}}) = \phi_\kappa(\widehat{\theta}, \pi_{\mathrm{DF}}) + \widehat{\mathrm{cv}}_{1-\alpha}/\sqrt{n}$ and our confidence sets for the sensitivity and robust region $\widehat{\mathrm{SR}}_L(\kappa)$ and $\widehat{RR}_L(\kappa)$ are constructed based on this mapping as explained in Section 1.5. We then have the following result.

Proposition A.1.6. *Suppose Assumptions 1.1 and Assumptions A.1.3–A.1.5 hold. It then follows that* $\lim_{n\to\infty} \mathbb{P}(\widehat{RR}_L(\kappa) \subseteq RR,\ \mathrm{SR} \subseteq \widehat{\mathrm{SR}}_L(\kappa)) \geq 1-\alpha$.

1.C.7. **Population Smoothing.**

1.C.7.1. *General Introduction.* We now show how to construct these smoothed mapping ϕ_κ. As our mapping ϕ is a mapping of many non-differentiable mappings, we prove a chain-rule argument, which allows us to consider simpler mappings.

Lemma A.1.3. *Let ψ and ϕ be two positive and nondecreasing mappings and denote by $\psi^U(\kappa)$ and by $\phi^U(\kappa)$ there respectively SLA, then $\psi^U(f,\kappa)$ and by $\phi^U(\psi^U(f,\kappa),\kappa)$ is a SLA of $\psi(\phi)$. Accordingly, $\phi^U(\psi^U(f,\kappa),\kappa)$ denotes the SUA.*

Based on these definitions, we argue that the mapping $\phi(\theta, \pi_{\mathrm{DF}})$ is a composition of non-smooth random functions, where we replace each of them with a respective SLA and SUA. We first consider these mapping separately, and we then show how to use them to construct our bounds. Let $\kappa > 1$ be the smoothing parameter. Let $(\Theta, \|\cdot\|_\Theta)$ and $(\mathcal{H}, \|\cdot\|_\mathcal{H})$ be Banach spaces, where $\leq$ is a partial order on $\mathcal{H}$. we consider two mappings $f, g : \Theta \to \mathcal{H}$ in the following, which are both Hadamard-differentiable.

Maximum and minimum: We first consider the function $\psi_{av}(f) = |f|$, where a SLA and SUA is given by $\psi_{av}^U(f;\kappa) = \sqrt{f^2 + 1/\kappa}$ and $\psi_{av}^L(f;\kappa) = f^2/(\sqrt{f^2+1/\kappa})$.

Lemma A.1.4. *$\psi_{av}^L(f;\kappa)$ is a SLA and $\psi_{av}^U(f;\kappa)$ a SUA for the mapping $\psi_{av}(f)$ and .*

Let $\psi_{\min}(f,g) = \min(f,g)$ and $\psi_{\max}(f,g) = \max(f,g)$. A SLA of $\psi_{\max}(f,g)$ is clearly given by $\psi_{\max}^L(f,g;\kappa) = f + g + \psi_{av}^L(f-g;\kappa)$ and a SUA is given by $\psi_{\max}^U(f,g;\kappa) =$

$f+g+\psi^U_{av}(f-g;\kappa)$. It follows from a simple induction argument, that one can generalize this procedure to the maximum of a set of finitely many mappings.

Supremum and infimum: In the following, we consider the mapping $\psi_{\sup,\leq}(f,g)(\cdot) = \sup_{z\leq\cdot} f(z) - g(z)$ and the equally binned set $\mathbb{Y} = \bigcup_{k=1}^{\kappa}[\underline{y} + (k-1)d_Y, \underline{y} + kd_Y]$, where $d_Y = \frac{1}{\kappa}(\overline{y} - \underline{y})$. Let $k_j = \underline{y} + j \cdot d_y$, where $j \in \{0, 1, 2, \ldots, \kappa\}$.

$$\psi^L_{\sup,\leq}(f,g;\kappa)(\cdot) = \psi^L_{\max}(\{g(k_j) - f(k_j)\}_{j:k_j\leq\cdot};\kappa).$$
$$\psi^U_{\sup,\leq}(f,g;\kappa)(\cdot) = \psi^U_{\max}(\{g(k_j) - f(\min(\cdot, k_{j+1}))\}_{j:k_j<\cdot}\,;\kappa).$$

We similarly define for the mapping $\psi_{\inf,\leq}(f,g)(\cdot) = \inf_{z\leq\cdot} f(z) - g(z)$ that

$$\psi^U_{\inf}(f,g;\kappa)(\cdot) = \psi^U_{\min}(\{g(k_j) - f(k_{j+1}))\}_{j:k_j\leq\cdot};\kappa).$$
$$\psi^L_{\inf}(f,g;\kappa)(\cdot) = \psi^L_{\min}(\{g(\min(\cdot, k_{j+1}) - f(k_j))\}_{j:k_j<\cdot};\kappa).$$

Lemma A.1.5. *If f and g are monotone increasing, $\psi^L_{\inf,\leq}(f,g;\kappa)$ is a SLA and $\psi^U_{\sup,\leq}(f,g;\kappa)$ a SUA to the function $\psi_{\sup,\leq}(f,g)$, and $\psi^U_{\inf,\leq}(f,g;\kappa)$ a SUA and $\psi^L_{\inf,\leq}(f,g;\kappa)$ a SLA to the function $\psi_{\inf,\leq}(f,g)$.*[42]

1.C.7.2. *Smoothing the Sensitivity and Robust Regions.* We derive the smoothed mapping

$$\phi_\kappa(\theta, \pi_{\mathrm{DF}}) = \left(\underline{\pi}^L_{DF}(\kappa),\ -\overline{\pi}^U_{DF}(\kappa),\ \underline{\delta}^L(\pi_{\mathrm{DF}};\kappa),\ -\overline{\delta}^U(\pi_{\mathrm{DF}};\kappa),\ BP^L(\pi_{\mathrm{DF}};\kappa))\right).$$

Since our sharp bounds $\underline{F}_{Y_d^{CO}}$ and $\overline{F}_{Y_d^{CO}}$ are the key elements in our construction, we consider them first. We show how to smooth the lower bound from above. we note that $G^{\sup}_d(y) = \sup_{z\leq y} G^+_d(z) - G^-_d(z)$, where $G^-_d(z) =$. We denote the upper bound by

$$G^{\sup,U}_d(y;\kappa) = \psi^U_{\sup,\leq}(G^+_d - G^-_d;\kappa)(y)$$

The bound on the outcome distribution of compliers are therefore bounded by

$$\underline{H}^U_{Y_d^{CO}}(y, \pi_{\mathrm{DF}}, \delta;\kappa) =$$
$$\frac{1}{\pi_{\mathrm{CO}}}\,\psi^U_{\max}\left(\left\{0, \pi_\Delta G^{\sup,U}_d(y;\kappa), Q_{\mathrm{dd}}(y) - \pi_d, \frac{\pi_{\mathrm{CO}}}{\pi_\Delta}(G^{\sup,U}_d(y;\kappa) - \pi_{\mathrm{DF}}\delta)\right\};\kappa\right)$$

A SUA of $\underline{F}_{Y_d^{CO}}(y, \pi_{\mathrm{DF}}, \delta)$ is given by

$$\underline{F}^U_{Y_d^{CO}}(y, \pi_{\mathrm{DF}}, \delta;\kappa) = \frac{1}{\pi_{\mathrm{CO}}}Q_{\mathrm{dd}}(y)$$
$$-\frac{1}{\pi_{\mathrm{CO}}}\psi^L_{\inf,\geq}\left(Q_{dd}(\widetilde{y}) - \pi_\Delta G^+_d(\widetilde{y}) - \psi^L_{\inf,\leq}\left(\pi_\Delta G^+_d(\widehat{y}) - \pi_{\mathrm{CO}}\underline{H}^U(y, \pi_{\mathrm{DF}}, \delta;\kappa);\kappa\right)(\widehat{y})\,;\kappa\right)(\widetilde{y}).$$

[42] By similar reasoning, the functions $\psi_{\sup,\geq}(f,g)$ and $\psi_{\inf,\geq}(f,g)$ can be smoothly approximated.

A SLA of $\underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$ can be similarly constructed as well as a smooth lower and upper approximation of $\overline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$. We now turn to the sensitivity region. The lower bounds on the sensitivity parameter $\overline{\pi}_{DF}$ can be constructed by

$$\overline{\pi}^L_{DF}(\kappa) = \psi^L_{\min}(\{\mathbb{P}(D=1|Z=0), \mathbb{P}(D=0|Z=1)\}; \kappa)$$

and similarly upper bound on the sensitivity parameter $\underline{\pi}_{DF}$ by

$$\overline{\pi}^U_{DF}(\kappa) = \frac{\pi_\Delta}{\pi_{\text{CO}}} \psi^U_{\max}(\{G_1^+(\overline{y}), G_0^+(\overline{y})\}; \kappa) - 1.$$

We therefore note that just by construction the smoothed upper and lower bounds of $\underline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$ and $\overline{F}_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta)$ are Hadamard-differentiable in $(y, \pi_{\text{DF}}, \delta)$. Moreover, they are strictly increasing in δ if δ is small. Taking the inf is Hadamard-differentiable by Lemma 21.4 van der Vaart (1998b). A SLA is given by

$$\underline{\delta}^L(\pi_{\text{DF}}) = \psi^L_{\max}(\inf\{\delta : \inf_y \overline{F}^U_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) - \underline{F}^L_{Y_d^{CO}}(y, \pi_{\text{DF}}, \delta) \geq 0\}; \kappa).$$

A SUA of $\overline{\delta}(\pi_{\text{DF}})$ is given by

$$\overline{\delta}^U(\pi_{\text{DF}}; \kappa) = \psi^U_{\max}(\Big\{\psi^U_{\sup,\leq}(\overline{F}^U_{Y_d^{T_1}}(y, \pi_{\text{DF}}, 1), \overline{F}^U_{Y_d^{T_2}}(y, \pi_{\text{DF}}, 1))(\overline{y}; \kappa), \\ \psi^U_{\sup,\leq}(\underline{F}^U_{Y_d^{T_1}}(y, \pi_{\text{DF}}, 1), \underline{F}^U_{Y_d^{T_2}}(y, \pi_{\text{DF}}, 1))(\overline{y}; \kappa)\Big\}_{T_1,T_2 \in \{CO;DF\}}; \kappa).$$

Based on these definitions, we can smoothly bound the treatment effect of compliers, where the lower bound $\underline{\Delta}^L_{CO}(\pi_{\text{DF}}, \delta; \kappa)$ is given by

$$\underline{\Delta}^L_{CO}(\pi_{\text{DF}}, \delta; \kappa) = \int_{\mathbb{Y}} y \, d\overline{F}^U_{Y_1^{CO}}(y, \pi_{\text{DF}}, \delta) - \int_{\mathbb{Y}} y \, d\underline{F}^L_{Y_0^{CO}}(y, \pi_{\text{DF}}, \delta).$$

A SLA of the breakdown point BP^Lcan be derived by

$$BP^L(\pi_{\text{DF}}; \kappa) = \psi^L_{\min}\left(\sup\{\underline{\Delta}^L_{CO}(\pi_{\text{DF}}, \delta; \kappa) \geq \mu\}, \overline{\delta}(\pi_{\text{DF}}); \kappa\right).$$

Since $\Delta^L_{CO}(\pi_{\text{DF}}, \delta; \kappa)$ is strictly increasing in δ, $BP(\pi_{\text{DF}})$ is Hadamard-differentiable and as $\underline{\Delta}^L_{CO}(\pi_{\text{DF}}, \delta; \kappa)$ is Hadamard differentiable in both π_{DF} and δ and strictly increasing in δ. It follows that $\phi_L(\pi_{\text{DF}})$ is a SLA of $\phi(\pi_{\text{DF}})$, which concludes this subsection.

1.D. PROOFS OF ADDITIONAL RESULTS

1.D.1. **Proof of Proposition A.1.6.** We remind that $\theta_b = (P_{11}, P_{10}, P_{01}, P_{00}, P_1, P_0)$. It follows from standard central limit arguments that, under Assumption A.1.3, the estima-

tor $\widehat{\theta}_b$, the arithmetic mean of binary variables, satisfies that

$$\sqrt{n}(\widehat{\theta}_b - \theta_b) \to \mathcal{N}(0, \Sigma),$$

where we specify the elements of the variance-covariance matrix Σ in the following. Let Σ_{ij} denote the i-th row and j-th column element of Σ. $\Sigma_{ii} = \theta_{b,i}(1-\theta_{b,i})$ for $i \in \{1, \ldots, 6\}$; $\Sigma_{ij} = 0$, if $i \in \{1, 2, 5\}$ and $j \in \{3, 4, 6\}$; $\Sigma_{ij} = -\theta_{bi}\theta_{bj}$ for $i = 1$ and $j = 3$ or $i = 2$ and $j = 4$; and $\Sigma_{ij} = \theta_{bi}(1 - \theta_{bj})$ and for $i \in \{1, 3\}$ and $j = 5$, or $i \in \{2, 4\}$ and $j = 6$.

Using the arguments of Example 2.2 of Fang and Santos (2018) and Corollary 3.2 of Fang and Santos (2018), it follows that $\hat{\phi}'_{b,\theta}$ is a consistent estimator of $\hat{\phi}'_{b,\theta}$ when evaluated at finitely many π^k_{DF} so that the bootstrap procedure is consistent, as we have chosen a slightly larger value than $1 - \alpha$.

The final result of the Proposition then follows from applying the reasoning of Appendix 1.B.7.3.

1.D.2. **Proof of Proposition A.1.5.** Let A_y be the interval $[\underline{y}, y]$. We first consider the following parameters, that is similar to θ,

$$\tilde{\theta}(y) = (Q_{11}(y), Q_{10}(y), Q_{01}(y), Q_{00}(y), B_1(y), B_0(y)), \tag{A.1.38}$$

where $B_d(y) = \int_{C_{1d}(y)} q_{ss}(z)dz + \int_{C_{0d}(y)} q_{s(1-s)}(z)dz + o_P(1)$, and $C_{1d}(y) = \{z \in A_y : q_{dd}(z) < q_{d(1-d)}(z)\}$, $C_{0d}(y) = \{z \in A_y : q_{dd}(z) > q_{d(1-d)}(z)\}$. By Assumption A.1.4 (iii), we can consider the class of functions

$$\{\mathbf{1}\{Y \leq y\} \cdot \mathbf{1}\{Y \in C_{sd}\} \cdot \mathbf{1}\{D = d\} : y \in Y, d \in \{0, 1\}, s \in \{0, 1\}\},$$

Using Assumption (iii), it then follows by Example 19.6 van der Vaart (1998b) that $\mathcal{F}$ is Donsker. Based on this reasoning, using our assumptions, and using Theorem 4 in Giné and Nickl (2008), it then follows from simple variance calculations, that

$$\sqrt{n}(\widehat{\tilde{\theta}}(y) - \theta(y)) \to \mathcal{Z}_1(y),$$

$\mathcal{Z}_1$ is a tight Gaussian process with continuous paths in $\ell^\infty(\mathbb{R} \times \{0, 1\}, \mathbb{R}^6)$ with zero mean and covariance kernel given by the following expressions.[43] It holds that $[\Sigma(y, y')]_{ii} = \theta_i(y \wedge y') - \theta_i(y)\theta_i(y')$ for $i \in \{1, 2, 3, 4\}$, $[\Sigma(y, y')]_{24} = -\theta_2(y)\theta_4(y')$, and $[\Sigma(y, y')]_{13} = -\theta_2(y)\theta_4(y')$. For $i \in \{5, 6\}$, it holds that

$$[\Sigma(y, y')]_{ii} = \theta_{(i-2)}(y \wedge y') - \theta_{(i-2)}(y')\theta_{(i-2)}(y) + \theta_{(i-4)}(y \wedge y') - \theta_{(i-4)}(y')\theta_{(i-4)}(y);$$

[43] Let $\min(a, b) = a \wedge b$.

and for $i \in \{6\}$and $j \in \{2,4\}$ or for $i \in \{5\}$and $j \in \{1,3\}$

$$[\Sigma(y,y')]_{ij} = \theta_{s(1-s)}(y \wedge y') - p_{s(1-s)}(y)p_{s(1-s)}(y');$$

and otherwise $[\Sigma(y,y')]_{ij} = 0$.

It is left to show that

$$\sup_{y\in\mathbb{Y}} \sqrt{n}(\widehat{\theta}(y) - \widehat{\widetilde{\theta}}(y)) = o_P(1) \tag{A.1.39}$$

It suffices to analyze $\sup_{y\in\mathbb{Y}} \sqrt{n}\widehat{G}_d^+(y) - \widehat{B}_d(y) = o_P(1)$. We note that this results follows from using the same arguments used in the proof of Theorem 1 of Anderson et al. (2012). □

1.D.3. **Proof of Proposition A.1.6.** By construction, the mapping ϕ_κ is a composition of SLA by Lemma A.1.4 and Lemma A.1.5. It therefore follows that ϕ_κ is SLA of ϕ_κ by Lemma A.1.3. It then follows by the Delta method of Hadamard differentiable mappings that (see, e.g., Theorem 20.8 van der Vaart, 1998a) that $\sqrt{n}(\phi_\kappa(\theta,\pi_{\text{DF}}) - \phi_\kappa(\theta,\pi_{\text{DF}}))$ converges to a Gaussian distribution. Let

$$\text{cv}_{1-\alpha} = \inf\left\{ z \in \mathbb{R} : \mathbb{P}\left((\sup_{\pi_{\text{DF}}\in[0,0.5], l\leq 5} \sqrt{n} e_l(\phi_\kappa(\widehat{\theta},\pi_{\text{DF}}) - \phi_\kappa(\theta,\pi_{\text{DF}})) \leq z \right) \geq 1-\alpha \right\}$$

We can conclude that $\widehat{\text{cv}}_{1-\alpha} = \text{cv}_{1-\alpha} + o_p(1)$ as the nonparametric bootstrap is consistent for the mapping ϕ_κ by Dümbgen (1993) and Fang and Santos (2018). Based on this reasoning it follows that

$$\lim_{n\to\infty} \mathbb{P}(\max_{l\leq 5} \inf_{\pi_{\text{DF}}\in[0,0.5]} e_l^\top \left(\phi_{\kappa,L}(\widehat{\theta},\pi_{\text{DF}}) + \widehat{\text{cv}}_{1-\alpha} - \phi_{\kappa,L}(\theta,\pi_{\text{DF}})\right) \leq 0) \geq 1-\alpha.$$

It then follows by construction of ϕ_κ that $\phi_\kappa(\theta,\pi_{\text{DF}}) \leq \phi_\kappa(\theta,\pi_{\text{DF}})$ for each component and for all $\pi_{\text{DF}} \in [0,0.5]$, so that

$$\lim_{n\to\infty} \mathbb{P}(\max_{k\leq 5} \inf_{\pi_{\text{DF}}\in[0,0.5]} e_k^\top \left(\phi_{\kappa,L}(\widehat{\theta},\pi_{\text{DF}}) + \widehat{\text{cv}}_{1-\alpha} - \phi(\theta,\pi_{\text{DF}})\right) \leq 0) \geq 1-\alpha.$$

Using $\phi_{\kappa,L}(\widehat{\theta},\pi_{\text{DF}}) = \phi_{\kappa,L}(\widehat{\theta},\pi_{\text{DF}}) + \widetilde{\text{cv}}_{1-\alpha}$ as mappings to construct both $\widehat{SR}_L$ and $\widehat{RR}_L$ then yields in turn by the reasoning of Appendix 1.B.7.3 that

$$\lim_{n\to\infty} \mathbb{P}(\widehat{RR}_{\kappa,L} \subseteq RR,\ \text{SR} \subseteq \widehat{\text{SR}}_{\kappa,L}) \geq 1-\alpha,$$

which concludes this proof. □

1.D.4. **Proof of Lemma A.1.3.**

1. It holds by monotonicity that $\phi^U(\psi^U(f,\kappa),\kappa) \leq \phi^U(\psi(f,\kappa),\kappa) \leq \phi(\psi(f,\kappa))$

2. It is clear that if $|\phi^U(\psi^U(f,\kappa),\kappa) - \phi(\psi(f,\kappa))| \leq |\phi^U(\psi^U(f,\kappa),\kappa) - \phi(\psi^U(f,\kappa),\kappa)| + |\phi(\psi^U(f,\kappa),\kappa) - \phi(\psi(f,\kappa))|$ where the right hand side of the equation can be made uniformly arbitrarily small for κ large enough by assumption by continuity of the mapping

3. It follows from the chain-rule of HD mappings (Theorem 20.9 van der Vaart, 1998b) □

1.D.5. **Proof of Lemma A.1.4.** We first consider ψ^U_{av}.

1. Trivial, as $|f(y)| = \sqrt{f(y)}^2 \leq \sqrt{f(y) + 1/\kappa}^2$ for all $y \in \mathbb{Y}$ and any $f \in l^\infty(\mathbb{Y})$.
2. It is also clear that $\sqrt{f(y) + 1/\kappa}^2 \to |f(y)|$ uniformly for all $x \in \mathbb{R}$ as $\kappa \to \infty$.
3. ψ^U_{av} is HD as $(\psi^U_{av}(f;\kappa))'(y) = (f(y)^2 + \frac{1}{\kappa})^{-1/2} \cdot f'(y)$ and $(f(y)^2 + \frac{1}{\kappa}) \geq \frac{1}{\kappa} \geq 0$.

ψ^L_{av} satisfies the above criterion by similar arguments. □

1.D.6. **Proof of Lemma A.1.5.** We first consider $\psi^L_{\text{sup},\leq}$.

1. Follows immediately, as for any $y \in \mathbb{Y}$ and any $\kappa \in \mathbb{N}$

$$\psi^L_{\text{sup},\leq}(f,g;\kappa)(y) \leq \max(\{g(k_j) - f(k_j)\}_{j:k_j \leq y}) = \psi_{\text{sup},\leq}(f,g).$$

The second inequality follows as g and f are nondecreasing.

2. It follows that for $\kappa \to \infty$, $\psi^L_{\text{sup},\leq}(f,g;\kappa) \to \psi_{\text{sup},\leq}(f,g)$

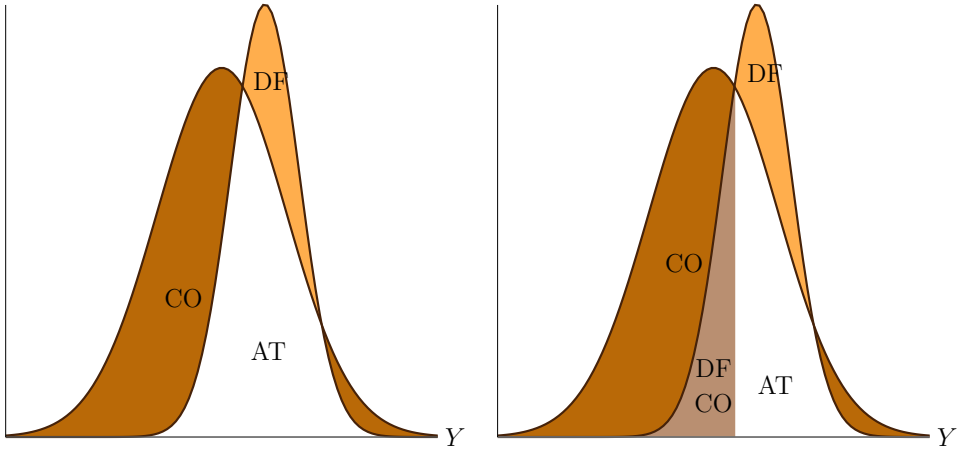
3. HD follows from the chain rule of HD functions (Theorem 20.9 van der Vaart, 1998b) and as the difference is a linear operator.

Similar arguments apply to the other mappings, which concludes this proof. □

1.E. FURTHER ILLUSTRATIONS

We now give intuitive explanations on how the bounds on the distribution function $F_{Y_d^{CO}}$ are derived and how the sensitivity parameter δ is bounded. For simplicity, we just consider the presence of treatment in both cases.

1.E.1. **Illustration of Derivation of Bounds on Outcome Distributions.** In Figure A.1.1, we give some intuition on how the outcome distribution of compliers is constructed. We plot the functions q_{11} and q_{10}. Based on the reasoning of the main text, the function q_{11} is a weighted average of the densities of $f_{Y_1^{CO}}$ and $f_{Y_1^{AT}}$, and the function q_{10}



(a) If $\pi_{\text{DF}} = \underline{\pi}_{DF}$ (b) If $\pi_{\text{DF}} > \underline{\pi}_{DF}$

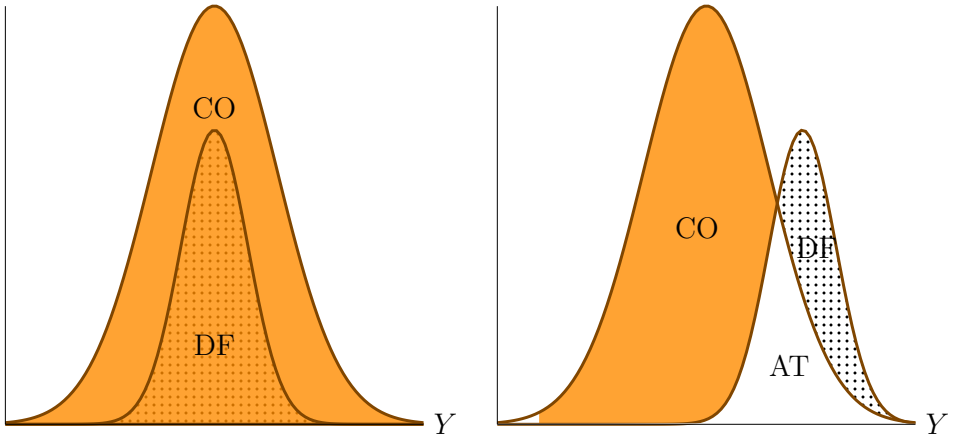
Figure A.1.1: Derivation of the compliers outcome distributions

of the densities of $f_{Y_1^{DF}}$ and $f_{Y_1^{AT}}$. It is clear that to be a feasible candidate of $f_{Y_1^{CO}}$, any density has to satisfy that

$$\frac{1}{\pi_\Delta} \max\{0, q_{11}(y) - q_{10}(y)\} \leq f_{Y_1^{CO}}(y) \leq \frac{1}{\pi_{\text{CO}}} q_{11}(y).$$

In Figure A.1.1 (a), the density of $f_{Y_1^{CO}}$ is point identified for the sensitivity parameter π_{DF} that is the smallest when ignoring the distribution functions in the absence of treatment. However, if π_{DF} increases the density of $f_{Y_1^{CO}}$ is in general not point identified. The corresponding probability mass of the tails of the function $\min\{q_{11}(y), q_{10}(y)\}$ is then imputed to belong to the compliers and defiers. Figure A.1.1 (b) gives such an example for a possible candidate of density function of $f_{Y_1^{CO}}$ implying an upper bound on the distribution function of compliers.

1.E.2. **Intuition for Lower and Upper Bound on Outcome Heterogeneity.** We give some intuition on how the bounds on the sensitivity parameters δ are derived. Let us first consider the largest value $\overline{\pi}_{DF}$ ignoring the distribution functions in the absence of treatment. In Figure A.1.2 (a), this value implies that both the outcome distributions of compliers and defiers are point identified, as the population size of always takers would be zero. Thus the outcome distribution function of defiers equals $Q_{10}(y)$, and of compliers equals $Q_{11}(y)$ up to normalization. In this specific example, the outcome heterogeneity would be point identified but especially bounded from above by 0.5. In Figure A.1.2 (b), we consider the smallest possible value of outcome heterogeneity $\pi_{\text{DF}} = \underline{\pi}_{DF}$. The two outcome distributions are again point identified, and the outcome heterogeneity would be close to one, but especially it would be bounded from below. A similar reasoning then also applies to the absence of treatment.



(a) above for large values of π_{DF} (b) below for small values of π_{DF}.

Figure A.1.2: Illustration of sensitivity region.

Chapter 2

Bias-Aware Inference in Fuzzy Regression Discontinuity Designs

with Christoph Rothe

2.1. INTRODUCTION

The regression discontinuity design is a popular empirical strategy for estimating causal treatment effects from observational data. In sharp (SRD) designs units receive a treatment if and only if a running variable falls above a known cutoff value, whereas in fuzzy (FRD) designs the treatment probability jumps at the threshold, but generally not from zero to one. Methods for estimation and inference based on local linear regression are widely used in empirical research for both kinds of designs, and their theoretical properties have been studied extensively; see Imbens and Lemieux (2008) or Lee and Lemieux (2010) for surveys, and Cattaneo et al. (2019) for a textbook treatment.

A key issue for SRD confidence intervals (CIs) is the handling of the estimator's smoothing bias, with undersmoothing (cf. Imbens and Lemieux, 2008) and robust bias correction (Calonico et al., 2014) being popular approaches in applications. However, Armstrong and Kolesár (2020) show that common implementations of such CIs can have coverage issues in practice, mostly due to the way they select the bandwidth,[1] and that "bias-aware" CIs, which adjust the critical value to take possible bias into account, are more efficient than their counterparts based on either undersmoothing or robust bias correction, even at infeasible bandwidths. A further advantage of bias-aware SRD CIs relative to these alternatives is that they do not require a continuously distributed running variable.

In an FRD design, the usual point estimator is the ratio of two SRD estimators, and due to this nonlinearity one cannot directly use the same bias-handling techniques as in SRD setups. The CIs reported in empirical FRD papers therefore typically build

[1] Both methods typically take an estimate of a pointwise-MSE-optimal bandwidth (Imbens and Kalyanaraman, 2012) as an input. This bandwidth can be large even if the underlying function is highly nonlinear, which then leads to large smoothing biases in finite samples. Estimators of this bandwidth generally involve a regularization step to prevent extreme values, the result can depend critically on tuning parameters that are difficult to pick (Armstrong and Kolesár, 2020).

on a delta method (DM) argument. This entails approximating the FRD estimator with a term that behaves like an SRD estimator, imposing conditions under which the corresponding error is negligible in large samples, and applying an SRD bias-handling approach to the leading term. Proceeding like this can exasperate the practical issues of undersmoothing and robust bias correction known from SRD contexts; and it can also create problems for the bias-aware approach, as bias-aware FRD DM CIs only account for an approximate bias. Moreover, any type of DM CI can only be asymptotically valid if the running variable is continuous with positive density around the cutoff, and the jump in treatment probabilities at the cutoff is "large". DM CIs generally break down in empirical settings that do not exhibit these properties, in the sense that their actual coverage can deviate substantially from the nominal level; and they cannot be salvaged by adjusting the method used to control the bias. This is important because empirical researchers often face running variables that take only a limited number of distinct values, like test scores or class sizes (Angrist and Lavy, 1999; Oreopoulos, 2006; Urquiola and Verhoogen, 2009; Fredriksson et al., 2013; Clark and Martorell, 2014; Hinnerich and Pettersson-Lidbom, 2014; Card and Giuliano, 2016; Jepsen et al., 2016); "donut designs" that exclude units close to the cutoff to increase the credibility of causal estimates (Almond and Doyle, 2011; Dahl et al., 2014; Dube et al., 2019; Le Barbanchon et al., 2019; Scott-Clayton and Zafar, 2019); or weakly identified setups with small jumps in treatment probabilities (Malenko and Shen, 2016; Coviello et al., 2018).

In this chapter, we propose new confidence sets (CSs) for the FRD parameter that are not subject to such shortcomings. Our CSs avoid the use of the FRD point estimator, and are instead based on auxiliary statistics that can be computed directly via local linear regression. The construction avoids the approximation errors of the DM, and is somewhat analogous to that of an Anderson-Rubin (AR) statistic in an exactly identified linear instrumental variable model (Staiger and Stock, 1997). We then apply the bias-aware approach to these statistics, which allows us to account exactly for the possible smoothing bias. The resulting CSs are easy to compute; an `R` package is available on the authors' website.

We derive two main results under the common assumption that the second derivatives of the conditional expected outcome and the conditional treatment probability are bounded by some constant on either side of the cutoff. First, we show that our CSs are honest in the sense of Li (1989), meaning that they have correct asymptotic coverage uniformly over the class of functions satisfying our assumption, irrespective of the distribution of the running variable or the strength of identification. This property implies good CS performance across the entire range of plausible data generating processes, and

is thus necessary for good finite-sample coverage. The novel insight here is not so much that AR CSs can accommodate weak identification, but that combining this construction with a bias-aware approach provides robustness to other deviations from the canonical setup, like discreteness of the running variable and "donut" designs.[2]

Second, we show that bias-aware AR CSs are asymptotically equivalent to bias-aware DM CIs if the running variable is continuous and identification is strong, which are conditions needed for DM CIs to be honest in the first place. The robustness of bias-aware AR CSs does thus not come with a cost in terms of power relative to DM CIs in a canonical setup. Moreover, since Armstrong and Kolesár (2020) show that bias-aware DM CIs outperform DM CIs based on undersmoothing and robust bias correction, the equivalence result implies that the same is true for our bias-aware AR CSs. These predictions are confirmed by simulation results reported in this chapter.

We also make three contributions regarding the implementation of bias-aware inference that are not only important for our CSs, but can also be used more generally. First, we provide a new standard error for local linear regression estimates that is uniformly consistent over the class of functions with bounded second derivatives. It is a variation of the nearest-neighbor variance estimator (e.g. Abadie and Imbens, 2006) commonly used in the RD literature. Our proposal replaces the usual local average with a local linear projection among the nearest neighbors, which removes a bias term that is proportional to the underlying function's first derivative. Second, we propose a new empirical bandwidth that enforces an upper bound on Lindeberg weights to ensure that a normal approximation works well for our local linear estimates in finite samples. Third, we provide new graphical tools and an analysis of "rules of thumb" that can help guide the choice of the bounds on second derivatives, which are the main tuning parameters required for bias-aware inference.

As an extension, we also derive new bias-aware CSs for the fuzzy regression kink design (Card et al., 2015), and establish theoretical properties analogous to those we obtain for the FRD case. These results also apply more generally to settings in which the parameter of interest is the ratio of jumps in the vth-order derivatives of two conditional expectation functions at some threshold value.

Our paper contributes to a growing literature on "bias-aware" inference. Building

[2]Feir et al. (2016) already showed that undersmoothing AR CSs can have correct pointwise asymptotic coverage if the jump in treatment probabilities tends to zero with the sample size at an appropriate rate, while undersmoothing DM CIs generally do not have this property. Such CIs require a continuous running variable with positive density around the cutoff, and may, depending on the implementation of undersmoothing, not be honest. After circulating the first draft of the paper, we were also made aware that Huang and Zhan (2020) have discussed combining a bias-aware approach with an AR-type statistic. Since they misinterpret the results from Armstrong and Kolesár (2020), however, their proposed methods do not yield valid inference.

on classical work (e.g. Sacks and Ylvisaker, 1978; Donoho, 1994), such methods, which take bias explicitly into account rather than trying to remove it, have recently been shown to yield powerful and practical CSs in a wide range of non- and semiparametric problems (Armstrong and Kolesár, 2018, 2021a, 2020; Kolesár and Rothe, 2018; Imbens and Wager, 2019; Ignatiadis and Wager, 2020; Roth and Rambachan, 2019; Schennach, 2020; Armstrong et al., 2020). A concern sometimes raised with these methods is that, in contrast to traditional approaches such as undersmoothing or robust bias correction, they require specifying explicit bounds on the smoothness of the underlying functions. However, this view neglects such bounds are implicitly required for traditional methods to work well in practice.[3] Following the literature on bias-aware inference, we recommend to vary the values of smoothness bounds in the construction of our CS in empirical practice as a form of sensitivity analysis. We also provide a number of tools to guide and communicate the choices.

The remainder of this chapter is structured as follows. Section 2 describes our setup. Section 3 describes existing approaches to SRD and FRD inference, and discusses issues with DM CIs. Section 4 describes our bias-aware AR CSs, and Section 5 establishes their theoretical properties. Section 6 discusses implementation issues. Section 7 contains a simulation study, and Section 8 an empirical application. Section 9 concludes. The appendix contains the proofs of our main theorems. Further technical arguments, extensions and additional materials are given in the online appendix.

2.2. SETUP AND PRELIMINARIES

2.2.1. **Fuzzy RD Designs.** Let $Y_i \in \mathbb{R}$ be the outcome, $T_i \in \{0,1\}$ be the actual treatment status, $Z_i \in \{0,1\}$ be the assigned treatment, and $X_i \in \mathbb{R}$ be the running variable of the ith unit in a random sample of size n from a large population. Treatment is assigned if the running variable falls above a known cutoff. We normalize this threshold to zero, so that $Z_i = \mathbf{1}\{X_i \geq 0\}$. Because of limited compliance, it could be that $Z_i \neq T_i$. For a generic random variable W_i (which could be equal to Y_i or T_i, for example), we then write $\mu_W(x) = \mathbb{E}(W_i|X_i = x)$ for its conditional expectation function given the running variable; $\mu_{W+} = \lim_{x\downarrow 0} \mu_W(x)$ and $\mu_{W-} = \lim_{x\uparrow 0} \mu_W(x)$ for its right and left limit at zero; and $\tau_W = \mu_{W+} - \mu_{W-}$ for the jump in μ_W at the cutoff. The parameter of interest is

$$\theta = \frac{\tau_Y}{\tau_T},$$

[3]For example, in order for standard implementations of robust bias correction SRD CIs to have approximately correct coverage in finite samples, one must have a "sufficiently small" bound on the underlying function's third derivative (Kamat, 2018). Researchers that report such CIs and consider them reliable thus implicitly impose a smoothness bound. Moreover, if that bound was made explicit, a more efficient CI could be constructed through a bias-aware approach (Armstrong and Kolesár, 2020).

which, in a potential outcomes framework with certain continuity and monotonicity conditions (e.g. Hahn et al., 2001; Dong, 2018), has a causal interpretation as the local average treatment effect among "compliers" at the cutoff, where "compliers" are units whose treatment decision is affected by the assignment rule (Imbens and Angrist, 1994).

2.2.2. **Honest Confidence Sets.** Our goal is to construct confidence sets (CSs) that cover the parameter θ in large samples with at least some pre-specified probability, uniformly over (μ_Y, μ_T) in some function class $\mathcal{F}$ that embodies shape restrictions that the analyst is willing to impose. That is´, we want to construct data-dependent sets $\mathcal{C}^\alpha \subset \mathbb{R}$ that satisfy

$$\liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1-\alpha \tag{2.1}$$

for some $\alpha > 0$.[4] Following Li (1989), we refer to such CSs as *honest with respect to* $\mathcal{F}$. This is a much stronger requirement than correct pointwise asymptotic coverage:

$$\liminf_{n\to\infty} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1-\alpha \text{ for all } (\mu_Y, \mu_T) \in \mathcal{F}. \tag{2.2}$$

In particular, under (2.1) we can always find a sample size n such that the coverage probability of $\mathcal{C}^\alpha$ is not below $1-\alpha$ by more than an arbitrarily small amount for every $(\mu_Y, \mu_T) \in \mathcal{F}$. Under (2.2) there is no such guarantee, and even in very large samples the coverage probability of $\mathcal{C}^\alpha$ could be poor for some $(\mu_Y, \mu_T) \in \mathcal{F}$. Since we do not know in advance which function pair is the correct one, honesty as in (2.1) is necessary for good finite sample coverage of $\mathcal{C}^\alpha$ across data generating processes. Of course, we also want CSs that are efficient, in the sense that they are "small" while maintaining honesty.

2.2.3. **Smoothness Conditions.** Following Armstrong and Kolesár (2018, 2020), we specify the class $\mathcal{F}$ of plausible candidates for (μ_Y, μ_T) as a smoothness class. Specifically, let

$$\mathcal{F}_H(B) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_w''\|_\infty \leq B, w = 0, 1\}$$

be the Hölder-type class of real functions that are potentially discontinuous at zero, are twice differentiable almost everywhere on either side of the threshold, and have second derivatives uniformly bounded by some constant $B > 0$; and let

$$\mathcal{F}_H^\delta(B) = \{f \in \mathcal{F}_H(B) : |f_+ - f_-| > \delta\},$$

[4]Note that we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on μ_Y and μ_T implicit in our notation. Each function pair (μ_Y, μ_T) corresponds to a single distribution of $(Y, T, X, Z) = (\mu_Y(X) + \epsilon_M, \mathbf{1}\{\mu_T(X) \geq \epsilon_T\}, X, Z)$, where (ϵ_M, ϵ_T) is some fixed random vector.

for some $\delta \geq 0$, be a similar class of functions whose discontinuity at zero exceeds δ in absolute magnitude. We then assume that

$$(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^0(B_T) \equiv \mathcal{F}, \tag{2.3}$$

for some constants B_Y and B_T whose choice in empirical practice we discuss in Section 2.6.4. Note that in addition to imposing smoothness, condition (2.3) also rules out cross-restrictions between the shapes of μ_Y and μ_T, since $\mathcal{F}$ is a Cartesian product. This seems reasonable for applications in economics. Also note that we impose $\mu_T \in \mathcal{F}_H^0(B_T)$, and thus that $\tau_T \neq 0$, only to ensure that the parameter of interest $\theta = \tau_Y/\tau_T$ is well-defined. Our setup explicitly allows τ_T to be arbitrarily close to zero.

2.2.4. **Discrete Settings.** Conditional expectation functions are only well-defined over the support of the conditioning variable. One must therefore clarify the meaning of (2.3) if X_i is discrete, or more generally such that there are gaps in its support. Following Kolesár and Rothe (2018) and Imbens and Wager (2019), we understand this condition to mean that there exists a single "true" function pair $(\mu_Y, \mu_T) \in \mathcal{F}$ such that $(\mu_Y(X_i), \mu_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i))$ with probability 1. This pair is then obviously point identified on the support of the running variable, and partially identified everywhere else through the shape restrictions implied by it being an element of $\mathcal{F}$. This reasoning further implies that θ must be contained in the identified set

$$\Theta_I = \left\{ \frac{m_{Y+} - m_{Y-}}{m_{T+} - m_{T-}} : (m_Y, m_T) \in \mathcal{F},\ (m_Y(X_i), m_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i)) \text{ w.p.1} \right\}.$$

This set is a singleton if X_i is supported on an open neighborhood around the cutoff, but generally it is either (i) a closed interval $[a_1, a_2]$; (ii) the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$; (iii) the entire real line; or, as a knife-edge case (iv) a half-line $[a_1, \infty)$ or $(-\infty, -a_1]$, with $a_1 > 0$. This holds because the range of $(m_{Y+} - m_{Y-}, m_{T+} - m_{T-})$ over $(m_Y, m_T) \in \mathcal{F}$ is a Cartesian product of two intervals $I_Y \times I_T$. The four cases then obtain depending on which of these two intervals contain zero, possibly as a boundary value.

Note that while it is not possible to consistently estimate either τ_Y, τ_T, or θ if Θ_I is not a singleton, inference is not futile in such cases. Indeed, our CSs described below are valid in the sense of Imbens and Manski (2004) irrespective of whether θ is point or partially identified, and without applied researchers having to decide which of the two notions of identification more accurately describes their particular setting.

2.2.5. **Local Linear Estimation.** Local linear regression (Fan and Gijbels, 1996) is arguably the most popular empirical strategy for estimation and inference in RD designs. Formally, for a generic dependent variable W_i (which could be equal to Y_i or T_i, for example), the local linear estimator of the jump $\tau_W = \mu_{W+} - \mu_{W-}$ is

$$\widehat{\tau}_W(h) = e_1^\top \arg\min_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K(X_i/h)(W_i - \beta'(Z_i, X_i, Z_i X_i, 1))^2, \tag{2.4}$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$, $h > 0$ is a bandwidth, and $e_1 = (1, 0, 0, 0)'$ is the first unit vector. The natural point estimator of θ is then given by $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$, for some value of h. A key feature of $\widehat{\tau}_W(h)$ is that it can be written as a weighted average of the W_i, with weights $w_i(h)$ that depend on the data through the realizations $\mathcal{X}_n = (X_1, \ldots, X_n)'$ of the running variable only:

$$\widehat{\tau}_W(h) = \sum_{i=1}^{n} w_i(h) W_i.$$

The exact form of the weights follows from standard least squares algebra, and is given explicitly in Appendix 2.A. Estimators of the form (2.4) are the building blocks of our honest CSs described below, and we refer to $\widehat{\tau}_W(h)$ as an SRD-type estimator of τ_W in the following, as it is the conventional estimator in a hypothetical SRD design with outcome W_i.

2.3. EXISTING METHODS FOR RD INFERENCE

2.3.1. **SRD Inference.** We first review some techniques for inference based on SRD-type estimators, which are by now well-understood. To describe the bias-aware SRD CIs of Armstrong and Kolesár (2018, 2020), let $b_W(h)$ and $s_W(h)$ denote the bias and standard deviation, respectively, of a generic SRD-type estimator $\widehat{\tau}_W(h)$ conditional on the realizations of the running variable; and let $\widehat{s}_W(h)$ be a standard error. Under mild conditions, the large sample distribution of the t-ratio $(\widehat{\tau}_W(h) - \tau_W)/\widehat{s}_W(h)$ is then that of the sum of a standard normal random variable and the ratio $b_W(h)/\widehat{s}_W(h)$. While the latter is unknown in practice, a bound $\widehat{r}_W(h) = (\sup_{\mu_W \in \mathcal{F}_H(B_W)} |b_W(h)|)/\widehat{s}_W(h)$ on $|b_W(h)/\widehat{s}_W(h)|$ can be calculated explicitly. One can then construct the bias-aware CI

$$\mathcal{C}_W^\alpha = [\widehat{\tau}_W(h) \pm \mathrm{cv}_{1-\alpha}(\widehat{r}_W(h))\widehat{s}_W(h)],$$

where the critical value $\mathrm{cv}_{1-\alpha}(r)$ is the $(1-\alpha)$-quantile of $|N(r, 1)|$, the distribution of the absolute value of a normal random variable with mean r and unit variance. Armstrong and Kolesár (2018, 2020) show that this CI is honest with respect to $\mathcal{F}_H(B_W)$ irrespec-

tive of the distribution of the running variable, valid for any bandwidth (for which the quantities involved in its construction are well-defined), and highly efficient if the running variable is continuous and the bandwidth is chosen to minimize the length of $\mathcal{C}_W^\alpha$.

Other popular approaches to SRD inference include undersmoothing, or using a "small" bandwidth for which the "bias to standard error" ratio is asymptotically negligible (cf. Imbens and Lemieux, 2008); and robust bias correction, which involves subtracting a bias estimate from $\widehat{\tau}_W(h)$, and adjusting the standard error (Calonico et al., 2014). In either case, CIs are formed with the usual critical value $\mathrm{cv}_{1-\alpha}(0)$. Both approaches assume a continuously distributed running variable, but Armstrong and Kolesár (2020) show that common implementations of undersmoothing and robust bias correction can still have finite-sample issues in such settings. One reason is that both methods typically take an estimate of a pointwise-MSE-optimal bandwidth (Imbens and Kalyanaraman, 2012) as an input. This bandwidth can be very large even if the underlying function is highly nonlinear, which then leads to large smoothing biases in finite samples. While estimators of the pointwise-MSE-optimal bandwidth generally involve a regularization step to prevent extreme bandwidth values, in practice the result is often still unstable and depends critically on the values of tuning parameters, which are difficult to pick. Armstrong and Kolesár (2020) also show that undersmoothing and robust bias correction CIs are inefficient, in that they tend to be much longer than bias-aware counterparts, even with infeasible bandwidths.

2.3.2. **Delta Method FRD Inference.** The above mentioned methods for SRD inference critically rely on the "weighted average" representation of local linear regression estimators. Since the FRD estimator $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$ is a nonlinear transformation of two SRD-type estimators, such methods cannot simply be applied directly. Instead, the CIs commonly reported in empirical FRD studies are based on a "delta method" (DM) argument. From a simple Taylor expansion, it follows that $\widehat{\theta}(h) - \theta$ can be written as the sum of an SRD-type estimator $\widehat{\tau}_U(h)$ as in (2.4), with an unobserved dependent variable U_i, and a remainder $\widehat{\rho}(h)$:

$$\widehat{\theta}(h) - \theta = \widehat{\tau}_U(h) + \widehat{\rho}(h), \quad \widehat{\tau}_U(h) = \sum_{i=1}^n w_i(h) U_i, \quad U_i = \frac{Y_i - \tau_Y}{\tau_T} - \frac{\tau_Y(T_i - \tau_T)}{\tau_T^2},$$
$$\widehat{\rho}(h) = \frac{\widehat{\tau}_Y(h)(\widehat{\tau}_T(h) - \tau_T)^2}{2\widehat{\tau}_T^*(h)^3} - \frac{(\widehat{\tau}_Y(h) - \tau_Y)(\widehat{\tau}_T(h) - \tau_T)}{\tau_T^2},$$

with $\widehat{\tau}_T^*(h)$ an intermediate value between τ_T and $\widehat{\tau}_T(h)$. With DM inference, one then imposes regularity and bandwidth conditions under which $\widehat{\rho}(h)$ is an asymptotically negligible relative to $\widehat{\tau}_U(h)$, and forms a CI for θ by applying some method for SRD inference

to $\widehat{\tau}_U(h)$. Since U_i is unobserved, any such method is must be made feasible by using an estimate $\widehat{U}_i$ in which τ_Y and τ_T are replaced by suitable preliminary estimators. Versions of such CIs are proposed, for example, by Calonico et al. (2014) and Armstrong and Kolesár (2020) in combination with robust bias correction and a bias-aware approach, respectively.[5]

An obvious downside of such constructions, to which we refer as DM CIs, is that they only control the bias of a first-order approximation of $\widehat{\theta}(h)$, and not the bias of $\widehat{\theta}(h)$ itself. Moreover, replacing U_i with an estimate $\widehat{U}_i$ introduces additional uncertainties in finite samples. In practice, all DM FRD CIs are thus subject to additional distortions relative to conventional SRD CIs. A more principal, and more practically important issue with DM CIs is that the central condition for their validity, namely that $\widehat{\rho}(h)$ is asymptotically negligible relative to $\widehat{\tau}_U(h)$, is not innocuous. In particular, this condition is not compatible with a discrete running variable, or more generally one with support gaps around the cutoff.

To see this last point, recall from Section 2.2.4 that consistent estimation of τ_T and τ_Y is generally not possible with a discrete running variable. The terms $\widehat{\tau}_U(h)$ and $\widehat{\rho}(h)$ therefore have non-zero probability limits in this case, and $\widehat{\rho}(h)$ cannot be ignored for the purpose of inference on θ. This issue occurs irrespective of the method chosen to control the bias of $\widehat{\tau}_U(h)$, including bias-aware inference. Since running variables with discrete or irregular support are ubiquitous in practice, this is an important limitation.

Another issue for DM CIs is that the conditions for their validity rule out weakly identified settings with τ_T close to zero. This issue occurs even if the running variable is continuously distributed. To see this, note that for any DM CI to be honest with respect to $\mathcal{F}$, the term $\widehat{\rho}(h)$ must be of smaller order than $\widehat{\tau}_U(h)$ not only at the "true" function pair (μ_Y, μ_T), but uniformly over all $(\mu_Y, \mu_T) \in \mathcal{F}$. But since τ_T can be arbitrarily close to zero over $(\mu_Y, \mu_T) \in \mathcal{F}$, we have that $\sup_{\mu_Y, \mu_T} |\widehat{\rho}(h)| = \infty$, which means that DM CIs break down.[6]

[5] In empirical papers, FRD estimates are sometimes obtained through the two-stage least squares regression $Y_i = \theta T_i + \beta_+ X_i Z_i + \beta_- X_i (1 - Z_i) + \varepsilon_i$ with Z_i as an instrument for T_i, using only data in some window around the cutoff. This is numerically equivalent to a ratio of local linear regressions with a uniform kernel, and the resulting CI is thus of the DM type (Hahn et al., 2001; Imbens and Lemieux, 2008).

[6] Feir et al. (2016) also point out covarage issues of DM CIs under weak identification, although through a different technical argument. Specifically, they show that DM CIs based on infeasible "undersmoothing" bandwidths do not have correct asymptotic coverage under pointwise asymptotics when τ_T tends to zero with the sample size at an appropriate rate.

We propose an alternative approach to FRD inference that avoids the inherent shortcomings of DM CIs by directly considering an object that can be estimated by an SRD-type estimator. We define the "auxiliary" parameter $\tau_M(c) = \tau_Y - c\tau_T$, which can be written as

$$\tau_M(c) = \mu_{M+}(c) - \mu_{M-}(c), \quad \mu_M(x,c) = \mathbb{E}(M_i(c)|X_i = x), \quad M_i(c) = Y_i - cT_i.$$

That is, $\tau_M(c)$ is the jump in the conditional expectation $\mu_M(x,c)$ of the constructed outcome $M_i(c)$ given the running variable X_i at the cutoff $x = 0$. We can form a bias-aware CI for $\tau_M(c)$ based on the SRD-type estimator $\widehat{\tau}_M(h,c)$, which is as in (2.4) but with $M_i(c)$ replacing W_i, and a bandwidth that might depend on c. Note that to keep the notation simple, the estimator $\widehat{\tau}_M(h,c) = \widehat{\tau}_Y(h) - c\widehat{\tau}_T(h)$ uses the same bandwidth on each side of the cutoff, and also the same bandwidth for estimating τ_Y and τ_T. It is straightforward to accommodate more general bandwidth choices; see Online Appendix 2.B for details.

Our CS for the actual parameter of interest θ is then obtained by collecting all values of c for which the "auxiliary" CI contains zero:

$$\mathcal{C}^{\alpha}_{\text{ar}} = \{c \in \mathbb{R} : \text{a } (1-\alpha) \text{ bias-aware CI for } \tau_M(c) \text{ contains } 0\}. \tag{2.5}$$

This construction shares similarities with that of Anderson and Rubin (1949) for inference in exactly identified linear IV models, and Fieller (1954) for inference on ratios. Emphasizing the former connection, we refer to such CSs as bias-aware AR CSs for θ.

To describe the approach in more detail, recall the notation from Section 2.2.5 and denote the conditional bias and standard deviation of $\widehat{\tau}_M(h,c) = \sum_{i=1}^n w_i(h)M_i(c)$ given $\mathcal{X}_n = (X_1, \ldots, X_n)'$ by $b_M(h,c) = \mathbb{E}(\widehat{\tau}_M(h,c)|\mathcal{X}_n) - \tau_M(c)$ and $s_M(h,c) = \mathbb{V}(\widehat{\tau}_M(h,c)|\mathcal{X}_n)^{1/2}$, respectively. These quantities can be written more explicitly as

$$b_M(h,c) = \sum_{i=1}^n w_i(h)\mu_M(X_i,c) - (\mu_{M+}(c) - \mu_{M-}(c)),$$
$$s_M(h,c) = \left(\sum_{i=1}^n w_i(h)^2 \sigma^2_{M,i}(c)\right)^{1/2},$$

with $\sigma^2_{M,i}(c) = \mathbb{V}(M_i(c)|X_i)$ the conditional variance of $M_i(c)$ given X_i. The bias depends on (μ_Y, μ_T) through the transformation $\mu_M = \mu_Y - c \cdot \mu_T$ only, and $\mu_Y - c \cdot \mu_T \in \mathcal{F}_H(B_Y + |c|B_T)$ by (2.3) and linearity of the second derivatives operator. Following Armstrong and Kolesár (2020), we can bound $b_M(h,c)$ in absolute value over the functions

contained in $\mathcal{F}$, for any value of the bandwidth h, by

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_M(h,c)| \leq \bar{b}_M(h,c) \equiv -\frac{B_Y + |c|B_T}{2} \cdot \sum_{i=1}^{n} w_i(h) X_i^2 \cdot \text{sign}(X_i),$$

with the supremum being achieved by a pair of piecewise quadratic functions with second derivatives equal to $(B_Y \cdot \text{sign}(x), B_T \cdot \text{sign}(x))$ over $x \in [-h, h]$.[7] Under standard regularity conditions, the statistic

$$\frac{\widehat{\tau}_M(h,c) - \tau_M(c)}{s_M(h,c)} = \frac{\widehat{\tau}_M(h,c) - \tau_M(c) - b_M(h,c)}{s_M(h,c)} + \frac{b_M(h,c)}{s_M(h,c)}$$

is then the sum of a term that is approximately standard normal in large samples conditional on $\mathcal{X}_n$, and a term that is bounded in absolute value by $r_M(h,c) = \bar{b}_M(h,c)/s_M(h,c)$, the "worst case" bias to standard deviation ratio. For every $c \in \mathbb{R}$ we can thus construct an (infeasible) auxiliary bias-aware CI for the pseudo parameter $\tau_M(c)$ as $C_M^\alpha(h,c) = [\widehat{\tau}_M(h,c) \pm \text{cv}_{1-\alpha}(r_M(h,c)) s_M(h,c)]$, where $\text{cv}_{1-\alpha}(r)$ is again the $(1-\alpha)$-quantile of the $|N(r,1)|$ distribution. Since the construction of this CI is conditional on the realizations of the running variable, it is valid irrespective of whether the distribution of the latter is continuous or discrete; and since it takes into account the exact conditional bias, it is also valid for any choice of bandwidth $h = h(c)$, including fixed ones that do not depend on the sample size. Its asymptotic length is minimized by

$$h_M(c) = \arg\min_h \text{cv}_{1-\alpha}(r_M(h,c)) s_M(h,c).$$

Following the idea from (2.5), an efficient infeasible CS for θ is then given by the collection of all values of c for which the auxiliary CI $C_M^\alpha(h,c)$, evaluated at $h_M(c)$, contains zero:

$$\mathcal{C}_*^\alpha = \left\{ c : |\widehat{\tau}_M(h_M(c),c)| \leq \text{cv}_{1-\alpha}(r_M(h_M(c),c)) s_M(h_M(c),c)) \right\}. \tag{2.6}$$

Our proposed class of CSs for θ are then feasible versions of (2.6) that replace $s_M(h,c)$ and $h_M(c)$ with suitable empirical analogues $\widehat{s}_M(h,c)$ and $\widehat{h}_M(c)$, respectively:

$$\mathcal{C}_{\text{ar}}^\alpha = \left\{ c : |\widehat{\tau}_M(\widehat{h}_M(c),c)| \leq \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c),c)) \widehat{s}_M(\widehat{h}_M(c),c)) \right\}, \tag{2.7}$$

with $\widehat{r}_M(h,c) = \bar{b}_M(h,c)/\widehat{s}_M(h,c)$. Such CSs could in principle be implemented in a variety of ways, and our theoretical analysis below therefore only imposes some weak "consistency" conditions. However, we propose a specific standard error $\widehat{s}_M(h,c)$ that

[7] Note that this bound may not be sharp if no such pair of piecewise quadratic functions is a feasible candidate for (μ_Y, μ_T). For example, there is no function μ_T with $\mu_T''(x) = B_T \cdot \text{sign}(x)$ and $\mu_T(x) \in [0,1]$ for all $x \in [-h,h]$ if $h > (2/B_T)^{1/2}$. Still, the bias bound is valid in such cases.

substitutes appropriate nearest-neighbor estimates $\widehat{\sigma}^2_{M,i}(c)$ into the above expression for $s_M(h,c)$ in Section 2.6.1; and a feasible bandwidth $\widehat{h}_M(c)$ that combines a plug-in construction with a safeguard against certain small sample distortions in Section 2.6.2.

In Online Appendix 2.D, we present an extension of our approach that allows constructing a bias-aware AR CSs for the ratio of the jumps in the vth-order derivatives of two conditional expectation functions at some threshold value, using pth-order local polynomial regression. The most prominent example of such setup is the Fuzzy Regression Kink Design (Card et al., 2015), where the parameter of interest is the ratio of jumps in first derivatives, and the CSs are typically based on local quadratic regression.

2.5. THEORETICAL PROPERTIES

2.5.1. **Assumptions.** To study the theoretical properties of our proposed CSs, we introduce the following assumptions.

Assumption 2.1. *(i) The data* $\{(Y_i, T_i, X_i), i = 1, \ldots, n\}$ *are an i.i.d. sample from a fixed population; (ii)* $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^q|X_i = x)$ *exists and is bounded uniformly over* $x \in$ *supp*(X_i) *and* $(\mu_Y, \mu_T) \in \mathcal{F}$ *for some* $q > 2$ *and every* $c \in \mathbb{R}$*; (iii)* $\mathbb{V}(M_i(c)|X_i = x)$ *is bounded away from zero uniformly over* $x \in$ *supp*(X_i) *and* $(\mu_Y, \mu_T) \in \mathcal{F}$ *for every* $c \in \mathbb{R}$*; (iv) the kernel function* K *is a continuous, unimodal, symmetric density function that is equal to zero outside some compact set, say* $[-1, 1]$*.*

Assumption 2.1 is standard in the literature on local linear regression. Part (i) could be weakened to allow for certain forms of dependent sampling, such as cluster sampling. Parts (ii)–(iii) are standard moment conditions. Since $M_i(c) = Y_i - cT_i$ and T_i is binary, these conditions mainly restrict the conditional moments of the outcome variable. Part (iv) is satisfied by most kernel functions commonly used in applied RD analysis, such as the triangular or the Epanechnikov kernels.

Assumption 2.2. *The following holds uniformly over* $(\mu_Y, \mu_T) \in \mathcal{F}$*: (i)* $\widehat{h}_M(c) = h_M(c)(1 + o_P(1))$*; and (ii)* $\widehat{s}_M(\widehat{h}_M(c), c) = s_M(h_M(c), c)(1 + o_P(1))$*.*

Part (i) of Assumption 2.2 states that the empirical bandwidth is consistent for the infeasible optimal one, and part (ii) states that that the empirical standard error is consistent for the true standard deviation at the infeasible optimal bandwidth. We discuss specific implementations in Sections 2.6.1 and 2.6.2.

Assumption LL1. The support of the running variable X_i is finite and symmetric, in the sense that it is of the form $\{\pm x_1, \ldots, \pm x_k\}$, for positive constants $(x_1, \ldots, x_k)$ over some open neighborhood of the cutoff.

Assumption LL2. (i) The running variable X_i is continuously distributed with density f_X that is bounded and bounded away from zero over an open neighborhood of the cutoff; (ii) $\mathbb{V}(M_i(c)|X_i = x)$ is Lipschitz continuous uniformly over $x \in \text{supp}(X_i)\backslash 0$ and $(\mu_Y, \mu_T) \in \mathcal{F}$ for every $c \in \mathbb{R}$; and (iii) $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^4|X_i = x)$ is uniformly bounded over $x \in \mathbb{R}$ and $(\mu_Y, \mu_T) \in \mathcal{F}$ for every $c \in \mathbb{R}$.

Assumptions LL1–LL2 are standard descriptions of setups with a discrete and a continuously distributed running variable, respectively.[8] In Lemma A.2.1 in the Appendix, we show that these assumptions have two main implications that we use in the proofs of the main results below: (i) using an estimate of the optimal bandwidth instead of its population version has a minor impact, in some appropriate sense, on the quantities involved in the construction of our CS; (ii) the magnitude of each of the weights $w_i(h_M(c))$ becomes arbitrarily small relative to the others' in large samples, in the sense that $w_{\text{ratio}}(h_M(c)) = o_P(1)$, where $w_{\text{ratio}}(h) = \max_{j=1,\dots,n} w_j(h)^2 / \sum_{i=1}^n w_i(h)^2$, which means that a CLT applies to an appropriately standardized version of the estimator of $\tau_M(c)$.

2.5.2. **Honesty.** Our main theoretical result in this chapter is that $\mathcal{C}^\alpha_{\text{ar}}$ is an honest CS for θ with respect to $\mathcal{F}$ as defined in (2.1) under the rather weak conditions introduced in the previous subsection. As mentioned above, such a property is necessary to guarantee that a CS has good finite sample coverage.

Theorem 2.1. *Suppose that Assumptions 2.1–2.2 and either LL1 or LL2 hold. Then* $\mathcal{C}^\alpha_{\text{ar}}$ *is honest with respect to* $\mathcal{F}$ *in the sense of* (2.1).

2.5.3. **Shape.** Since $\mathcal{C}^\alpha_{\text{ar}}$ is defined through an inversion argument, it is interesting to study its general shape. Recall that $c \in \mathcal{C}^\alpha_{\text{ar}}$ if and only if

$$|\widehat{\tau}_M(\widehat{h}_M(c), c)| - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c), c))\widehat{s}_M(\widehat{h}_M(c), c) \leq 0.$$

A simple sufficient condition for $\mathcal{C}^\alpha_{\text{ar}}$ to be non-empty is that $h_M(c)$ is continuous in c, but beyond that it is difficult to make general statements. This is because the quantities involved in the above inequality depend on c directly, but also indirectly through the bandwidth $\widehat{h}_M(c)$. While the former dependence is rather simple in structure, the latter introduces complicated nonlinearities that make it impossible to give a simple analytical result regarding the shape of our CS. Such a characterization is possible, however, for a version that uses bandwidth that does not depend on c.

[8] A discrete running variable with asymmetric support can easily be accommodated by using a different bandwidth on each side of the cutoff, as in described Appendix 2.B.

Theorem 2.2. *Let $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ be a version of $\mathcal{C}^{\alpha}_{\text{ar}}$ that uses a bandwidth h that does not depend on c. Then either $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, a_2]$, or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1] \cup [a_2, \infty)$, or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, \infty)$, or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, \infty)$ or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1]$, for some constants $a_1 < a_2$.*

This result mirrors the identification analysis in Section 2.2.4, and suggests that our actual CS should also take one of these general shapes as long as $\widehat{h}_M(c)$ does not vary "too much" with c. We found this to be the case in every simulation run and every empirical analysis that we conducted in the context of this chapter. The last two cases in Theorem 2.2, in which $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ is a half-line, are also "knife-edge" cases: they only occur if one of the boundaries of a bias-aware CI for τ_T is exactly equal to zero. Since this is a probability zero event under standard asymptotics, these two cases are largely irrelevant for empirical practice.

2.5.4. **Comparison with Bias-Aware Delta Method CIs.** Armstrong and Kolesár (2020) study bias-aware DM CIs under conditions for which such DM CIs are asymptotically valid. These include Assumption LL2, which implies that is X_i continuously distributed, and that $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}^{\delta}_H(B_T) \equiv \mathcal{F}^{\delta}$ for some $\delta > 0$, which means that τ_T is well-separated from zero. Armstrong and Kolesár (2020) show that in this case bias-aware DM CIs honest with respect to $\mathcal{F}^{\delta}$, and also near-optimal, in the sense that no other method can substantially improve upon its length in large samples. This construction thus dominates others commonly used in empirical practice, such as robust bias correction (Calonico et al., 2014).

The next theorem shows that our bias-aware AR CSs are as efficient as their DM counterparts in settings for which DM CIs are specifically designed. In order to avoid introducing additional high-level assumptions about the implementation details we consider an infeasible version of the bias-aware DM CI from Armstrong and Kolesár (2020), and compare them to our infeasible counterpart $\mathcal{C}^{\alpha}_{*}$; see the proof for further discussion and the exact construction of $\mathcal{C}^{\alpha}_{\Delta}$. Equal efficiency is established in the sense that both CSs have the same local asymptotic coverage for a drifting parameter within a neighborhood of θ; the most interesting being of order $O(n^{-2/5})$, as the length of $\mathcal{C}^{\alpha}_{\Delta}$ is $O_P(n^{-2/5})$ uniformly over $\mathcal{F}^{\delta}$.

Theorem 2.3. *Suppose that Assumptions 2.1–2.2 and LL2 hold, and put $\theta^{(n)} = \theta + \kappa \cdot n^{-2/5}$ for some constant κ. Then*

$$\limsup_{n\to\infty} \sup_{(\mu_Y,\mu_T)\in\mathcal{F}^{\delta}} \left|\mathbb{P}\left(\theta^{(n)} \in \mathcal{C}^{\alpha}_{*}\right) - \mathbb{P}\left(\theta^{(n)} \in \mathcal{C}^{\alpha}_{\Delta}\right)\right| = 0.$$

This result parallels the well-known finding that there is no loss of efficiency when using the AR approach in exactly identified IV models relative to one based on a con-

ventional t-test (e.g. Andrews et al., 2019). It is not a simple corollary, however, as there are, for example, no analogues to the bandwidth and the smoothing bias in an IV model. Note that bias-aware DM CIs do not account for the actual bias of the estimator of interest, but only for the bias of the leading term in a stochastic approximation; and even that bound needs to be estimated. They are thus subject to additional higher-order distortions that could affect their finite sample performance relative to that of our AR CSs.

2.6. IMPLEMENTATION DETAILS

2.6.1. **Standard Errors.** Given the form of the conditional standard deviation $s_M(h,c)$, it is natural to use a standard error of the form $\widehat{s}_M(h,c) = (\sum_{i=1}^n w_i(h)^2 \widehat{\sigma}^2_{M,i}(c))^{1/2}$, with $\widehat{\sigma}^2_{M,i}(c)$ some estimate of $\sigma^2_{M,i}(c)$. Nearest-neighbor estimators that defines $\widehat{\sigma}^2_{M,i}(c)$ as the squared difference between the outcome of unit i and the average outcome among its nearest neighbors in terms of the running variable (Abadie and Imbens, 2006; Abadie et al., 2014) are a common recommendation in the RD literature for this purpose (e.g. Calonico et al., 2014; Armstrong and Kolesár, 2018, 2020). However, such a standard error is actually not uniformly consistent over $\mathcal{F}$ because the leading bias of $\widehat{\sigma}^2_{M,i}(c)$ is proportional to the first derivative of $\mu_M(\cdot,c)$ at X_i (Abadie and Imbens, 2006), which is unbounded over $\mathcal{F}$. We therefore propose a novel nearest-neighbor procedure in which the local sample average is replaced with a best linear predictor.

Specifically, let R be a small fixed integer, denote the rank of $|X_j - X_i|$ among the elements of the set $\{|X_s - X_i| : s \in \{1,\dots,n\} \setminus \{i\}, X_s X_i > 0\}$ by $r(j,i)$, let $\mathcal{R}_i$ be the set of indices such that $r(j,i) \leq Q_i$, where Q_i is the smallest integer such that $\mathcal{R}_i$ contains at least R elements, and let R_i be the resulting cardinality of $\mathcal{R}_i$. If every realization of X_i is unique, then $R = Q_i = R_i$, and $\mathcal{R}_i$ is the set of unit i's R nearest neighbors' indices; but with ties in the data R_i could be greater than R. We then define $\widehat{\sigma}^2_{M,i}(c)$ as the scaled squared difference between $M_i(c)$ and its best linear predictor given its R_i nearest neighbors:

$$\widehat{\sigma}^2_{M,i}(c) = \frac{1}{1+H_i}\left(M_i(c) - \widehat{M}_i(c)\right)^2, \text{ with}$$
$$\widehat{M}_i(c) = \widetilde{X}_i \left(\sum_{j\in\mathcal{R}_i} \widetilde{X}_j^\top \widetilde{X}_j\right)^{-1} \sum_{j\in\mathcal{R}_i} \widetilde{X}_j^\top M_j(c), \quad H_i = \widetilde{X}_i \left(\sum_{j\in\mathcal{R}_i} \widetilde{X}_j^\top \widetilde{X}_j\right)^{-1} \widetilde{X}_i^\top.$$

Here $\widetilde{X}_i = (1, X_i)^\top$ if the running variable takes at least two distinct values among the R_i nearest neighbors of unit i, and $\widetilde{X}_i = 1$ otherwise. The scaling term H_i, whose form follows from standard regression theory, ensures that $\widehat{\sigma}^2_{M,i}(c)$ is approximately unbiased in large samples. The next result shows that our new standard error is indeed uniformly

consistent.

Theorem 2.4. *Suppose that Assumption 2.1, Assumption 2.2(i), and either Assumption LL1 or Assumption LL2 are satisfied. Then Assumption 2.2(ii) holds for the standard error described in this subsection.*

This result holds because the bias of $\widehat{\sigma}^2_{M,i}(c)$ is proportional to the second derivative of $\mu_M(\cdot, c)$ at X_i, which is bounded in absolute value over $\mathcal{F}$ by $B_Y + |c|B_T$. In contrast, the result would not hold for the conventional nearest-neighbor estimator, whose bias is proportional to the first derivative of $\mu_M(\cdot, c)$ at X_i and therefore unbounded. We therefore recommend using our standard error not just the construction of our CS, but more generally in bias-aware inference problems that work with bounds on second derivatives. We use $R = 5$ in the simulations and the empirical application in this chapter.

2.6.2. **Bandwidth Choice.** An obvious candidate for a feasible bandwidth is the empirical analogue of $h_M(c)$, which minimizes the length of the auxiliary CI in Section 4:

$$\widehat{h}^*_M(c) = \arg\min_h \mathrm{cv}_{1-\alpha}(\widehat{r}_M(h, c))\widehat{s}_M(h, c).$$

While this choice is attractive in principle, in finite samples it could yield some coverage distortions if $B_Y + |c|B_T$ is very large relative to sampling uncertainty. To see why, recall from the discussion at the end of Section 2.5.1 that asymptotic normality of $\widehat{\tau}_M(h, c) = \sum_{i=1}^n w_i(h)M_i(c)$ follows from a CLT if $w_{\mathrm{ratio}}(h) = o_P(1)$. Normality should thus be a "good" finite-sample approximation if $w_{\mathrm{ratio}}(h)$ is "close" to zero. If $B_Y + |c|B_T$ is large, however, $\widehat{h}^*_M(c)$ is typically small in order to control the bias. The weights $w_i(\widehat{h}^*_M(c))$ then concentrate on few observations close to the cutoff, $w_{\mathrm{ratio}}(\widehat{h}^*_M(c))$ is large, and CLT approximations could be inaccurate as $\widehat{\tau}_M(\widehat{h}^*_M(c), c)$ then effectively behaves like a sample average of a small number of observations.

To address this issue, we propose imposing a lower bound on the bandwidth, chosen such that the value of $w_{\mathrm{ratio}}(h)$ remains below some reasonable threshold constant $\eta > 0$:

$$\widehat{h}_M(c) = \max\left\{\widehat{h}^*_M(c), h_{\min}(\eta)\right\}, \quad h_{\min}(\eta) = \min\left\{h : w_{\mathrm{ratio}}(h) < \eta\right\}.$$

To motivate a choice for η, suppose that $\mathcal{X}_n = \{\pm.02, \pm.04, \ldots, \pm1\}$, that $K(t) = (1 - |t|)\mathbf{1}\{|t| < 1\}$ is the triangular kernel, and that $h = 1$. In this case $w_{\mathrm{ratio}}(h) \approx .075$, and a CLT approximation should be reasonably accurate for $\widehat{\tau}_M(h, c)$, which is a weighted least squares estimator with 50 observations on each side of cutoff. Choosing $\eta \in [0.05, 0.1]$ therefore seems reasonable in practice; and we actually use $\eta = .075$ in our simulations.

As $\widehat{h}_M(c) \geq \widehat{h}^*_M(c)$, the constrained bandwidth could over-smooth the data relative to the one that would be asymptotically optimal for inference. If that happens, the

resulting increase in finite-sample bias is the cost for normality being a better finite-sample approximation. This trade-off seems worthwhile since our CS construction explicitly accounts for the exact bias through, while deviations from normality cannot be captured. Under standard conditions like Assumption LL1 or LL2 the lower bound on the bandwidth clearly never binds asymptotically, but it can improve the finite-sample coverage of our CSs. The same idea can also be used for SRD inference, and more generally in settings where the finite-sample accuracy of inference faces a similar "bias vs. normality" trade-off. For example, Armstrong and Kolesár (2021a) use our approach for inference on average treatment effects under unconfoundedness with limited overlap.

2.6.3. **Computation.** Although our CS is defined through an inversion argument, it can be computed rather efficiently. We start by noting that our CS can be written as

$$\mathcal{C}_{\text{ar}}^{\alpha} = \{c : \widehat{p}(c) \geq 0\}, \text{ where } \widehat{p}(c) = 1 - \alpha - F\left(\left|\frac{\widehat{\tau}_M(\widehat{h}_M(c), c)}{\widehat{s}_M(\widehat{h}_M(c), c)}\right|, \widehat{r}_M(\widehat{h}_M(c), c)\right), \quad (2.8)$$

and $F(\cdot, r)$ is the CDF of the $|N(r, 1)|$ distribution. Computing $\mathcal{C}_{\text{ar}}^{\alpha}$ thus reduces to finding the roots of $\widehat{p}(c)$. Algorithm 1 describes how this is implemented in the `R` package that we provide with this chapter. The main idea is to first evaluate $\widehat{p}(c)$ on a coarse grid over the plausible range of θ to get a "rough" picture of $\widehat{p}(c)$, and then search for a root between grid points where the sign of $\widehat{p}(c)$ changes. Following the discussion after Theorem 2.2, we assume that the boundaries of a bias-aware CI for τ_T are not exactly equal to zero, and exploit that $(-\infty, a_1] \cup [a_2, \infty) \subset \mathcal{C}_{\text{ar}}^{\alpha}$ for some $a_1 < a_2$ if zero is contained in such a CI (this holds because the t-ratios of $\widehat{\tau}_M(h, c)$ and $\widehat{\tau}_T(h)$ become equal for $|c| \to \infty$). In line with the conjecture after Theorem 2.2, $\widehat{p}(c)$ turned out to have either two or no roots in all of our numerical examples, but our algorithm does not assume that this is the case.

The runtime of Algorithm 1 is mostly driven by the computational cost of evaluating the function $\widehat{p}(c)$. This cost is rather low with efficient programming: even with $n = 10^5$ data points, our algorithm computes $\mathcal{C}_{\text{ar}}^{\alpha}$ in about 20 seconds on a standard desktop computer. For comparison, it takes the widely used `rdrobust` package about 45 seconds to compute a robust bias correction DM CI on the same machine with the same number of data points (with smaller samples there is generally no practically relevant difference between the computation times of the two packages). Much computation time can be saved by noting that the nearest-neighbor variance estimates do not have to be computed from scratch for every value of c. This is because $\widehat{\sigma}^2_{M,i}(c) = \widehat{\sigma}^2_{Y,i} + c^2\widehat{\sigma}^2_{T,i} - 2c\widehat{\sigma}_{YT,i}$ is a quadratic function in c, with coefficients given by two variance terms and one covariance term that need to be computed only once. Also note that computing $\widehat{h}_M(c)$ is not too

Algorithm 1. Computes the CS $\mathcal{C}^\alpha_{\text{ar}}$ for θ given bounds B_Y and B_T on the second derivatives of μ_Y and μ_T, respectively, and the number R of nearest neighbors to be used for the variance estimates that enter standard error.

1. Pick an interval $[c_L, c_U]$ that covers the plausible range of θ, and define grid points $c_j = c_L + j(c_U - c_L)/J$ for $j = 0, \ldots, J$ and some integer $J \geq 2$.
2. Compute $\widehat{p}(c_j)$ as in (2.8) for $j = 0, \ldots, J$. If $\widehat{p}(c_j)$ and $\widehat{p}(c_{j+1})$ have different sign, use the `uniroot` algorithm to find a root of $\widehat{p}(\cdot)$ over the interval (c_j, c_{j+1}). Denote the number of roots found by $S \geq 0$, and the roots themselves by $a_1, \ldots, a_S$.
3. Compute $\mathcal{C}^\alpha_T$, a bias-aware CI for τ_T, the jump in treatment probability.
4. Return the bias-aware AR CS $\mathcal{C}^\alpha_{\text{ar}}$ according to the following rules.
   (a) If $0 \in \mathcal{C}^\alpha_T$ and $S = 0$, then return $\mathcal{C}^\alpha_{\text{ar}} = (-\infty, \infty)$.
   (b) If $0 \in \mathcal{C}^\alpha_T$, S is positive and even, $\widehat{p}$ is decreasing at a_s if s is odd, and increasing if s is even, then return $\mathcal{C}^\alpha_{\text{ar}} = (-\infty, a_1] \cup [a_2, a_3] \cup \ldots \cup [a_S, \infty)$.
   (c) If $0 \notin \mathcal{C}^\alpha_T$, S is increasing at a_s if s is odd, and decreasing if s is even, then return $\mathcal{C}^\alpha_{\text{ar}} = [a_1, a_2] \cup [a_3, a_4] \cup \ldots \cup [a_{S-1}, a_S]$.

   If none of the four conditions is satisfied, restart the algorithm with a larger interval $[c_L, c_U]$ and/or a larger number of grid points J.

costly, as the corresponding optimization problem only involves a single linear regression for every candidate value of the bandwidth. This step is much less involved than, say, leave-one-out cross validation, which would require n linear regressions for every candidate bandwidth.

2.6.4. **Choosing Smoothness Bounds.** In order to compute $\mathcal{C}^\alpha_{\text{ar}}$, one needs to specify values for the smoothness bounds B_Y and B_T. Such bounds cannot be estimated consistently without imposing strong additional assumptions; and without specifying such bounds it is generally not possible to conduct inference on θ that is both valid and informative, even in large samples (Low, 1997; Armstrong and Kolesár, 2018; Bertanha and Moreira, 2020).

Roughly speaking, small values of B_Y and B_T amount to the assumption that the respective functions are "close" to linear on either side of the cutoff, whereas for larger values they are allowed to be increasingly "curved". This choice should be guided by subject knowledge, but in empirical applications there will generally be no single objectively right one. We hence recommend considering a range of plausible values as a form of sensitivity analysis. In the following subsections, we give some suggestions for how to determine such ranges, and for how to communicate their implications. For simplicity,

we focus on the choice of B_Y, but the choice of B_T follows from analogous considerations.

We note that, as pointed out in the introduction, the need to specify smoothness bounds arises generally with bias-aware inference, but not with other popular methods like undersmoothing or robust bias correction. While at first glance this might seem like a disadvantage, in effect other methods also require such bounds to guarantee approximately correct CI coverage in practice.[9] Having to specify B_Y and B_T is thus not a meaningful impediment of our approach, but helps clarifying the assumptions on which inferential statements are based.

2.6.4.1. *Visualizing Smoothness Bounds.* Determining whether a particular value of B_Y is plausible in practice requires intuition for what functions are actually contained in $\mathcal{F}_H(B_Y)$. We suggest a procedure that visualizes some "extreme" elements of $\mathcal{F}_H(B_Y)$ to convey such intuition. Specifically, our proposal is to pick functions that match the scale of the data, and whose second derivative is equal to B_Y near the cutoff, through the following algorithm. Let $g(X_i)$ be a vector of basis transformations of X_i and its interaction with $\mathbf{1}\{X_i \geq 0\}$, with sufficiently many entries for an OLS regression of Y_i on $g(X_i)$ to result in an erratic overfit of the data; and consider functions of the form $\widetilde{\mu}_Y(x) = g(x)^\top \widehat{\gamma}$, where $\widehat{\gamma}$ solves

$$\min_{\gamma} \sum_{i=1}^{n} (Y_i - g(X_i)^\top \gamma)^2 \text{ s.t. } \|g''(\cdot)^\top \gamma\|_\infty \leq B_Y, |g''(x_0)^\top \gamma| = |g''(-x_0)^\top \gamma| = B_Y,$$

for some $x_0 \geq 0$. The function $\widetilde{\mu}_Y$ is thus obtained by a constrained regression of Y_i on $g(X_i)$ in which the absolute second derivative is bounded by B_Y overall, and equal to B_Y near the cutoff. This optimization can easily be solved via quadratic programming.

We stress that $\widetilde{\mu}_Y$ is not supposed to be a good estimate of μ_Y, but simply an example of an "extreme" element of $\mathcal{F}_H(B_Y)$. The idea is to plot this function for various values of B_Y (and possibly x_0) to obtain a better understanding for what kind of functions are contained in $\mathcal{F}_H(B_Y)$. For example, one could start with a very small B_Y, implying an almost linear function, and then increase the value in small steps until the resulting $\widetilde{\mu}_Y$ becomes implausibly erratic. Figure 2.1 illustrates this approach for a hypothetical data set.

[9]For example, an undersmoothing SRD CI can only be expected to have approximately correct coverage in finite samples if the bias is "small" relative to the standard error. With local linear estimation, this can only be the case if the underlying function is "close" to linear, which is equivalent to its maximum second derivative being "close" to zero. A similar point applies to robust bias correction, which in its standard implementation can only be expected to deliver CIs with approximately correct coverage in finite samples if the maximum third derivative of the underlying function is "close" to zero (Kamat, 2018). A researcher that reports such a CI and considers it reliable thus implicitly imposes a smoothness bound. If that bound was made explicit, however, a more efficient CI could be constructed through a bias-aware approach. See Armstrong and Kolesár (2020) for more details on this point.

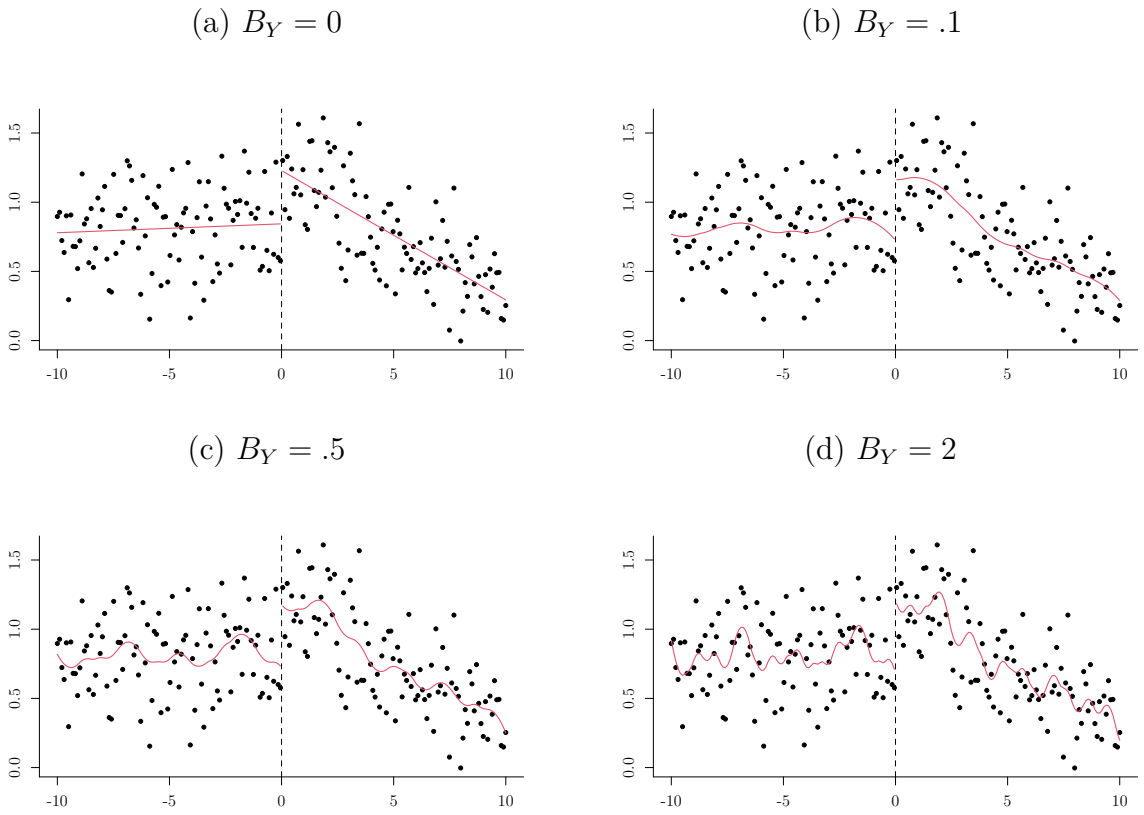


Figure 2.1: Examples of elements of $\mathcal{F}_H(B_Y)$ for various values of B_Y, each superimposed over the same hypothetical data set. Examples are constructed with $g(x)$ containing splines of order $k = 2$ and 50 knots on each side, and $\epsilon = .1$. Applied researchers can produce such graphs, and then pick the largest value of B_Y for which the resulting plot is empirically plausible. here the linear case in panel (a) is given for reference, panel (b) could be seen as adequate, panel (c) as a borderline case at best, and panel (d) would probably be considered implausible in economic applications.

2.6.4.2. *One-Sided CI for Smoothness Bound.* While it is not possible to obtain a valid data-driven upper bound on the curvature of μ_Y, it is possible to estimate a lower bound $\widehat{B}_{Y,\text{low}}$ for B_Y, and to compute a one-sided CI $[\widehat{B}^{\alpha}_{Y,\text{low}}, \infty)$ that covers B_Y with probability $1 - \alpha$ in large samples (cf. Armstrong and Kolesár, 2018; Kolesár and Rothe, 2018). We recommend computing these quantities in empirical practice to guard against overly optimistic choices of the smoothness bounds.

2.6.4.3. *Rules of Thumb.* While it is not possible to consistently estimate the smoothness bounds from data, we are aware of two heuristic "rules of thumb" (ROT) that have been suggested as a way of determining plausible values in practice. Both rules are based on fitting global polynomial specifications $\widetilde{\mu}_{Y,k}$ of order k on either side of the cutoff by conventional least squares. Armstrong and Kolesár (2020) consider fourth-oder

Figure 2.2: Conditional expectation function μ_Y for $\tau_Y = 1$ and various values of the smoothness bounds (solid line: $B_Y = 1$; dashed line: $B_Y = 10$; dotted line: $B_Y = 100$).

polynomials, and propose the ROT bound $\widehat{B}_{Y,\text{ROT1}} = \sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{Y,4}(x)|$, where $\mathcal{X}$ denotes the support of the running variable. Imbens and Wager (2019) consider a ROT in which the maximal curvature implied by a quadratic fit is multiplied by some moderate factor, say 2, to guard against overly optimistic values, yielding $\widehat{B}_{Y,\text{ROT2}} = 2\sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{Y,2}(x)|$.

Such rules can provide a useful first guidance to choosing smoothness bounds, but they should be complemented with other approaches in a sensitivity analysis. We strongly recommend to always check the fit of the respective polynomial specification, and to dismiss the ROT value if the fit is obviously poor. In Online Appendix 2.C, we compare the properties of ROT1 and ROT2 in a simple simulation study. We argue that in "roughly quadratic" settings the fourth-order polynomial specification that underlies ROT1 tends to produce quite erratic over-fits of the data. This leads to vast over-estimates of the true smoothness bounds, and corresponding CSs with poor statistical power. ROT2, on the other hand, tends to produce more reasonable values many such setups. See also our main Monte Carlo results in Section 2.7 for further details on this points.

2.7. SIMULATIONS

2.7.1. **Setup.** We now compare the practical performance of our bias-aware AR CS to that of alternative procedures though a Monte Carlo Study. We consider a number of data generating processes with varying curvature of the conditional expectation functions, richness of the running variable's support, strength of identification. Specifically, we simulate X_i from either a continuous uniform distribution over $[-1, 1]$ or a discrete uniform distribution over $\{\pm 1/15, \pm 2/15, \ldots \pm 1\}$; and let

$$\begin{aligned} Y_i &= (B_Y/2)\text{sign}(X_i)f(X_i) + \mathbf{1}\{X_i \geq 0\}\tau_Y + 0.1 \cdot \varepsilon_{1i}, \\ T_i &= \mathbf{1}\{-(B_T/2)\text{sign}(X_i)f(X_i) + \mathbf{1}\{X_i \geq 0\}\tau_T + 0.3 \geq \Phi(\varepsilon_{2i})\}, \end{aligned}$$

where $(\varepsilon_{1i}, \varepsilon_{2i})$ are bivariate standard normal with correlation 0.5, and $f(x) = x^2 - 1.5 \cdot \max(0, |x| - .1)^2 + 1.25 \cdot \max(0, |x| - .6)^2$. The functions μ_Y and μ_T are then second order splines with maximal absolute second derivative B_Y and B_T, respectively, over $[-1, 1]$. Figure 2.2 shows μ_Y for different values of B_Y. We consider the parameter values $(\tau_Y, \tau_T) \in \{(1, .2), (.5, .1)\}$, so that $\theta = 2$ in all settings, $B_T \in \{.2, 1\}$, and $B_Y \in \{1, 10, 100\}$; and set the sample size to $n = 1,000$. We refer to DGPs with $\tau_T = .1$ as weakly identified, and those with $\tau_T = .5$ as strongly identified.

We consider the performance of eight different AR CSs in our simulations: (i) our

bias-aware CS, using the true B_Y and B_T; (ii) our bias-aware CS, using twice the true B_Y and B_T; (iii) our bias-aware CS, using half the true B_Y and B_T; (iv) our bias-aware CS, using ROT1 estimates of B_Y and B_T; (v) our bias-aware CS, using ROT2 estimates of B_Y and B_T; (vi) a naive CS that ignores bias, using an estimate of the "pointwise-MSE optimal" bandwidth (Imbens and Kalyanaraman, 2012, henceforth IK); (vii) an undersmoothing CS, using $n^{-1/20}$ times the estimated IK bandwidth;[10] (viii) a robust bias correction CS, using local quadratic regression to estimate the bias, and estimated IK bandwidths. In addition, we also consider the performance of eight different DM CIs using the just-mentioned approaches to handling bias. Note that DM CIs based on undersmoothing and robust bias correction are currently the most common CSs in empirical FRD studies.[11]

2.7.2. **Simulations Results.** Table 2.1 shows simulated coverage rates of the various CSs we consider for $\theta = 2$. We first discuss results for AR CSs, shown in the left panel. With the true smoothness bounds, coverage rates our bias-ware CSs are close to and mostly slightly above the nominal level irrespective of the distribution of the running variable, the degree of nonlinearity of the unknown functions, and the degree of identification strength. The slight overcoverage occurs because the function $\mu_Y(x) - \theta\mu_T(x)$ is not exactly quadratic, and thus does not achieve the worst-case bias. Using twice or half the true bounds mostly leads to minor increases and decreases in simulated coverage, respectively. Using ROT1 for the smoothness bounds leads to over-coverage, especially for setups with a discrete running variable. This is because the underlying global quadratic approximation tends to severely over-estimate the smoothness bounds in our DGPs. ROT2 bounds generally lead to good coverage except for DGPs with $B_Y = 100$, where the underlying quadratic approximation leads to severe under-estimates of the smoothness bounds. Combining a naive approach, undersmoothing, or robust bias correction with an AR construction leads to CSs that undercover in all DGPs we consider,

[10] This CS corresponds to the one proposed by Feir et al. (2016) with a particular implementation of undersmoothing. Undersmoothing could in principle be implemented in a variety of ways, and hence the performance of the resulting CS must be interpreted accordingly.

[11] All computations are carried out with the statistical software package `R`. All bias-aware CSs are computed using our own software, which builds on the package `RDHonest`. All other CSs are computed using functions from the package `rdrobust`. A triangular kernel is used in all cases. Note that all CSs are only well-defined if the respective bandwidths are such that positive kernel weights are assigned to at least two (or three, in case of robust bias correction) distinct points on either side of the cutoff. In our simulations, the IK bandwidth estimates computed by `rdrobust` often do not satisfy this criterion if the running variable is discrete. We then manually set the bandwidth to 4/15, so that positive weights are given to three support points on each side of the cutoff. We also carried out a variant of our simulations in which we replace the IK bandwidth with the "coverage error optimal" bandwidth proposed by Calonico et al. (2018), using again the implementation in `rdrobust`. The results, which are qualitatively very similar to the ones reported in this section, are reported in Appendix 2.F.

with the distortions being more severe (up to about 30 percentage points) for those with larger values of B_Y and B_T.

Turning to result for DM CIs in the right panel of Table 2.1, we see that combining a bias-aware approach with this construction does not necessarily lead to a CI with correct coverage even under strong identification. This is because bias-aware DM CIs only control the bias of a first-order approximation of the estimator on which they are based. Such coverage distortions are further amplified by weak identification in our simulations. Discreteness of the running variable does not have a strong detrimental effect on bias-aware DM CIs in this particular setup though. Using the ROT choices for the smoothness bounds leads to further distortions in some cases. The coverage of DM CIs that use the naive approach, undersmoothing, or robust bias correction is again distorted for most DGPs, with particularly severe deviations for weak identification and large values of the smoothness constants.

To show that our bias-aware AR CSs not only have good coverage properties, but also yield comparatively powerful inference, we simulate the rates at which the various CSs we consider cover parameter values other than the true one. We report the results for the DGP with $(B_Y, B_T) = (1, .2)$ and strong identification in Figure 2.3.[12] To avoid having all 16 coverage curves in one plot, we split the results into four panels: the five bias-aware AR CSs in (a), the three other AR CSs in (b), the five bias-aware DM CIs in (c), and the three other DM CIs in (d). Panels (b)–(d) also show the curve for our bias-aware AR CS with the true constants to have a common point of reference.

Panel (a) then shows that the coverage rate of bias-aware AR CSs drops very quickly to zero away from the true parameter, except for the CS based on ROT1 (which, as mentioned above, severely overestimates the smoothness bounds). Panels (b)–(d) show that the coverage of bias-aware AR CSs is also below that of all competing procedures over almost all the parameter space. This confirms that the accurate coverage of our CSs in settings with discrete running variables and weak identification does not come at the expense of statistical power in a canonical setup, for which most competing CS are specifically constructed.

2.8. EMPIRICAL APPLICATION

In this section, we apply our methods to data from Oreopoulos (2006, 2008), who studies the effects of a 1947 education reform in Great Britain that raised the minimum school-leaving age from 14 to 15 years. The data are a sample of $n = 73,954$ workers who

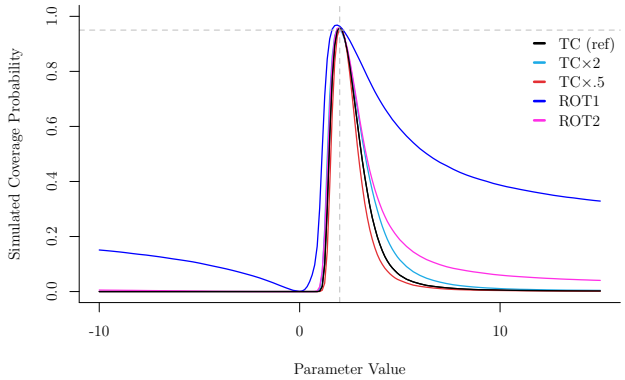
[12]We focus on these results because the coverage of the true parameter is reasonably close to the nominal level for all procedures, and thus comparison of coverage rates at "non-true" parameter values is meaningful across CSs. Analogous plots for other DGPs are available from the authors.

Table 2.1: Simulated coverage rate (in %) of true parameter for various types of confidence sets

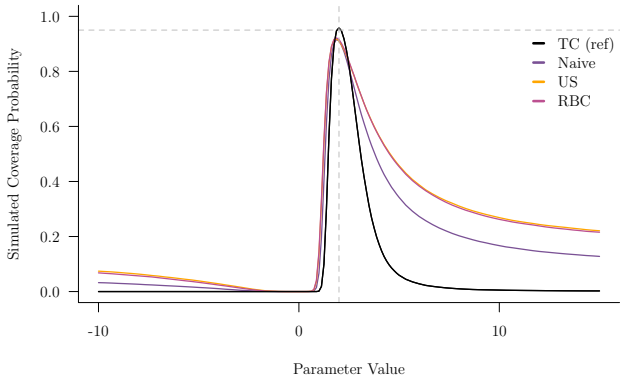
			Anderson-Rubin								Delta Method							
			Bias-Aware								Bias-Aware							
τ_T	B_Y	B_T	TC	TC×2	TC×.5	ROT1	ROT2	Naive	US	RBC	TC	TC×2	TC×.5	ROT1	ROT2	Naive	US	RBC
Running Variable with Continuous Distribution																		
0.5	1	0.2	97.2	97.1	96.9	96.4	96.9	93.1	93.4	93.4	97.3	96.6	97.6	92.6	95.3	90.8	90.4	91.3
0.5	1	1.0	96.7	96.5	96.7	96.4	96.5	93.0	93.3	93.3	95.8	94.5	97.5	92.6	95.3	90.4	89.9	91.0
0.5	10	0.2	95.8	95.6	96.3	96.8	96.4	92.4	93.0	92.5	95.0	94.5	95.3	93.2	94.3	88.3	88.3	88.7
0.5	10	1.0	95.5	95.4	96.2	96.6	96.1	92.2	92.8	92.2	94.5	93.9	95.7	93.0	94.0	87.9	88.2	88.4
0.5	100	0.2	95.1	98.9	88.8	99.5	86.1	78.5	87.9	74.7	94.5	98.0	86.0	98.1	79.1	72.8	80.8	72.2
0.5	100	1.0	95.1	99.0	88.6	99.5	86.0	78.2	88.0	74.4	93.3	97.8	87.1	98.1	79.9	72.6	80.9	72.1
0.1	1	0.2	97.2	97.3	96.8	97.1	97.3	93.7	94.0	94.0	92.3	90.0	92.0	79.0	87.0	76.6	74.0	79.2
0.1	1	1.0	97.3	97.1	96.8	97.1	96.9	93.4	93.8	93.8	89.6	86.5	93.2	78.7	87.6	76.5	73.8	79.0
0.1	10	0.2	96.9	96.6	97.1	97.4	97.1	93.4	94.0	93.7	84.1	85.8	82.8	79.3	82.6	71.0	69.6	73.3
0.1	10	1.0	96.9	96.7	97.1	97.5	97.1	93.2	93.9	93.5	85.0	83.0	87.2	78.8	83.5	70.3	69.2	72.7
0.1	100	0.2	96.3	99.2	91.5	99.6	89.7	83.2	91.0	79.0	83.6	96.0	65.4	92.3	54.4	36.7	47.4	37.1
0.1	100	1.0	96.4	99.2	91.7	99.6	89.8	83.2	91.0	79.1	82.4	94.4	72.6	92.3	58.1	36.5	47.4	37.0
Running Variable with Discrete Distribution																		
0.5	1	0.2	97.4	97.6	96.9	99.3	97.9	94.3	94.6	94.6	97.6	97.2	97.8	95.4	96.0	89.9	88.9	91.0
0.5	1	1.0	97.5	97.7	96.9	99.2	97.5	94.0	94.2	94.4	96.5	95.5	98.0	95.2	96.2	89.6	88.5	90.5
0.5	10	0.2	97.7	98.2	97.6	99.5	95.8	93.6	93.9	93.7	95.9	96.0	95.7	96.0	95.2	85.3	84.8	85.4
0.5	10	1.0	97.7	98.2	97.7	99.5	94.6	93.6	93.7	93.6	96.5	96.6	96.9	95.8	95.1	84.6	84.4	84.6
0.5	100	0.2	96.9	100.0	91.2	100.0	86.2	67.9	60.4	57.8	95.0	97.5	48.1	98.8	25.3	26.8	27.9	17.1
0.5	100	1.0	96.8	100.0	91.0	100.0	85.8	67.2	59.5	57.2	93.1	98.0	54.0	98.7	26.8	26.5	27.6	16.7
0.1	1	0.2	97.5	97.9	96.6	99.5	98.1	94.7	94.9	95.1	92.7	89.4	92.1	79.2	87.5	71.2	64.8	75.6
0.1	1	1.0	97.8	98.1	96.9	99.4	97.9	94.5	94.7	94.7	90.0	86.8	93.5	78.6	88.3	70.5	64.5	74.8
0.1	10	0.2	98.5	99.0	98.1	99.6	96.2	94.5	94.5	94.6	82.3	88.9	78.9	75.0	86.3	55.9	52.6	59.9
0.1	10	1.0	98.5	99.0	98.2	99.6	95.3	94.5	94.5	94.6	82.7	84.7	85.0	74.9	86.8	55.8	51.8	59.6
0.1	100	0.2	97.2	100.0	93.6	100.0	91.5	73.7	66.9	63.9	94.9	96.3	94.3	96.6	89.9	68.6	69.5	65.3
0.1	100	1.0	97.3	100.0	93.8	100.0	91.5	73.2	66.4	63.0	95.3	96.1	96.3	96.8	91.1	68.1	69.0	64.8

Notes: Results based on 50,000 Monte Carlo draws for a nominal confidence level of 95%. Columns show results for bias aware approach with true constants (TC), two times true constants (TC×2), half true constants (TC×.5), and with rule of thumb estimates (ROT1) and (ROT2); naive approach that ignores bias (Naive); undersmoothing (US); and robust bias correction (RBC).

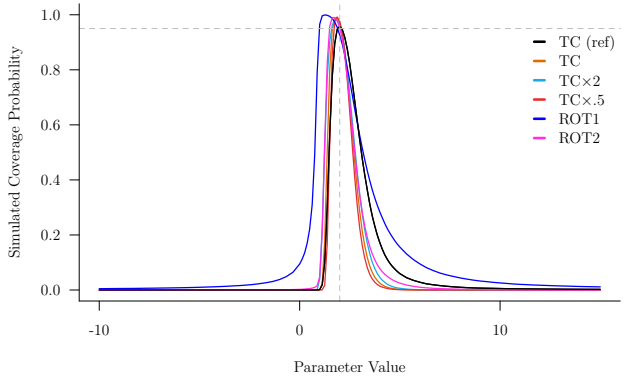
(a) Bias-aware Anderson-Rubin CSs



(b) Other Anderson-Rubin CSs



(c) Bias-Aware Delta Method CI



(d) Other Delta Method CIs

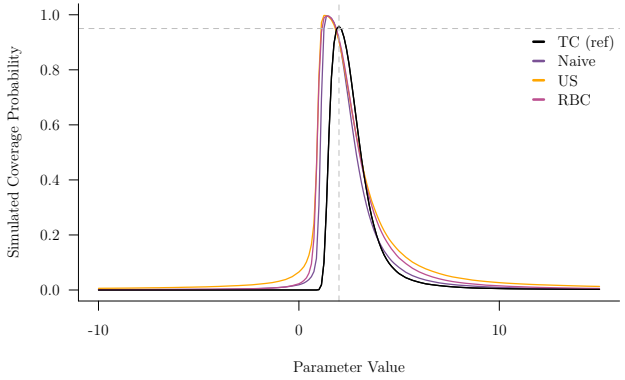


Figure 2.3: Simulated coverage rates of various values of parameter values and for different types of confidence sets. Based on the DGP described in the main text with $\tau_T = .5$, $B_T = .2$, and $B_Y = 1$. Bias aware approach with true constants (TC (ref); as reference function in all graphs), two times true constants (TC×2), 0.5 times true constants (TC×.5), and with rule of thumb smoothness bounds (ROT1) and (ROT2); naive approach that ignores bias (Naive); undersmoothing (US); and robust bias correction (RBC).

turned 14 between 1935 and 1965, obtained by combining the 1984-2006 waves of the UK General Household Survey. We take the effect of attending school beyond age 14 on annual earnings measured in 1998 UK pounds as the parameter of interest. The running variable is the year in which the worker turned 14, and the threshold is 1947. Figure 2.4 shows the average of log annual earnings and the empirical proportions of students who attended school beyond age 14 as a function of the running variable. The RD design is clearly seen to be fuzzy.

For reasons explained below, we conduct the analysis for both the entire data and the subset that excludes the 1947 cohort. Oreopoulos (2006) uses a parametric approach in which the respective dependent variable is regressed on a dummy for turning 14 in or after 1947 and a 4th order polynomial in age. This yields the estimate $\widehat{\theta} = .146$ with a 95% DM CI $[-.009; .300]$ based on a heteroscedasticity-robust standard error for the entire data, and $\widehat{\theta} = .111$ with a 95% DM CI $[-.032; .255]$ if the 1947 cohort is excluded.[13] These CIs, however, do not account for the model misspecification bias one should expect here.

To compute our bias-aware AR CSs, we first have to determine plausible values for the smoothness constants B_Y and B_T. To do that, we compute the ROT values, the lower bound estimates and one-sided CIs, and various graphs of candidate functions, all as described in Section 2.6.4. All graphs are shown in Appendix 2.E. Regarding the value of B_Y, inspection of the top panel of Figure 2.4 suggests that the function μ_Y should not be too erratic. Indeed, we estimate a lower bound of $\widehat{B}_{Y,\text{low}} = 0$ for B_Y, meaning that the data cannot rule out that μ_Y is linear. We also have $\widehat{B}_{Y,\text{ROT1}} = .023$ and $\widehat{B}_{Y,\text{ROT2}} = .012$, with the fit of the underlying polynomials seeming adequate in both cases. Including also some conservative values, we then consider $[0; .04]$ as a plausible range for B_Y.

Regarding the choice of B_T, one has to be more careful. From the bottom panel of Figure 2.4, we see that the empirical share of "treated" students increases very slowly after 1948, but jumps sharply from 0.724 for 1947 to 0.909 for 1948. If we consider the latter change to be natural variation in treatment probabilities, then only rather large values of B_T are consistent with the data. Indeed, we estimate a lower bound $\widehat{B}_{T,\text{low}} = .158$, with a 95% one-sided CI of $[0.126; \infty)$. The two ROTs yield much smaller values, namely $\widehat{B}_{T,\text{ROT1}} = .031$ and $\widehat{B}_{T,\text{ROT2}} = .011$. But since the fit of both underlying polynomial

[13]The numerical result here differ from those in Oreopoulos (2006) because (i) we use the data set from its online corrigendum (Oreopoulos, 2008), which includes additional waves of the UK General Household Survey; (ii) Oreopoulos (2006) considers a slightly different parameter of interest; and (iii) Oreopoulos (2006) uses Lee and Card (2008) standard errors that are clustered by the running variable. Kolesár and Rothe (2018) show that such clustering does not alleviate the issues caused by a discrete running variable, but tends to produce CIs with poor coverage properties, and hence such standard errors should not be used.

Table 2.2: Bias-aware Anderson-Rubin confidence sets for the effect of one additional year of compulsory schooling for various values of the smoothness bounds

	B_Y				
B_T	0	.01	.02	.03	.04
	Panel A: Results for full data set				
.12	[-.239; 1.841]	[-.366; 1.953]	[-.458; 2.068]	[-.555; 2.183]	[-.655; 2.301]
.14	[-.343; 2.395]	[-.448; 2.554]	[-.569; 2.716]	[-.694; 2.881]	[-.824; 3.049]
.16	[-.432; 3.608]	[-.591; 3.887]	[-.762; 4.174]	[-.941; 4.467]	[-1.131; 4.767]
.18	[-.637; 10.049]	[-.907; 11.152]	[-1.217; 12.279]	[-1.575; 13.427]	[-1.995; 14.590]
.20	$(-\infty; \infty)$	$(-\infty; \infty)$	$(-\infty; \infty)$	$(-\infty; \infty)$	$(-\infty; \infty)$
	Panel B: Results excluding data for 1947				
0	[-.108; .080]	[-.152; .441]	[-.237; .546]	[-.313; .619]	[-.386; .687]
.01	[-.100; .224]	[-.168; .496]	[-.257; .589]	[-.338; .665]	[-.415; .733]
.02	[-.117; .406]	[-.187; .554]	[-.280; .638]	[-.367; .714]	[-.459; .778]
.03	[-.125; .495]	[-.208; .606]	[-.307; .692]	[-.400; .765]	[-.489; .825]
.04	[-.126; .566]	[-.232; .664]	[-.340; .749]	[-.439; .814]	[-.522; .879]

Notes: All CSs have 95% nominal level. Results based on 73,954 data points for Panel A and 73,954 data points for Panel B. See main text for a justification of the smoothness bounds value considered.

specifications is poor we choose to disregard these values, and consider $[.12; .2]$ as a plausible range for B_T. The upper end was chosen because it turns out that for $B_T \geq .2$ our CS is always equal to the real line, and thus considering larger values would not affect the results.

If we take the arguably more realistic position that the change in treatment probabilities between 1947 and 1948 was largely caused by delayed implementation of the reform, a more natural approach is to exclude the 1947 cohort and conduct a "donut" analysis. We then estimate a lower bound $\widehat{B}_{T,\text{low}} = 0$ for B_T, meaning that linearity of μ_T cannot be ruled out, and the ROTs yield $\widehat{B}_{T,\text{ROT1}} = .013$ and $\widehat{B}_{T,\text{ROT2}} = .009$, with the fitted polynomial being adequate in both cases. To also include some conservative values, we then consider $[0; .04]$ as a plausible range for B_T in this donut setup.

In Table 2.2, then we report bias-aware AR CSs with nominal level 95%, separately for the entire data (top panel) and for the subsample that excludes the 1947 cohort (bottom panel), and for values of B_Y and B_T in regular grids over the ranges motivated above. All CSs in panel (a) are extremely wide, in the sense that even the shortest one is much larger than all plausible values for the return to increased compulsory schooling. This is because treating the sharp increase in treatment probability from 1947 to 1948 as natural variation implies that the parameter of interest is only weakly identified. In panel (b), which excludes 1947 cohort data, and considers an appropriate range for B_T, the CSs become much shorter, but they still all cover zero and many contain the full plausible

parameter space.

Our overall preferred specification is the one that excludes the 1947 cohort, and uses $B_Y = .02$ and $B_T = .01$ (the grid values in between the respective ROT estimates), which yields the bias-aware AR CS $[-.257, .589]$. This CS is almost three times as large as the reference CS $[-.032; .255]$ based on the parametric specification. Overall, the data are not very informative about the returns to schooling.

2.9. CONCLUSIONS

FRD designs occur frequently in many areas of applied economics. Motivated by the various shortcomings of existing methods of inference, we propose new confidence sets for the causal effect in such designs, which are based on a bias-aware AR construction. Our CSs are simple to compute, highly efficient, and have excellent coverage properties in finite samples because they explicitly take into account the exact smoothing bias from the local linear regression steps. They are also valid under weak identification and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

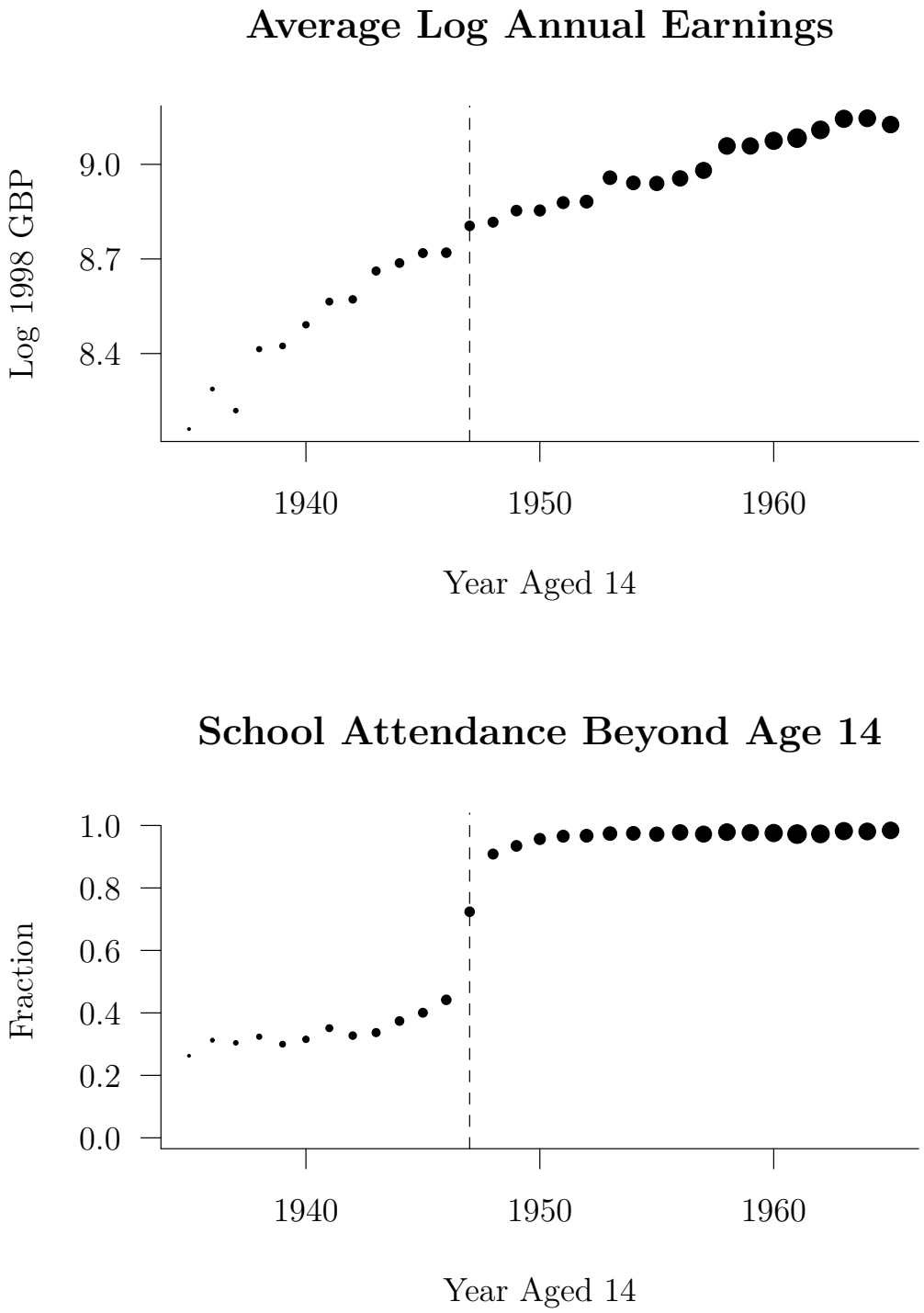


Figure 2.4: Average log annual earnings (top panel) and fraction of individuals in full time education beyond age 14 by birth year cohort. Dashed vertical lines indicate the year 1947, in which the minimum school leaving age changed from 14 to 15 years. Size of dots is proportional to the cohort size in the data.

Appendix to Chapter 2

2.A. PROOFS OF MAIN RESULTS

In this Appendix, we prove the main results from Section 2.5. We use repeatedly that, using basic least squares algebra, the statistic $\widehat{\tau}_M(h,c)$ can be written as

$$\widehat{\tau}_M(h,c) = \sum_{i=1}^{n} w_i(h) M_i(c), \quad w_i(h) = w_{i,+}(h) - w_{i,-}(h),$$

$$w_{i,+}(h) = e_1^\top Q_+^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_+ = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_i \widetilde{X}_i' \mathbf{1}\{X_i \geq 0\}$$

$$w_{i,-}(h) = e_1^\top Q_-^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_- = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_i \widetilde{X}_i' \mathbf{1}\{X_i < 0\},$$

with $\widetilde{X}_i = (1, X_i)'$. To simplify the notation, throughout the proofs we write $A_n(\mu) = o_{P,\mathcal{F}}(1)$ if $\sup_{\mu \in \mathcal{F}} P(|A_n(\mu)| > \epsilon) = o(1)$ for all $\epsilon > 0$ and a generic sequence $A_n(\mu)$ of random variables indexed by $\mu \in \mathcal{F}$. We also drop the dependency on c from the notation for the optimal bandwidth in most instances, writing h_M instead of $h_M(c)$.

2.A.1. **Proof of Theorem 2.1.** We first establish the following lemma.

Lemma A.2.1. *Suppose that Assumption 2.1–2.2 and either Assumption LL1 or Assumption LL2 are satisfied. Then the following holds uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}$: (i) $w_{\text{ratio}}(h_M(c)) = o_P(1)$; (ii) $(\widehat{\tau}_M(\widehat{h}_M(c), c) - \widehat{\tau}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$; and (iii) $(\bar{b}_M(\widehat{h}_M(c), c) - \bar{b}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$.*

Proof. We first show part (i). Suppose that Assumption LL1 is satisfied. With probability approaching 1, we have that

$$\max_{i \in \{1,\ldots,n\}} \frac{w_i(h_M)^2}{\sum_{j=1}^{n} w_j(h_M)^2} \leq \max_{i \in \{1,\ldots,n\}} \frac{w_i(h_M)^2}{\sum_{j: X_j = X_i} w_j(h_M)^2} = \max_{i \in \{1,\ldots,n\}} \frac{1}{\sum_{j: X_j = X_i} \mathbf{1}\{X_i = X_j\}}.$$

As $n \to \infty$, the number of units whose realization of the running variable is equal to any particular value in its support tends to infinity, and we obtain the statement of the lemma.

Now suppose that Assumption LL2 is satisfied. First, it is easy to see that the minimizer of $\text{cv}_{1-\alpha}(r_M(h,c)) \cdot s_M(h,c)$ must satisfy $h_M \to 0$ and $nh_M \to \infty$ as $n \to \infty$.

Under these conditions, the bias and variance of the local linear regression estimator scale as h_M^2 and $1/(nh_M)$, respectively. From the properties of the function $\text{cv}_{1-\alpha}(\cdot)$, it then follows that $h_M \propto n^{-1/5}(1 + o_P(1))$. It also holds that

$$\max_{i\in\{1,\ldots,n\}} \frac{w_i(h_M)^2}{\sum_{j=1}^n w_j(h_M)^2} \leq \max_{i:Z_i=1} \frac{w_i(h_M)^2}{\sum_{j:Z_j=1} w_j(h_M)^2} + \max_{i:Z_i=0} \frac{w_i(h_M)^2}{\sum_{j:Z_j=0} w_j(h_M)^2}.$$

It then suffices to show that the first term on the right-hand side of the last inequality tends to zero in probability uniformly over $\mathcal{F}$, as the same arguments can be used to prove an analogous result for the second term. Note that

$$\begin{aligned}
&\max_{i:Z_i=1} \frac{w_i(h_M)^2}{\sum_{j:Z_j=1} w_j(h_M)^2} \\
&\quad = \max_{i:Z_i=1} \frac{K(X_i/h_M)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2 - X_i \sum_{l:Z_l=1} X_l K(X_l/h_M)]^2}{\sum_{j:Z_j=1} K(X_j/h_M)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h_M)]^2}.
\end{aligned}$$

Treating the numerator of the right-hand side of the second line as a function of X_i, it follows from the fact that the kernel is bounded from above by Assumption 2.1 that this function is bounded from above by a quadratic function in $X_i \in [0, h]$. The maximum of this quadratic function is bounded by a constant multiplied by $[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2]^2 + h_M^2[\sum_{l:Z_l=1} X_l K(X_l/h_M)]^2$. Taken together, this means that

$$\begin{aligned}
&\max_{i:Z_i=1} \frac{w_i(h_M)^2}{\sum_{j:Z_j=1} w_j(h_M)^2} \\
&\quad \leq C \frac{(\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2)^2 + h_M^2(\sum_{l:Z_l=1} X_l K(X_l/h_M))^2}{\sum_{j:Z_j=1} K(X_j/h_M)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h_M)]^2}.
\end{aligned}$$

for some finite constant C, and for n sufficiently large. Standard kernel calculations than yield that the numerator on the right-hand side of the last inequality is an $O_P(n^2 h_M^2)$ term, while the denominator an $O_P(n^3 h_M^3)$ term. As $nh_M \to \infty$ as $n \to \infty$, this completes part (i).

Now consider part (ii)–(iii). Suppose Assumption LL1 holds. With a discrete running variable, it is clear that the optimal bandwidth h_M shrinks with the sample size, but it cannot tend to zero as it has to be greater than the support point second closest to the cutoff in order for the local linear regression estimator to be well-defined. Furthermore, any bandwidth h between the second and third support point closest to the cutoff implies the same local linear regression weights $w_i(h)$ for all i. Hence any bandwidth between the second and third support point closest to the cutoff is asymptotically optimal. Part (ii)–(iii) then follow trivially, as each expression under consideration depends on h only through $w_i(h)$.

Now suppose that Assumption LL2 holds. Statements (ii)–(iii) of Lemma A.2.1 then follow as in the proof of Theorem E.1 in Armstrong and Kolesár (2020). □

We now proceed with the proof of the core statement of Theorem 1. Since $\theta \in \mathcal{C}^{\alpha}_{\text{ar}}$ if and only if $\tau_M(\theta) \in \mathcal{C}^{\alpha}(\theta)$, it suffices to show that for any $c \in \mathbb{R}$

$$\liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\tau_M(c) \in \mathcal{C}^{\alpha}(c)) \geq 1-\alpha.$$

Note that it follows from Lemma A.2.1 (ii)–(iii) and uniform continuity of $\text{cv}_{1-\alpha}(\cdot)$ that

$$\begin{aligned}
&\frac{|\widehat{\tau}_M(\widehat{h}_M, c) - \tau_M(c)|}{\widehat{s}_M(\widehat{h}_M, c)} - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M, c)) \\
&\quad = \left| \frac{\widehat{\tau}_M(h_M, c) - \mathbb{E}\left[\widehat{\tau}_M(h_M, c)|\mathcal{X}_n\right]}{s_M(h_M, c)} + \frac{b_M(h_M, c)}{s_M(h_M, c)} \right| - \text{cv}_{1-\alpha}(r_M(h_M, c)) + o_{P,\mathcal{F}}(1).
\end{aligned}$$

We now apply Lyapunov's CLT to show that $(\widehat{\tau}_M(h_M, c) - \mathbb{E}\left[\widehat{\tau}_M(h_M, c)|\mathcal{X}_n\right])/s_M(h_M, c)$ converges in distribution to a standard normally distributed random variable, uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}$. Specifically, let C be a positive constant, let $\delta > 2$, and recall that $\widehat{\tau}_M(h_M, c) = \sum_{i=1}^n w_i(h_M) M_i(c)$. Lyapunov's CLT can be applied conditional on $\mathcal{X}_n$ since

$$\begin{aligned}
\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}\left[|w_i(h_M)(M_i(c) - \mathbb{E}[M_i(c)|\mathcal{X}_n])|^{\delta} |\mathcal{X}_n\right]}{\left(\sqrt{\sum_{i=1}^n w_i(h_M)^2 \sigma^2_{M,i}}\right)^{\delta}} &\leq \lim_{n\to\infty} C \sum_{i=1}^n \frac{|w_i(h_M)|^{\delta}}{\left(\sqrt{\sum_{i=1}^n w_i(h_M)^2}\right)^{\delta}} \\
&\leq \lim_{n\to\infty} C \max_{i=1,\ldots n} \left(\frac{|w_i(h_M)|}{\sqrt{\sum_{i=1}^n w_i(h_M)^2}} \right)^{\delta-2} = o_{P,\mathcal{F}}(1)
\end{aligned}$$

by Assumption 2.1(i)–(iii) and Lemma A.2.1(i). Standard arguments then yield that

$$\begin{aligned}
\liminf_{n\to\infty} \Bigg(&\inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\tau_M(c) \in \mathcal{C}^{\alpha}(c)) \\
&- \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}\left(\left| S + \frac{b_M(h_M, c)}{s_M(h_M, c)} \right| \leq \text{cv}_{1-\alpha}(r_M(h_M, c)) \right) \Bigg) = 0,
\end{aligned}$$

with S a generic standard normal random variable. The statement of the theorem now follows from the definition of the critical value function $\text{cv}_{1-\alpha}(\cdot)$ if

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_M(h_M, c)/s_M(h_M, c)| \leq r_M(h_M, c).$$

Note that Armstrong and Kolesár (2020, Theorem B.3) show that the last statement holds with equality if μ_Y and μ_T have unbounded domain. In our setup, we only have a potentially weak inequality because μ_T is naturally constrained to take values in $[0, 1]$,

and the supremum is thus taken over a smaller set of functions. This completes our proof. □

2.A.2. **Proof of Theorem 2.2.** To simplify the exposition, we emphasize the dependence of various estimators on c in our notation, but suppress their dependency on the bandwidth h (which does not depend on c under the conditions of this theorem). The CS $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ is given by the set of all values of c satisfying

$$\vartheta(c) \leq 0, \quad \text{where} \quad \vartheta(c) \equiv |\widehat{\tau}_Y - c\widehat{\tau}_T| - \text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c).$$

The function $\vartheta(c)$ is continuous in c, as $\text{cv}_{1-\alpha}(\cdot)$ is a uniformly continuous function, and both the standard error $\widehat{s}_M(c) = (\widehat{s}_Y^2 - 2c\widehat{s}_{TY} + c^2\widehat{s}_T^2)^{1/2}$ and the worst case bias $\overline{b}_M(h,c) = -(B_Y + |c|B_T)/2 \cdot \sum_{i=1}^n w_i(h)X_i^2 \cdot \text{sign}(X_i)$ are continuous in c. Moreover, the term $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is also strictly convex in c, because both the standard error and the worst-case bias are convex in c and $\text{cv}_{1-\alpha}(\cdot)$ is strictly convex and increasing. The shape of $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ is then determined by the roots of $\vartheta(c)$. While one can in principle solve analytically for the roots of $\vartheta(c)$, doing so is very tedious.

To prove the theorem, it suffices to show that the function $\vartheta(c)$ always fits into one of the following four categories: (i) $\vartheta(c) \leq 0$ for all c; (ii) $\vartheta(c)$ has two roots, and there exists $c^* > 0$ such that $\vartheta(c) < 0$ for all $|c| > c^*$; (iii) $\vartheta(c)$ has two roots, and there exists $c^* > 0$ such that $\vartheta(c) > 0$ for all $|c| > c^*$, and (iv) $\vartheta(c)$ has one root. Then $\mathcal{C}^{\alpha}_{\text{ar}}(h) = \mathbb{R}$ in case (i), $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1] \cup [a_2, \infty)$ for some $a_1 < a_2$ in case (ii); and by $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, a_2]$ for some $a_1 < a_2$ in case (iii), and $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_2]$ or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, \infty)$ in case (iv). We now go through a number of case distinctions.

If $\widehat{\tau}_T = 0$, then $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ is a constant function in c. As $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is strictly convex in c and unbounded, $\vartheta(c)$ must be either of form (i) or (ii). We threfore suppose that $\widehat{\tau}_T \neq 0$ from now on , and write $\widehat{\theta} = \widehat{\tau}_Y/\widehat{\tau}_T$. Since $\vartheta(\widehat{\theta}) < 0$ by construction, the function $\vartheta(c)$ cannot be strictly positive. As $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ is a piecewise linear function and $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is strictly convex, the function $\vartheta(c)$ can also have at most two roots for $c \leq \widehat{\theta}$, and at most two roots for $c > \widehat{\theta}$. If it does not have any root, $\vartheta(c)$ is of the form (i).

Let us first assume that $\lim_{c\to\pm\infty} \vartheta(c) \neq 0$. It follows from basic algebra that there exists some c^* sufficiently large such that $\text{sign}(\vartheta(c)) = \text{sign}(\vartheta(-c)) = 1$ or $\text{sign}(\vartheta(c)) = \text{sign}(\vartheta(-c)) = -1$ and $\vartheta(c) \neq 0$ for all $c > c^*$. The function $\vartheta(c)$ therefore cannot have one or three roots; so it must have either four roots or two roots or none. If $\text{sign}(\vartheta(c)) = -1$ for all $|c| > c^*$, which means that $|\widehat{\tau}_Y - c\widehat{\tau}_T| > \text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$. The function $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ intersects once with the function $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ for $c < \widehat{\theta}$, and once for $c > \widehat{\theta}$. Therefore $\vartheta(c)$ must be of form (iii) in this case. If $\text{sign}(\vartheta(c)) = -1$ for

all $|c| > c^*$, the above reasoning only yields that $\vartheta(c)$ has at most four roots. However, note that for $|c| \to \infty$ the absolute value of the first derivative of $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ with respect to c converges to some constant ϖ, and that for any value of $\varsigma \in \mathbb{R}$ the expression $\text{sign}(c) \cdot (\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M - |\varsigma + \varpi \cdot c|)$ converges to a constant. Choose ς such that the latter constant is zero, and set $\varrho(c) = |\varsigma + \varpi c|$. By construction, $\varrho(c)$ intersects with $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ twice either for $c \leq \widehat{\theta}$ or $c \geq \widehat{\theta}$. It also holds that $\varrho(c) \leq \text{cv}_{1-\alpha}(\widehat{r}(c)) \cdot \widehat{s}_M(c)$ for all c by strict convexity of $\text{cv}_{1-\alpha}(\widehat{r}(c)) \cdot \widehat{s}_M(c)$. This reasoning implies that $\vartheta(c)$ can have at most two roots, and must be of form (ii) in this case.

Now suppose that $\lim_{c\to\pm\infty} \vartheta(c) = 0$, which is an event that only occurs if $\widehat{\tau}_T = \pm\text{cv}_{1-\alpha}(\widehat{r}_T(c)) \cdot \widehat{s}_T(c)$. It then follows from strict convexity of $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ that $\vartheta(c)$ cannot have three roots. $\vartheta(c)$ is therefore of form (i) if it does not have any root, and otherwise of form (iv). This completes the proof. □

2.A.3. **Proof of Theorem 2.3.** We begin by giving a formal description of a bias-aware DM CI. Recall the definition of U_i from Section 2.3.2, and let $b_U(h) = \mathbb{E}(\widehat{\tau}_U(h)|\mathcal{X}_n)$ and $s_U(h) = \mathbb{V}(\widehat{\tau}_U(h)|\mathcal{X}_n)^{1/2}$ denote conditional bias and standard deviation, respectively, of the SRD-type estimator $\widehat{\tau}_U(h)$. Exploiting linearity, one can write

$$b_U(h) = \sum_{i=1}^n w_i(h)(\mu_U(X_i) - \tau_U) \text{ and } s_U(h) = \left(\sum_{i=1}^n w_i(h)^2\sigma_{U,i}^2\right)^{1/2},$$

where $\mu_U(x) = (\mu_Y(x) - \tau_Y)/\tau_T - \tau_Y(\mu_T(x) - \tau_T)/\tau_T^2$ is a linear combination of the functions μ_Y and μ_T, and $\sigma_{U,i}^2 = \mathbb{V}(U_i|X_i)$ is the conditional variance of U_i given X_i. Since the bias depends on (μ_Y, μ_T) through the function $\mu_U \in \mathcal{F}_H(B_Y/|\tau_T| + |\tau_Y|B_T/\tau_T^2)$ only, its "worst case" magnitude over the functions contained in $\mathcal{F}^\delta$ is

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} |b_U(h)| = \overline{b}_U(h) \equiv -\frac{1}{2}\left(\frac{B_Y}{|\tau_T|} + \frac{|\tau_Y|B_T}{\tau_T^2}\right)\sum_{i=1}^n w_i(h)X_i^2\text{sign}(X_i).$$

An infeasible bias-aware DM CI is then given by

$$\mathcal{C}_\Delta^\alpha = \left[\widehat{\theta}(h_U) \pm \text{cv}_{1-\alpha}\left(\overline{b}_U(h_U)/s_U(h_U)\right)s_U(h_U)\right],$$

where $h_U = \arg\min_h \text{cv}_{1-\alpha}\left(\overline{b}_U(h)/s_U(h)\right)s_U(h)$ is the bandwidth that minimizes its length.

Making this CI feasible would require three main modifications. First, replacing the unknown bias bound with an estimate $\widehat{\overline{b}}_U(h)$ which replaces τ_Y and τ_T with feasible estimates (obvious candidates would be local linear estimates $\widehat{\tau}_Y = \widehat{\tau}_Y(g_Y)$ and $\widehat{\tau}_T = \widehat{\tau}_T(g_T)$ based on preliminary bandwidths g_Y and g_T). Second, replacing the standard

deviation $s_U(h)$ with a valid standard error (this could be achieved as in Section 2.6.1, using estimates $\widehat{U}_i = (Y_i - \widehat{\tau}_Y)/\widehat{\tau}_T - \widehat{\tau}_Y(T_i - \widehat{\tau}_T)/\widehat{\tau}_T^2$ of the U_i). Third, replacing the bandwidth h_U with a suitable empirical analogue (such as an adaptation of the restricted plug-in procedure described in Section 6.2). Since such modifications can be shown not to affect the asymptotic coverage properties of the CI under standard additional regularity conditions, we simply base our result on a comparison of $\mathcal{C}_*^\alpha$ and $\mathcal{C}_\Delta^\alpha$.

To prove Theorem 2.3, we now make the dependence of quantities like $h_M(c)$ on c again explicit in our notation. We begin by noting that the events $\theta^{(n)} \in \mathcal{C}_\Delta^\alpha$ and $\theta^{(n)} \in \mathcal{C}_*^\alpha$ occur if and only if

$$\frac{|\widehat{\theta}(h_U) - \theta^{(n)}|}{s_U(h_U)} - \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_U(h_U)}{s_U(h_U)}\right) \leq 0 \tag{A.2.9}$$

$$\text{and} \quad \frac{|\widehat{\tau}_M(h_M(\theta^{(n)}), \theta^{(n)})|}{s_M(h_M(\theta^{(n)}), \theta^{(n)})} - \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_M(h_M(\theta^{(n)}), \theta^{(n)})}{s_M(h_M(\theta^{(n)}), \theta^{(n)})}\right) \leq 0, \tag{A.2.10}$$

respectively. Since the left-hand sides of the last two displays are both approximated by a constant plus the absolute value of a normal random variable with variance 1 in large samples, it suffices to show that the difference between the respective left-hand sides of the last two displays converges to zero in probability, uniformly over $\mathcal{F}^\delta$. To show this, note first that standard delta method arguments yield that the left-hand side of (A.2.9) is equal to

$$\frac{|\widehat{\tau}_U(h_U) - \kappa n^{-2/5}|}{s_U(h_U)} - \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_U(h_U)}{s_U(h_U)}\right) + o_{P,\mathcal{F}^\delta}(1).$$

Next, note that $U_i = M_i(\theta)/\tau_T$, and that we thus have that

$$\widehat{\tau}_U(h) = \frac{\widehat{\tau}_M(h,\theta)}{\tau_T}, \qquad s_U(h) = \frac{s_M(h,\theta)}{|\tau_T|}, \qquad \bar{b}_U(h) = \frac{\bar{b}_M(h,\theta)}{|\tau_T|},$$

for any $h > 0$. Substituting these identities into the definition of h_U, we also find that

$$h_U = \arg\min_h \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_M(h,\theta)}{s_M(h,\theta)}\right) \cdot \frac{s_M(h,\theta)}{|\tau_T|} = \arg\min_h \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_M(h,\theta)}{s_M(h,\theta)}\right) s_M(h,\theta) = h_M(\theta).$$

The left-hand side of (A.2.9) is thus equal to

$$\frac{|\widehat{\tau}_M(h_M(\theta),\theta) - \tau_T \kappa n^{-2/5}|}{s_M(h_M(\theta),\theta)} - \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_M(h_M(\theta),\theta)}{s_M(h_M(\theta),\theta)}\right) + o_{P,\mathcal{F}^\delta}(1).$$

Now consider the term on the left-hand side of (A.2.10). By simple algebra, we have that

$$\begin{aligned}\bar{b}_M(h,\theta^{(n)}) &= \bar{b}_M(h,\theta) + n^{-2/5}|\kappa|\bar{b}_T(h),\\ s_M(h,\theta^{(n)}) &= s_M(h,\theta) + n^{-2/5}|\kappa|(s_T(h) - 2\tilde{s}_{M(\theta),T}(h)),\end{aligned}$$

with $\tilde{s}_{M(\theta),T}(h) = (\sum_{i=1}^n w_i(h)^2\sigma_{M(\theta),T,i})^{1/2}$ a conditional covariance term of the same order as $s_T(h)$. These identities imply that evaluation at $\theta^{(n)}$ does not change the leading terms of the (conditional) bias and the standard deviation (which are of order h^2 and $1/\sqrt{nh}$, respectively) relative to evaluation at θ. Since the leading term of $h_M(\theta)$ is a smooth transformation of the leading terms of the bias and standard deviation, this means that $h_M(\theta^{(n)}) = h_M(\theta)(1+o_{P,\mathcal{F}^\delta}(1))$. Arguing as in the proof of Lemma A.2.1, the left-hand side of (A.2.10) is thus equal to

$$\frac{|\widehat{\tau}_M(h_M(\theta),\theta) - \tau_T\kappa n^{-2/5}|}{s_M(h_M(\theta),\theta)} - \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_M(h_M(\theta),\theta)}{s_M(h_M(\theta),\theta)}\right) + o_{P,\mathcal{F}^\delta}(1),$$

which completes the proof. □

2.A.4. **Proof of Theorem 2.4.** We split the proof into two parts, and first show that

$$s_M(h_M) = \widehat{s}_M(h_M)(1+o_{P,\mathcal{F}}(1)). \tag{A.2.11}$$

This part is similar in structure to that of Abadie and Imbens (2006, Theorem 6). To simplify the presentation, we suppress the dependence on c of various quantities that appear in this proof. For example, we write $\widehat{s}^2_M(h_M)$ instead of $\widehat{s}^2_M(h_M(c),c)$, etc. We also define

$$q_i(h_M) = \frac{w_i(h_M)^2}{\sum_{i=1}^n w_i(h_M)^2\sigma^2_{M,i}},$$

so that $\sum_{i=1}^n q_i(h_M)\widehat{\sigma}^2_{M,i} = \widehat{s}^2_M(h_M)/s^2_M(h_M)$. We note that $\max_{i=1,\ldots,n} q_i(h_M) = o_{P,\mathcal{F}}(1)$ and $\sum_{i=1}^n q_i(h_M) = O_{P,\mathcal{F}}(1)$ by the same arguments as in the proof of Theorem 2, and the fact that the variance terms $\sigma^2_{M,i}$ are uniformly bounded and bounded away from zero, respectively.

The proof for the case that Assumption LL1 holds is rather straightforward. As we the kernel has compact support by Assumption 2.1, and h_M is bounded as a function of n, the number of support points at which $q_i(h_M) > 0$ is finite. It follows that $\sum_{i=1}^n \mathbf{1}\{X_i = x\}$ tends to infinity for all support points x with $q_i(h_M) > 0$ if $X_i = x$. Moreover, it holds that

$$\max_{i:q_i(h_M)>0} |\widehat{\sigma}^2_{M,i} - \sigma^2_{M,i}| = o_{P,\mathcal{F}}(1).$$

Since $\sum_{i=1}^n q_i(h_M) = O_{P,\mathcal{F}}(1)$ and $q_i(h_M)$ is positive, the statement of the theorem then

follows because

$$\left|\frac{\widehat{s}^2_M(h_M)}{s^2_M(h_M)} - 1\right| = \left|\sum_{i=1}^n q_i(h_M)(\widehat{\sigma}^2_{M,i} - \sigma^2_{M,i})\right| \leq \max_{i:q_i(h_M)>0} |\widehat{\sigma}^2_{M,i} - \sigma^2_{M,i}| \cdot \sum_{i=1}^n q_i(h_M) = o_{P,\mathcal{F}}(1).$$

Now suppose that Assumption LL2 holds. In this case there are no ties in the data, and each unit has exactly $R_i = R$ nearest neighbors, with probability 1. We thus define the $R \times 2$ matrix $\widetilde{X}_{-i} = (\widetilde{X}'_{r_1}, \ldots, \widetilde{X}'_{r_R})'$, where $r_1, \ldots, r_R$ are the indices of the R nearest neighbors of unit i, and $\widetilde{X}_i = (1, X_i)$, let $H_i = \widetilde{X}_i(\widetilde{X}'_{-i}\widetilde{X}_{-i})^{-1}\widetilde{X}'_i$, and write $v_j(X_i) = \widetilde{X}_i(\widetilde{X}'_{-i}\widetilde{X}_{-i})^{-1}\widetilde{X}'_{-i}e_j$ with e_j the jth R-dimensional unit-vector. With W_i a generic random variable, we also write $\breve{W}_i = W_i - \sum_{j\in\mathcal{R}_i} v_j(X_i)W_j$. In the following, we use repeatedly that

$$\sum_{j\in\mathcal{R}_i} v_j(X_i) = 1, \quad \sum_{j\in\mathcal{R}_i} v_j(X_i)(X_j - X_i) = 0, \text{ and } \sum_{j\in\mathcal{R}_i} v_j(X_i)^2 = H_i,$$

which follows from basic algebra. Next, note that the variance estimators $\widehat{\sigma}^2_{M,i}$, $i = 1, \ldots, n$, are all well-defined with probability one, as the running variable is continuously distributed with a bounded density function. Also, recall that $M_i = Y_i - cT_i$, that $\mathbb{E}(M_i|X_i) = \mu_M(X_i) = \mu_Y(X_i) - c\mu_T(X_i)$, put $\varepsilon_i = M_i - \mu_M(X_i)$, and note that $\varepsilon_i = \varepsilon_{Y,i} - c\varepsilon_{T,i} = (Y_i - \mu_Y(X_i)) - c(T_i - \mu_T(X_i))$. The variance estimators can then be written as

$$\widehat{\sigma}^2_{M,i} = \frac{\breve{M}^2_i}{1+H_i} = \frac{1}{1+H_i}\left(\breve{\mu}_M(X_i) + \varepsilon_i - \sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_i\right)^2$$

It then suffices to show the following:

$$\left|\sum_{i=1}^n q_i(h_M)(\sigma^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n])\right| = o_{P,\mathcal{F}}(1) \text{ and} \tag{A.2.12}$$

$$\left|\sum_{i=1}^n q_i(h_M)(\widehat{\sigma}^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n])\right| = o_{P,\mathcal{F}}(1), \tag{A.2.13}$$

We begin by noting that (A.2.12) follows from the triangle inequality and the fact that $\sum_{i=1}^n q_i(h_M) = O_{P,\mathcal{F}}(1)$ if

$$\max_{i=1,\ldots,n} |\sigma^2_{M,i} - \mathbb{E}[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n]| = o_{P,\mathcal{F}}(1). \tag{A.2.14}$$

To show (A.2.14), note that

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2_{M,i}|\mathcal{X}_n\right] &= \frac{1}{1+H_i}\mathbb{E}\left[\left(\breve{\mu}_M(X_i)+\varepsilon_i-\sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_j\right)^2|\mathcal{X}_n\right] \\ &= \frac{1}{1+H_i}\left(\breve{\mu}_M(X_i)^2+\sigma^2_{M,i}+\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\sigma^2_{M,j}\right) \\ &= \sigma^2_{M,i}+\frac{1}{1+H_i}\left(\breve{\mu}_M(X_i)^2+\sum_{j\in\mathcal{R}_i} v_j(X_i)^2(\sigma^2_{M,j}-\sigma^2_{M,i})\right).\end{aligned}$$

Here the second equality holds because ε_i and ε_j are independent if $i \neq j$, and are zero in expectation; and the third equality holds because $\sum_{j\in\mathcal{R}_i} v_j(X_i)^2 = H_i$. As the running variable density is uniformly bounded away from zero, it follows from the proof of Theorem 6 in Abadie and Imbens (2006) that

$$x_{\max} \equiv \max_{i=1,\ldots,n}\max_{r\in\mathcal{R}_i}|X_i - X_r| = o_{P,\mathcal{F}}(1). \tag{A.2.15}$$

Since $\sigma^2_{M,i}$ is uniformly Lipschitz continuous with some constant L_σ by Assumption 2.1, we then have that

$$\begin{aligned}\max_i \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)^2(\sigma^2_{M,j}-\sigma^2_{M,i})\right) &\leq L_\sigma x_{\max}\max_i \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\right) \\ &\leq L_\sigma x_{\max}\max_i \frac{H_i}{1+H_i} = o_{P,\mathcal{F}}(1).\end{aligned}$$

To show (A.2.14), it thus only remains to show that

$$\max_i \frac{1}{1+H_i}\breve{\mu}_M(X_i)^2 = o_{P,\mathcal{F}}(1). \tag{A.2.16}$$

To do so, note that

$$\begin{aligned}&\max_{i\in\{1,\ldots,n\}}\left(\mu_M(X_i)-\sum_{j\in\mathcal{R}_i} v_j(X_i)\mu_M(X_j)\right) \\ &\quad= \max_{i\in\{1,\ldots,n\}}\left(\mu_M(X_i)-\sum_{j\in\mathcal{R}_i} v_j(X_i)(\mu_M(X_i)+\mu'_M(X_i)(X_j-X_i)\right. \\ &\qquad\qquad\left.+\frac{1}{2}\mu''_M(\mathring{X}_{i,j})(X_j-X_i)^2)\right) \\ &\quad= \frac{1}{2}\max_{i\in\{1,\ldots,n\}}\sum_{j\in\mathcal{R}_i} v_j(X_i)\mu''_M(\mathring{X}_{i,j})(X_j-X_i)^2.\end{aligned}$$

Here the first equality follows from a second order expansion, with $\mathring{X}_{i,j}$ some value between X_i and X_j, where $j \in \mathcal{R}_i$; and the second equality follows as $\sum_{j\in\mathcal{R}_i} v_j(X_i) = 1$ and $\sum_{j\in\mathcal{R}_i} v_j(X_i)(X_j - X_i) = 0$. We then find that

$$\begin{aligned}
\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\breve{\mu}_M(X_i)^2 &= \frac{1}{4}\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)\mu_M''(\mathring{X}_{i,j})(X_j - X_i)^2\right)^2 \\
&\leq \frac{R}{4}\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\mu_M''(\mathring{X}_{i,j})^2(X_j - X_i)^4 \\
&\leq \frac{RB_M^2 x_{\max}^4}{4}\max_{i\in\{1,\ldots,n\}} \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_j(X_i)^2\right) = o_{P,\mathcal{F}}(1).
\end{aligned}$$

Here first inequality follows from Cauchy-Schwarz as the cardinality of $\mathcal{R}_i$ is R; and the second inequality follows as all the terms of the sum are positive, $\mu''(\mathring{X}_{i,j})^2$ is bounded by B_M^2, and $(X_j - X_i)^4 \leq x_{\max}^4$ for all i and $j \in \mathcal{R}_i$. The final equality follows because $\sum_{j\in\mathcal{R}_i} v_j(X_i)^2 = H_i$, and $H_i/(1+H_i) \leq 1$ for all $i \in \{1,\ldots,n\}$, and $x_{\max} = o_{P,\mathcal{F}}(1)$. This completes the proof of the statement (A.2.12).

To show that (A.2.13) holds, write $\widetilde{q}_i(h_M) = q_i(h_M)(1+H_i)^{-1}$. Note that since $|\widetilde{q}_i(h_M)| \leq |q_i(h_M)|$, it follows from Theorem A.2.1 that $\max_{i=1,\ldots,n}\widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1)$ and $\sum_{i=1}^n \widetilde{q}_i(h_M) = O_{P,\mathcal{F}}(1)$. We write this quantity the sum of five terms:

$$\begin{aligned}
&\sum_{i=1}^n q_i(h_M)(\widehat{\sigma}_{M,i}^2 - \mathbb{E}[\widehat{\sigma}_{M,i}^2|\mathcal{X}_n]) \\
&= \sum_{i=1}^n \widetilde{q}_i(h_M)(\varepsilon_i^2 - \sigma_{M,i}^2) + \sum_{i=1}^n \widetilde{q}_i(h_M)\sum_{j\in\mathcal{R}_i} v_j^2(X_i)(\varepsilon_j^2 - \sigma_{M,j}^2) \\
&\quad + 2\sum_{i=1}^n \widetilde{q}_i(h_M)\varepsilon_i\sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_j + 2\sum_{i=1}^n \widetilde{q}_i(h_M)\breve{\mu}_M(X_i)\varepsilon_i \\
&\quad - 2\sum_{i=1}^n \widetilde{q}_i(h_M)\breve{\mu}_M(X_i)\sum_{j\in\mathcal{R}_i} v_j(X_i)\varepsilon_j \\
&\equiv G_1 + G_2 + 2G_3 + 2G_4 + 2G_5.
\end{aligned}$$

It is easy to see that these five terms all have mean zero conditional on $\mathcal{X}_n$. It thus suffices to show that their second moments converge uniformly over the function class $\mathcal{F}$ to zero. In the following derivations, we write C for a generic positive constant whose value might differ between equations.

For the first term, we have that

$$\mathbb{V}(G_1|\mathcal{X}_n) = \sum_{i=1}^{n} \widetilde{q}_i(h_M)^2 \mathbb{E}[(\varepsilon_i^2 - \sigma_{M,i}^2)^2|\mathcal{X}_n] \leq C \max_{i=1,\ldots,n} \widetilde{q}_i(h_M) \cdot \sum_{i=1}^{n} \widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1),$$

where the inequality follows from the bound on the fourth moment of ε_i and $\widetilde{q}_i(h_M)$ being positive, and the last equality follows since $\max_{i=1,\ldots,n} \widetilde{q}_i(h_M) \sum_{i=1}^{n} \widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1)$.

We now turn to the second term, and note that by independent sampling

$$\begin{aligned}
\mathbb{V}(G_2|\mathcal{X}_n) &= \sum_{i=1}^{n}\sum_{l=1}^{n} \widetilde{q}_l(h_M)\widetilde{q}_i(h_M) \sum_{j\in\mathcal{R}_i}\sum_{k\in\mathcal{R}_l} v_k^2(X_l)v_j^2(X_i)\mathbb{E}\left[(\varepsilon_j^2 - \sigma_{M,j}^2)(\varepsilon_k^2 - \sigma_{M,k}^2)|\mathcal{X}_n\right] \\
&= \sum_{i=1}^{n} \sum_{l:\mathcal{R}_i\cap R_l\neq\emptyset} \widetilde{q}_l(h_M)\widetilde{q}_i(h_M) \\
&\qquad \cdot \sum_{j\in\mathcal{R}_i}\sum_{k\in\mathcal{R}_l} v_k^2(X_l)v_j^2(X_i)\mathbb{E}\left[(\varepsilon_j^2 - \sigma_{M,j}^2)(\varepsilon_k^2 - \sigma_{M,k}^2)|\mathcal{X}_n\right] \\
&\leq \sum_{i=1}^{n} \sum_{l:\mathcal{R}_i\cap\mathcal{R}_l\neq\emptyset} \widetilde{q}_l(h_M)\widetilde{q}_i(h_M) \sum_{j\in\mathcal{R}_i}\sum_{k\in\mathcal{R}_l} v_k^2(X_l)v_j^2(X_i)\mathbb{E}\left[(\varepsilon_j^2 - \sigma_{M,j}^2)^2|\mathcal{X}_n\right].
\end{aligned}$$

Using that ε_i has bounded fourth moments, that $\sum_{k\in\mathcal{R}_l} v_k^2(X_i) = H_i$, and that $H_i/(1+H_i) \leq 1$ for all $i \in \{1,\ldots,n\}$, we further deduce that

$$\mathbb{V}(G_2|\mathcal{X}_n) \leq C \sum_{i=1}^{n} q_i(h_M) \sum_{l:\mathcal{R}_i\cap\mathcal{R}_l\neq\emptyset} q_l(h_M).$$

Finally, note that the cardinality of the set $\{l : \mathcal{R}_i \cap \mathcal{R}_l \neq \emptyset\}$, which contains the indices of those units that share at least one common R-nearest neighbor with unit i, is bounded by $3R+1$ (this can be seen through a simple counting exercise). We thus have that

$$\mathbb{V}(G_2|\mathcal{X}_n) \leq C \sum_{i=1}^{n} q_i(h_M)(3R+1) \max_{j\in\{1,\ldots,n\}} q_j(h_M) = o_{P,\mathcal{F}}(1).$$

We now consider the third term, which satisfies

$$\mathbb{V}(G_3|\mathcal{X}_n) = \sum_{i=1}^{n}\sum_{k=1}^{n} \widetilde{q}_i(h_M)\widetilde{q}_k(h_M) \sum_{j\in\mathcal{R}_i}\sum_{l\in\mathcal{R}_k} v_j(X_i)v_l(x_g)\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l|\mathcal{X}_n].$$

To proceed, note that $\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l|\mathcal{X}_n] = 0$ unless the four indices involved in this expression

can be grouped into two pairs that each have the same value. This means that

$$\begin{aligned}\mathbb{V}(G_3|\mathcal{X}_n) &\leq C\sum_{i=1}^{n}\left(\sum_{j\in\mathcal{R}_i}\widetilde{q}_i(h_M)^2 v_j(X_i)^2 + \sum_{j\in R_i: i\in R_j}\widetilde{q}_i(h_M)\widetilde{q}_j(h_M)v_i(X_j)v_j(X_i)\right)\\ &\leq C\max_{i\in\{1,\ldots,n\}}\widetilde{q}_i(h_M)\sum_{i=1}^{n}\widetilde{q}_i(h_M)\sum_{j\in\mathcal{R}_i}v_j(X_i)^2\\ &= C\max_{i\in\{1,\ldots,n\}}\widetilde{q}_i(h_M)\sum_{i=1}^{n}q_i(h_M)\frac{H_i}{1+H_i} = o_{P,\mathcal{F}}(1).\end{aligned}$$

For the fourth and fifth term, we can use arguments similar to those used for the three previous terms to show that that

$$\begin{aligned}\mathbb{V}(G_4|\mathcal{X}_n) &\leq CB_M^2 x_{\max}^4\sum_{i=1}^{n}\widetilde{q}_i(h_M)^2 = o_{P,\mathcal{F}}(1);\\ \mathbb{V}(G_5|\mathcal{X}_n) &\leq CB_M^2 x_{\max}^4\max_{i\in\{1,\ldots,n\}}\left(q_i(h_M)\frac{H_i}{1+H_i}\right)\sum_{i=1}^{n}\widetilde{q}_i(h_M) = o_{P,\mathcal{F}}(1).\end{aligned}$$

This completes the proof of the statement (A.2.13); and thus (A.2.11) holds, as claimed.

In the second part of our proof, we show that

$$\widehat{s}_M(\widehat{h}_M) = \widehat{s}_M(h_M)(1+o_{P,\mathcal{F}}(1)). \tag{A.2.17}$$

Equations (A.2.11) and (A.2.17) then imply together the statement of Theorem 2.4 and thus the proof is completed.

First suppose that Assumption LL1 holds. Equation (A.2.17) follows trivially in this case similarly to the arguments of Lemma A.2.1.

Now suppose that Assumption LL2 holds. Statements (ii)–(iii) of Lemma A.2.1 then follow from arguments analogous to those in the proof of Theorem E.1 in Armstrong and Kolesár (2020). A similar line of reasoning can be used to show Assumption 2.2(iv). We describe the latter argument in detail. Since $s_M^2(h_M) = O_P((nh_M)^{-1})$, it suffices to show that

$$nh_M(\widehat{s}_M^2(\widehat{h}_M) - \widehat{s}_M^2(h_M)) = o_{P,\mathcal{F}}(1).$$

To do so, write $\widehat{\eta}_i = \widehat{\sigma}_{M,i}^2 - \mathbb{E}[\widehat{\sigma}_{M,i}^2|\mathcal{X}_n]$, and note that

$$\widehat{s}_M^2(\widehat{h}_M) - \widehat{s}_M^2(h_M) = \sum_{i=1}^{n}(w_i(\widehat{h}_M)^2 - w_i(h_M)^2)\widehat{\eta}_i + \sum_{i=1}^{n}(w_i^2(\widehat{h}_M) - w_i^2(h_M))\mathbb{E}[\widehat{\sigma}_{M,i}^2|\mathcal{X}_n].$$

Above, we showed that $\max_{i=1,\ldots,n}|\sigma_{M,i}^2 - \mathbb{E}[\widehat{\sigma}_{M,i}^2|\mathcal{X}_n]| = o_{P,\mathcal{F}}(1)$, and by assumption the

conditional variance terms $\sigma^2_{M,i}$ are bounded. We thus only need to show that

$$nh \sum_{i=1}^{n} |w_i(\widehat{h}_M)^2 - w_i(h_M)^2| = o_{P,\mathcal{F}}(1), \tag{A.2.18}$$

$$nh \left| \sum_{i=1}^{n} (w_i(\widehat{h}_M)^2 - w_i(h_M)^2)\widehat{\eta}_i \right| = o_{P,\mathcal{F}}(1). \tag{A.2.19}$$

We show this using arguments analogous to those used in the proof of Theorem E.1 in Armstrong and Kolesár (2020), the main difference being that in their proof the analogue of $\widehat{\eta}_i$ is i.i.d., whereas in our case these terms are generally not independent. By the triangle inequality, it suffices to show that both (A.2.18) and (A.2.19) hold with $w_{+,i}(\cdot)^2$ replacing $w_i(\cdot)^2$, as the same arguments apply to $w_{-,i}(\cdot)^2$. These weights can be written as

$$w^2_{+,i}(h) = \frac{1}{nh}\varphi(h)'\psi_i(h)\varphi(h), \quad \text{where}$$
$$\varphi(h) = \left(\frac{1}{nh} \sum_{i:X_i>0} K(X_i/h)\widetilde{X}_i'\widetilde{X}_i \right)^{-1} e_1, \quad \psi_i(h_M) = K(X_i/h)^2\widetilde{X}_i'\widetilde{X}_i/(nh).$$

Let $\|\cdot\|$ be the L_1-norm of a vector or a matrix. By the triangular inequality, it follows that the left-hand side of (A.2.18) is bounded by

$$\left((2\,\|\varphi(h_M)| + \|\varphi(\widehat{h}_M) - \varphi(h_M)\|) \sum_{i:Z_i=1} \|\psi_i(\widehat{h}_M)\| + \|\varphi(h_M)\|^2 \sum_{i:Z_i=1} \|\psi_i(\widehat{h}_M) - \psi_i(h_M)\| \right)$$
$$\times \|\varphi(\widehat{h}_M) - \varphi(h_M)\|.$$

As shown in the proof of Lemma E.1 in Armstrong and Kolesár (2020), this term is of the order $o_{P,\mathcal{F}}(1)$. Moreover, equation (A.2.19) is bounded by

$$||\varphi(\widehat{h}_M)||^2 || \sum_{i=1}^{n} (\psi_n(x_i, \widehat{h}_M) - \psi_n(x_i, h_M))\hat{\eta}_i || +$$
$$\|\varphi_n(\widehat{h}_M) - \varphi_n(h_M)\| \left(2\|\varphi_n(\widehat{h}_M)\|| \sum_{i=1}^{n} \psi_n(x_i, \widehat{h}_M) - \psi_n(x_i, h_M)|| \right.$$
$$\left. + | \sum_{i=1}^{n} \psi_n(x_i, \widehat{h}_M)|| \left(||\varphi_n(\widehat{h}_M) - \varphi_n(h_M)|| + 2||\varphi(\widehat{h}_M)|| \right) \right).$$

By Lemma E.1 Armstrong and Kolesár (2020) it follows that $\|\varphi(h_M)\|^2 = O_{P,\mathcal{F}}(1)$ and $\|\varphi(\widehat{h}_M) - \varphi(h_M)\| = o_{P,\mathcal{F}}(1)$. It therefore suffices to show that $\| \sum_{i:Z_i=1} (\psi_i(\widehat{h}_M) - \psi_i(h_M))\widehat{\eta}_i \| = o_{P,\mathcal{F}}(1)$. The elements of $\psi_i(h)$ are given by the function $g(z) = z^v K(z)^w$

for $v,\ w\ \in \{0,1,2\}$. We therefore show that for all $\varepsilon > 0$

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{s \in [1-\delta, 1+\delta]} \left| \sqrt{nh} \sum_{i:Z_i=1} (g(X_i/(sh_M)) - g(X_i/h_M))\widehat{\eta}_i \right| > \varepsilon \right) = 0. \quad \text{(A.2.20)}$$

For δ small enough, it holds that for s and $\widetilde{s}$ in a neighborhood of 1, and C a positive constant that can take different values at different occurrences, that

$$\begin{aligned} &\mathbb{E}\left[\left(\sum_{i:Z_i=1} (g(X_i/sh_M) - g(X_i/\widetilde{s}h_M))\widehat{\eta}_i \right)^2 \right] \\ &\quad \leq \frac{C}{nh_M} \sum_{i:Z_i=1} \left(g(X_i/(sh_M)) - g(X_i/(\widetilde{s}h_M)) \right)^2 \\ &\quad \leq |1/s - 1/\widetilde{s}|^2 \frac{C}{nh_M} \sum_{i:Z_i=1} \mathbf{1}\{X_i/h_M \leq C\}. \end{aligned} \quad \text{(A.2.21)}$$

Here the first inequality holds because $\widehat{\eta}_i$ has a finite second moment, Cauchy-Schwarz, and the fact that for all i the cardinality of the set of indicies j such that $\widehat{\eta}_i$ contains data points that are also used in $\widehat{\eta}_j$ is bounded by a finite constant (this is shown in the proof of Theorem 2.4). The second inequality then holds because the function $g(\cdot)$ is Lipschitz continuous and the kernel is bounded from above with compact support. For n large enough, the term in (A.2.21) is bounded by $|1/s - 1/\widetilde{s}|^2$ times a constant that does not depend on the sample size. Equation (A.2.20) then follows from Example 2.2.12 in van der Vaart and Wellner (1996). This completes our proof. □

2.B. MORE GENERAL BANDWIDTH CHOICES

In the main body of the paper, the local linear regression estimators $\widehat{\tau}_M(h,c) = \widehat{\tau}_Y(h) - c\widehat{\tau}_T(h)$ on which our bias-aware AR CSs are based use the same bandwidth on each side of the cutoff, and also the same bandwidth for estimating τ_Y and τ_T. It is also imposed that the second derivatives of μ_Y and μ_T are bounded in absolute value by the same respective constant on either side of the cutoff. These features can all easily be relaxed. In particular, we can define a more general Hölder-type class of functions as

$$\mathcal{F}_H(B_+, B_-) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_1''\|_\infty \leq B_+, \|f_0''\|_\infty \leq B_-\},$$

define the class $\mathcal{F}_H^\delta(B_+, B_-)$ similarly, and then seek to obtain bias-aware AR CSs that are honest uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_{Y+}, B_{Y-}) \times \mathcal{F}_H^0(B_{T+}, B_{T-})$, based on the local

linear regression estimator

$$\widehat{\tau}_M(\mathbf{h}, c) = \sum_{i=1}^{n} \left(w_{i,+}(h_{Y+}) - w_{i,-}(h_{Y-})\right) Y_i - c \sum_{i=1}^{n} \left(w_{i,+}(h_{T+}) - w_{i,-}(h_{T-})\right) T_i,$$

where $\mathbf{h} = (h_{T+}, h_{T-}, h_{Y+}, h_{Y-})$ is a vector of side- and function-specific bandwidths, and the weights $w_{i,+}(h)$ and $w_{i,-}(h)$ are as defined in the beginning of Appendix 2.A in the main body of the paper. With such a setup, the explicit expression for the bound on the absolute value of the conditional bias of $\widehat{\tau}_M(\mathbf{h}, c)$ is

$$\begin{aligned}\overline{b}_M(\mathbf{h}, c) = &-\frac{B_{Y+}}{2} \sum_{i=1}^{n} w_{i,+}(h_{Y+}) X_i^2 - \frac{|c| B_{T+}}{2} \sum_{i=1}^{n} w_{i,+}(h_{T+}) X_i^2 \\ &+ \frac{B_{Y-}}{2} \sum_{i=1}^{n} w_{i,-}(h_{Y-}) X_i^2 + \frac{|c| B_{T-}}{2} \sum_{i=1}^{n} w_{i,-}(h_{T-}) X_i^2,\end{aligned}$$

and the conditional standard deviation of $\widehat{\tau}_M(\mathbf{h}, c)$ is

$$\begin{aligned}s_M(\mathbf{h}, c) = \Bigg(&\sum_{i=1}^{n} \left(w_{i,+}(h_{Y,+}) - w_{i,-}(h_{Y-})\right)^2 \sigma_{Y,i}^2 + c^2 \sum_{i=1}^{n} \left(w_{i,+}(h_{T+}) - w_{i,-}(h_{T-})\right)^2 \sigma_{T,i}^2 \\ &-2c \sum_{i=1}^{n} \left(w_{i,+}(h_{Y+}) - w_{i,-}(h_{Y-})\right) \left(w_{i,+}(h_{T+}) - w_{i,-}(h_{T-})\right) \sigma_{YT,i} \Bigg)^{1/2},\end{aligned}$$

with $\sigma_{Y,i}^2 = \mathbb{V}(Y_i|X_i)$, $\sigma_{T,i}^2 = \mathbb{V}(T_i|X_i)$, and $\sigma_{YT,i} = \mathbb{C}(Y_i, T_i|X_i)$ being conditional variance and covariance terms. A feasible standard error $\widehat{s}_M(\mathbf{h}, c)$ can be obtained by substituting nearest-neighbor estimates of the latter terms into the above expression for $s_M(\mathbf{h}, c)$. Letting $\widehat{\mathbf{h}}_M(c)$ be a feasible estimate of $\mathbf{h}_M(c) = \arg\min_{\mathbf{h}} \mathrm{cv}_{1-\alpha}(r_M(\mathbf{h}, c)) \cdot s_M(\mathbf{h}, c)$, with $r_M(\mathbf{h}, c) = \overline{b}_M(\mathbf{h}, c)/s_M(\mathbf{h}, c)$, a generalization of our proposed bias-aware AR CS for θ is then given by

$$\mathcal{C}_{\mathrm{ar}}^{\alpha} = \left\{ c : |\widehat{\tau}_M(\widehat{\mathbf{h}}_M(c), c)| \leq \mathrm{cv}_{1-\alpha}(\widehat{r}_M(\widehat{\mathbf{h}}_M(c), c)) \widehat{s}_M(\widehat{\mathbf{h}}_M(c), c)) \right\}.$$

A theoretical analysis of this CS would follow arguments that are fully analogous to those in the analysis of the CS in the main body of this chapter, which only uses a single bandwidth, and would yield fully analogous results.

2.C. PROPERTIES OF RULE-OF-THUMB SMOOTHNESS BOUNDS

In this appendix, we study the properties of two data-driven rules-of-thumb (ROT) for selecting the smoothness constants B_Y and B_T, which are both based on fitting global polynomial specifications on either side of the cutoff. For simplicity, we focus on the case of B_Y, but the arguments apply analogously to the case of B_T. To describe the two

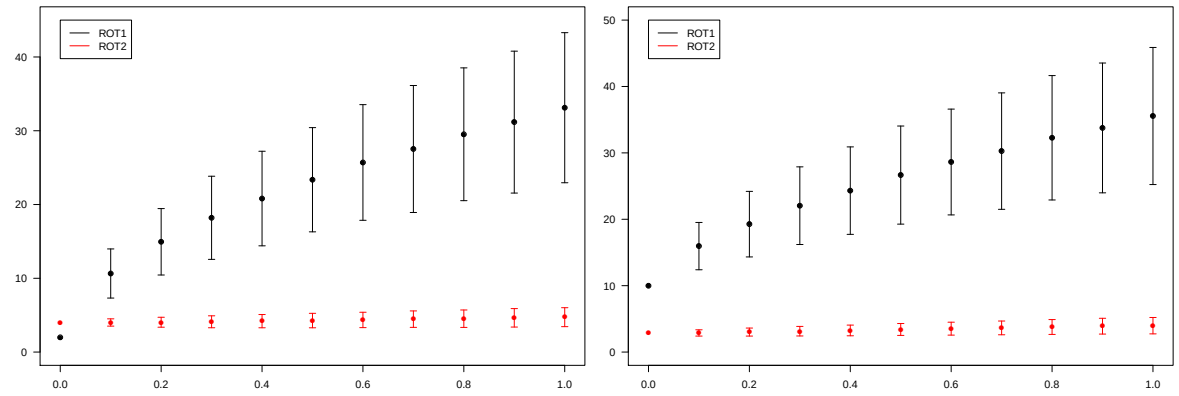


Figure A.2.1: Mean (dots) and interquartile range (bars) of simulated ROT1 (black) and ROT2 (red) "rule-of-thumb" estimates of bound on absolute second derivative for $\sigma^2 \in \{0, .1, \ldots, 1\}$ and $\mu_Y(x) = x^2$ (left panel) and $\mu_Y(x) = x^2 - x^4$ (right panel)

methods, let $g_k(x) = (1, x, \ldots, x^k, \mathbf{1}\{x \geq 0\}, \mathbf{1}\{x \geq 0\}x, \ldots, \mathbf{1}\{x \geq 0\}x^k)^\top$ be a vector of polynomials, define the function

$$\widetilde{\mu}_{Y,k}(x) = g_k(x)^\top \widehat{\gamma}_k, \text{ with } \widehat{\gamma}_k = \arg\min_\gamma \sum_{i=1}^{n} (Y_i - g_k(X_i)^\top \gamma)^2,$$

and write $\mathcal{X}$ for the range of the realizations of the running variable.Armstrong and Kolesár (2020) then consider fourth-order polynomials, and propose the ROT value

$$\widehat{B}_{Y,\text{ROT1}} = \sup_{x \in \mathcal{X}} |\widetilde{\mu}''_{Y,4}(x)|.$$

Imbens and Wager (2019) mention a ROT in which the maximal curvature implied by a quadratic fit is multiplied by some moderate factor, say 2, to guard against overly optimistic values, yielding the rule-of-thumb value

$$\widehat{B}_{Y,\text{ROT2}} = 2 \sup_{x \in \mathcal{X}} |\widetilde{\mu}''_{Y,2}(x)|.$$

We refer to these estimators ROT1 and ROT2 in the following. In principle, we would like any such rule to be close to the true smoothness bound, but not to underestimate it, so that the resulting CS has high power and correct coverage. Both Armstrong and Kolesár (2020) and Imbens and Wager (2019) caution that the respective rules cannot be expected to provide universally good smoothness bounds, and should rather serve as a first guidance that is complemented with other approaches in a sensitivity analysis.

To get a better understanding of the relative properties of these two rules, we conduct two small Monte Carlo experiments in which the conditional expectation function is either $\mu_Y(x) = x^2$ or $\mu_Y(x) = x^2 - x^4$. With each function and each $\sigma^2 \in \{0, .1, .2, \ldots, 1\}$, we

conduct $10,000$ runs in which we simulate $n = 1,000$ realizations of (Y_i, X_i) according to

$$Y_i = \mu_Y(X_i) + \varepsilon_i, \quad X_i \sim U[-1,1], \quad \varepsilon_i \sim N(0,\sigma^2), \quad X_i \perp \varepsilon_i,$$

and calculate both ROT values. If $\mu_Y(x) = x^2$, the true smallest upper bound on the absolute second derivative is $B_Y = 2$, whereas if $\mu_Y(x) = x^2 - x^4$, we have that $B_Y = 10$. In both cases, the corresponding values of "population R squared", defined as $R^2 = \mathbb{V}(\mu_Y(X_i))/\mathbb{V}(Y)$, are also within the range typically encountered in empirical studies.

We start by considering the case $\mu_Y(x) = x^2$, for which both a second and a fourth order polynomial obviously constitute a correct specification. It thus holds in this particular case that $\widehat{B}_{Y,\text{ROT1}} \xrightarrow{p} B_Y = 2$ and $\widehat{B}_{Y,\text{ROT2}} \xrightarrow{p} 2B_Y = 4$ as $n \to \infty$. That is, ROT1 consistently estimates B_Y here, while the probability limit of the ROT2 exceeds the true smoothness bound by a factor of two. A priori, one might therefore expect ROT1 to perform better than ROT2 rule in this setup. Our results, summarized in the left panel of Figure A.2.1, show that this is not the case. The distribution of ROT1 depends strongly on the error variance, and except for very small values of σ^2 the methods tends to produce vast over-estimates of B_Y. For $\sigma^2 = 1$, for example, the average across simulation runs is 33.58, which exceeds the true bound by a factor of almost 17. ROT1 is also quite volatile, which can be seen from its large interquartile range. ROT2, on the other hand, is much less affected by changes in the error variance: its mean across simulation runs increases from 4.01 for $\sigma^2 = 0.1$ to only 4.74 for $\sigma^2 = 1$, and its sampling variability is rather small.

Now consider the case $\mu_Y(x) = x^2 - x^4$, for which fourth order polynomial is clearly a correct specification. Indeed, a second order polynomial is particularly inadequate here, as the true function oscillates on either side of the cutoff. We have that $\widehat{B}_{Y,\text{ROT1}} \xrightarrow{p} 10 = B_Y$ and $\widehat{B}_{Y,\text{ROT2}} \xrightarrow{p} 2.753 \neq B_Y$ as $n \to \infty$, which means that ROT1 consistently estimates B_Y here, while the probability limit of ROT2 is about four times smaller than the true smoothness bound. Our simulation results for this setup are summarized in the right panel of Figure A.2.1. Again, ROT1 estimates are highly variable, and tend to be much larger than the true smoothness bound. The discrepancy is not as pronounced as in the previous setup though: for $\sigma^2 = 1$, for example, the average across simulation runs is 36.86, which is only 3.6 times larger than B_Y. ROT2 is again much less affected by changes in the error variance: its mean across simulation runs increases from 2.78 for $\sigma^2 = 0.1$ to only 3.99 for $\sigma^2 = 1$, and its sampling variability is rather small. But due to the severe misspecification of a second-order polynomial these values tend to severely under-estimate the true smoothness bounds.

These results first of all stress the theoretical point that no data-driven method for choosing smoothness bounds can be expected to work well under all circumstances. Still,

our exercise conveys some insight regarding under which condition one rule might be a better "first guess" than the other. Roughly speaking, the performance patterns of ROT1 can be explained by the fact that its underlying fourth order polynomial specification tends to produce erratic over-fits if the function $\mu_Y(x)$ is rather "simple", and there is a non-negligible level of noise in the data (this is a general feature of high-order polynomial regression, related to Runge's phenomenon in the literature on approximation theory). This is much less of an issue with a quadratic model. In practice, we therefore recommend using ROT2 over ROT1 in settings where one believes that μ_Y is "close" to being a "moderately" convex or concave function. If this is not the shape one has in mind there is no obvious ordering of the ROTs, and both should be considered within a more extensive sensitivity analysis.

2.D. EXTENSION TO FUZZY REGRESSION KINK DESIGNS

2.D.1. **Description.** Our approach to FRD inference described in the main body of the paper can easily be extended to the cases in which the parameter of interest is the ratio of jumps in the derivatives (of some order $v \geq 0$) of two conditional expectation functions $\mu_Y(x) = \mathbb{E}(Y|X = x)$ and $\mu_T(x) = \mathbb{E}(T|X = x)$ at the threshold value zero.[14] The most prominent example of such a setup is the Fuzzy Regression Kink Designs (Card et al., 2015), where the goal is to estimate the ratio of jumps in the first derivatives of these functions. We now sketch our extension using notation analogous to that in Section 2.4.

For a generic random variable W_i, we write $\mu_W^{(v)}(x) = \partial^v \mathbb{E}(W_i|X_i = x)/(\partial x)^v$ for the vth derivative of its conditional expectation given X_i; $\mu_{W,+}^{(v)} = \lim_{x \downarrow 0} \mu_W^{(v)}(x)$ and $\mu_{W,-}^{(v)} = \lim_{x \uparrow 0} \mu_W^{(v)}(x)$ denotes the left and right limits of the derivative at the threshold; and $\tau_{W,v} = \mu_{W,+}^{(v)} - \mu_{W,-}^{(v)}$ denotes the corresponding jump in $\mu_W^{(v)}$. Our parameter of interest is $\theta_v = \tau_{Y,v}/\tau_{T,v}$, and the goal is again to construct CSs with correct asymptotic coverage, uniformly in (μ_Y, μ_T) over some function class $\mathcal{F}$. That is, we want to construct data-dependent sets $\mathcal{C}^\alpha \subset \mathbb{R}$ that satisfy

$$\liminf_{n \to \infty} \inf_{(\mu_Y, \mu_T) \in \mathcal{F}} \mathbb{P}(\theta_v \in \mathcal{C}^\alpha) \geq 1 - \alpha \tag{A.2.22}$$

for some $\alpha > 0$. We again define $\mathcal{F}$ as a smoothness class. Specifically, let

$$\mathcal{F}_{H,p}(B) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_w^{(p+1)}\|_\infty \leq B, w = 0, 1\}$$

be the Hölder-type class of real functions that are potentially discontinuous at zero, $(p+1)$-times differentiable almost everywhere on either side of the threshold, and whose $(p+1)$th

[14] We could in principle allow the two derivatives to be of of different order, but as we are not aware of a setup that requires this we only consider identical orders here to keep the notation simple.

derivative is uniformly bounded by some constant $B > 0$. We also define the class

$$\mathcal{F}^{\delta}_{H,vp}(B) = \{f \in \mathcal{F}_{H,p}(B) : |f_+^{(v)} - f_-^{(v)}| > \delta\},$$

and assume that

$$(\mu_T, \mu_Y) \in \mathcal{F}^0_{H,vp}(B_T) \times \mathcal{F}_{H,p}(B_Y) \equiv \mathcal{F}.$$

Our CSs for the ratio of jumps in vth-order derivatives are based on pth order local polynomial regression, where $v \leq p$. Following standard results on the bias properties of local polynomial regression (Fan and Gijbels, 1996), it is generally recommended to use $p = v + 1$. For a generic dependent variable W_i, the local pth order polynomial estimator $\widehat{\tau}_{W,vp}(h)$ of $\tau_{W,v}$ is the $(p + v + 2)$th component of

$$\arg\min_{\beta \in \mathbb{R}^{2p}} \sum_{i=1}^{n} K(X_i/h)(W_i - \beta^\top (1, X_i, X_i^2/2, \ldots, X_i^p/(p!), Z_i, Z_i X_i, \ldots, Z_i X_i^p/(p!))^2,$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$ and $h > 0$ is a bandwidth. It follows from standard least squares algebra that this estimator can be written as

$$\widehat{\tau}_{W,vp}(h) = \sum_{i=1}^{n} w_{vp,i}(h) W_i, \quad w_{vp,i}(h) = w_{vp,i,+}(h) - w_{vp,i,-}(h),$$

$$w_{vp,i,+}(h) = e_{v+1}^\top Q_{p,+}^{-1} \widetilde{X}_{p,i} K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_{p,+} = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_{p,i} \widetilde{X}_{p,i}^\top \mathbf{1}\{X_i \geq 0\},$$

$$w_{vp,i,-}(h) = e_{v+1}^\top Q_{p,-}^{-1} \widetilde{X}_{p,i} K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_{p,-} = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_{p,i} \widetilde{X}_{p,i}^\top \mathbf{1}\{X_i < 0\},$$

with $\widetilde{X}_{p,i} = (1, X_i, X_i^2/2, \ldots, X_i^p/(p!))^\top$. We then obtain a bias-aware AR CS for θ_v by collecting those values of c for which an auxiliary bias-aware CI for $\tau_{M,v}(c) = \tau_{Y,v} - c\tau_{T,v}$ contains zero. To describe the construction, denote the conditional bias and standard deviation of $\widehat{\tau}_{M,vp}(h, c) = \sum_{i=1}^n w_{vp,i}(h) M_i(c)$ given $\mathcal{X}_n = (X_1, \ldots, X_n)'$ by $b_{M,vp}(h, c) = \mathbb{E}(\widehat{\tau}_{M,vp}(h, c)|\mathcal{X}_n) - \tau_{M,vp}(c)$ and $s_{M,vp}(h, c) = \mathbb{V}(\widehat{\tau}_{M,vp}(h, c)|\mathcal{X}_n)^{1/2}$, respectively. These quantities can be written more explicitly as

$$b_{M,vp}(h, c) = \sum_{i=1}^{n} w_{vp,i}(h) \mu_M(X_i, c) - (\mu_{M+}^{(v)}(c) - \mu_{M-}^{(v)}(c)),$$

$$s_{M,vp}(h, c) = \left(\sum_{i=1}^{n} w_{vp,i}(h)^2 \sigma_{M,i}^2(c) \right)^{1/2},$$

with $\sigma_{M,i}^2(c) = \mathbb{V}(M_i(c)|X_i)$ the conditional variance of $M_i(c)$ given X_i. The bias depends on (μ_Y, μ_T) through the transformation $\mu_M^{(v)} = \mu_Y^{(v)} - c \cdot \mu_T^{(v)}$ only, and $\mu_Y^{(v)} - c\mu_T^{(v)} \in$

$\mathcal{F}_{H,vp}(B_Y + |c|B_T)$. Our main contribution is to show that one can bound $b_{M,vp}(h,c)$ in absolute value over the functions contained in $\mathcal{F}$ by

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_{M,vp}(h,c)| \leq \overline{b}_{M,vp}(h,c) \equiv (-1)^{p-v}\frac{B_Y + |c|B_T}{(p+1)!}\sum_{i=1}^{n} w_{vp,i}(h)X_i^{p+1}\text{sign}(X_i), \tag{A.2.23}$$

assuming only that h is such that positive kernel weights are assigned to at least $(p+1)$ data points on either side of the threshold. An infeasible bias-aware AR CS for our parameter of interest θ_v is then given by

$$\mathcal{C}_{vp}^{\alpha} = \{c : |\widehat{\tau}_{M,vp}(h_{M,vp}(c),c)| \leq \text{cv}_{1-\alpha}(r_{M,vp}(h_{M,vp}(c),c))s_{M,vp}(h_{M,vp}(c),c)\},$$

where $h_{M,vp}(c) = \arg\min_h \text{cv}_{1-\alpha}(r_{M,vp}(h,c))s_{M,vp}(h,c)$ is again the efficiency-maximizing bandwidth and $r_{M,vp}(h,c) = \overline{b}_{M,vp}(h,c)/s_{M,vp}(h,c)$ the "worst case" bias to standard deviation ratio. We can then establish the following result.

Theorem A.2.5. *Suppose that Assumptions 2.1 and either LL1 or LL2 hold. Then $\mathcal{C}_{vp}^{\alpha}$ is honest with respect to $\mathcal{F}$ in the sense of* (A.2.22)*.*

It is also straightforward to obtain an analogous result for a feasible version of $\mathcal{C}_{vp}^{\alpha}$ that uses a valid standard error and an estimate of the optimal bandwidth, under appropriate regularity conditions.

2.D.2. **Proof of Theorem A.2.5.** The result follows from the same type of arguments as those used in the proof of Theorem 1 for the FRD case. The only step that requires particular attention is establishing the validity of the general bias bound in (A.2.23), as Armstrong and Kolesár (2020, Theorem B.3) give an explicit expression for the special case $p = 1$ and $v = 0$ only. We first prove a preliminary result. Let $\chi = \{x_0, x_1, \ldots, x_k\}$, with $0 \leq x_0 \leq x_1 \leq \ldots \leq x_k < h$ and $k \geq p$, be a generic set of at least $p+1$ constants from the interval $[0,h)$, write $\chi_{-i} = \chi \setminus \{x_i\}$ for the subset of χ that excludes its ith element, and define

$$\widehat{\beta}_{vp}(t,\chi) = \sum_{i=0}^{k} w_{vp,i,+}(h,\chi)\mathbf{1}\{x_i \geq t\}(x_i - t)^p,$$

where $w_{vp,i,+}(h,\chi)$ are local polynomial regression weights analogous to those defined above, but with χ taking the role of the data $\mathcal{X}_n$. Put differently, the term $\widehat{\beta}_{vp}(t,\chi)$ is the $(v+1)$th coefficient in a weighted least squares regression of $\mathbf{1}\{x_i \geq t\}(x_i - t)^p$ on $(1, x_i, x_i^2, \ldots, x_i^p)^\top$. This term is well-defined as long as χ contains at least $p+1$ distinct elements. We first establish the following preliminary result.

Lemma A.2.2. *Suppose that either (i) χ has $(p+1)$ elements, all which are distinct; or (ii) χ has at least $(p+2)$ distinct elements, and $\widehat{\beta}_{vp}(t, \chi_{-i})$ satisfies* (A.2.24) *for all $i = 1, \ldots, |\chi|$. Then it holds for all $t \in \mathbb{R}$ that*

$$\widehat{\beta}_{vp}(t, \chi) \leq 0 \quad \text{if } p - v \text{ odd and } \widehat{\beta}_{vp}(t, \chi) \geq 0 \quad \text{if } p - v \text{ even.} \tag{A.2.24}$$

To then establish the bias bound (A.2.23), note that the bias can be written as

$$\begin{aligned} b_{M,vp}(h, c) = &\left(\sum_{i:X_i \geq 0} w_{vp,i,+}(h) \mu_M(X_i, c) - \mu_{M+}^{(v)}(c) \right) \\ &- \left(\sum_{i:X_i < 0} w_{vp,i,-}(h) \mu_M(X_i, c) - \mu_{M-}^{(v)}(c) \right) \equiv T_1 + T_2. \end{aligned}$$

Since $\sum_{i:X_i \geq 0} w_{vp,i,+}(h) X_i^v = 1$ and $\sum_{i:X_i \geq 0} w_{vp,i,+}(h) X_i^j = 0$ for $j \neq v$ and $j \leq p$ by standard least squares algebra, it follows that

$$\begin{aligned} T_1 &= \sum_{i:X_i \geq 0} w_{vp,i,+}(h) \left(\sum_{j=0}^{p} \frac{1}{j!} X_i^j \mu_M^{(j)}(0, c) + \frac{1}{p!} \int_0^{X_i} \mu^{(p+1)}(X_i, c)(X_i - t)^j dt \right) - \mu_{M+}^{(v)}(c) \\ &= \frac{1}{p!} \sum_{i:X_i \geq 0} w_{vp,i,+}(h) \int_0^{X_i} \mu_M^{(p+1)}(t, c)(X_i - t)^p dt \\ &= \frac{1}{p!} \int_0^{\infty} \mu_M^{(p+1)}(t, c) \sum_{i:X_i \geq 0} w_{vp,i,+}(h) \mathbf{1}\{X_i \geq t\}(X_i - t)^p dt \\ &\equiv \frac{1}{p!} \int_0^{\infty} \mu_M^{(p+1)}(t, c) \widehat{\beta}_{vp}(t, \mathcal{X}_n^+), \end{aligned}$$

where $\mathcal{X}_n^+ = \{X_i \in \mathcal{X}_n : 0 \leq X_i \leq h\}$. This expression is clearly maximized in absolute value by any function $\mu_M(t, c)$ whose $(p+1)$th derivative is given by $\mu_M^{(p+1)}(t, c) = B_M \text{sign}(\widehat{\beta}_{vp}(t, \mathcal{X}_n^+))$ for $t \geq 0$.

We now construct a collection $\mathcal{X}_{n,k}^+$ of subsets of $\mathcal{X}_n^+$, with $k = p+1, \ldots, n$, as follows. Let $\mathcal{X}_{n,p+1}^+$ be an arbitrary subset of $p+1$ distinct elements of $\mathcal{X}_n^+$ (such a subset exists by assumption), and let $\mathcal{X}_{n,k}^+$, for $k > p+1$, be the union of $\mathcal{X}_{n,k-1}^+$ and an arbitrary element of $\mathcal{X}_n^+ \setminus \mathcal{X}_{n,k-1}^+$. Then Lemma A.2.2 implies that $\widehat{\beta}_{vp}(t, \mathcal{X}_{n,k}^+)$ satisfies (A.2.24) for any $k = p+1, \ldots, n$. Since $\mathcal{X}_{n,n}^+ = \mathcal{X}_n^+$, this means that $\text{sign}(\widehat{\beta}_{vp}(t, \mathcal{X}_n^+)) = (-1)^{p-v}$ for all t. The term T_1 is thus maximized in absolute value for any function μ_M such that $\mu_M(t, c) = (-1)^{p-v} B_M t^{p+1} \text{sign}(t)/((p+1)!)$ for $t \geq 0$. A similar reasoning implies that T_2 is maximized for any function μ_M such that $\mu_M(t, c) = (-1)^{p-v} B_M t^{p+1} \text{sign}(t)/((p+1)!)$ for $t < 0$. Together, these statements prove (A.2.23). □

2.D.3. **Proof of Lemma A.2.2.** To prove part (i), note that there is always a unique polynomial of order p that interpolates the points $\{(x, \mathbf{1}\{x \geq t\}(x-t)^p)\}_{x\in\chi}$. We denote this polynomial as a function of x by $P(x, \chi_k)$..Our proof comes down to determining the sign of the corresponding coefficients as a function of t. To do so, let $S(k) = k + |\{x \in \chi : x \leq t\}|$ be the sum of k and the number of elements of χ whose value does not exceed t, and consider subsets of χ of the form $\chi_k = \{x_i \in \chi : x_i \leq t\} \cup \{x_i \in \chi : S(1) \leq i \leq S(k)\}$ that contain those elements of χ whose value does not exceed t, and the k next largest ones. That is, $\chi_0 = \{x_i \in \chi : x_i \leq t\}$, and χ_1 is the union of χ_0 and the smallest element of χ that is larger than t, etc. We also note that if χ is such that $S(0) = 0$, then $\widehat{\beta}_{vp}(t, \chi) = (-1)^{p-v}\binom{p}{v}t^v$ clearly satisfies (A.2.24). It therefore suffices to restrict attention to sets χ such that $S(0) > 0$. It is also easy to see that $\widehat{\beta}_{vS(0)}(t, \chi_0) = 0$, and hence satisfies (A.2.24). It thus remains to show that if $\widehat{\beta}_{vS(k)}(t, \chi_k)$ satisfies (A.2.24), so does $\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1})$. The statement of the lemma then follows by induction.

To show the last step, assume that $\widehat{\beta}_{vS(k)}(t, \chi_k)$ satisfies (A.2.24), and write the polynomial that interpolates the points $\{x, \mathbf{1}\{x \geq t\}(x-t)^{S(k+1)}\}_{x\in\chi_{k+1}}$ as

$$P(x, \chi_{k+1})x^v = (x-t)P(x, \chi_k) + \bar{\iota}_{p+1}\prod_{x_l\in\chi_k}(x - x_l), \text{ where} \tag{A.2.25}$$
$$\bar{\iota}_{k+1} = (x_{S(k+1)} - t)\left((x_{S(k+1)} - t)^{S(k)} - P(x, \chi_k)\right)\prod_{x_l\in\chi_k}\frac{1}{x_{S(k+1)} - x_l}.$$

We can then express the $\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1})$ in terms of the $\widehat{\beta}_{vS(k)}(t, \chi_k)$ by comparing the appropriate terms on both sides of equation (A.2.25). This yields that

$$\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1}) = \begin{cases} \widehat{\beta}_{S(k)S(k)}(t, \chi_k) + \bar{\iota}_{k+1} & \text{if } v = S(k+1), \\ -t\widehat{\beta}_{0S(k)}(t, \chi_k) + (-1)^{S(k+1)}\bar{\iota}_{k+1}\prod_{0\leq j\leq S(k)} x_j & \text{if } v = 0, \\ \widehat{\beta}_{(v-1)S(k)}(t, \chi_k) - t\widehat{\beta}_{vS(k)}(t, \chi_k) + (-1)^{S(k+1)-v}\bar{\iota}_{k+1}\sum_{M\in\mathcal{M}_{S(k+1)-v}}\prod_{m_s\in M} x_{m_s} & \text{else.} \end{cases}$$

where $\mathcal{M}_v$ is the set of all subsets $M = \{m_1, \ldots, m_v\}$ of $\{1, \ldots, S(k+1)\}$ that contain exactly v elements. Careful inspection of the last display shows that $\widehat{\beta}_{vS(k+1)}(t, \chi_{k+1})$ satisfies (A.2.24) if $\bar{\iota}_{k+1} \geq 0$. We proof this claim by a simple argument about the number of zeros of polynomials. Let $\chi_{k\setminus 0} = \chi_k \setminus x_0$ that is the set χ_k without its smallest element. We note that $\bar{\iota}_{k+1} \geq 0$ if

$$P(x, \chi_k) < P(x, \chi_{k\setminus 0} \cup x) \text{ for all } x > x_{S(k)}. \tag{A.2.26}$$

To show (A.2.26), we fix some arbitrary $x_l > x_{S(k)}$ and consider the two different polynomials $P(x, \chi_k)$ and $P(x, \chi_{k\backslash 0} \cup x_l)$. These polynomials are of degree $S(k)$ and they intersect $S(k)$ times at all $x \in \chi_k \setminus x_0$, so that they cannot intersect for any $x \notin \chi_k \setminus x_0$.

As the set χ_k was arbitrarily chosen, we note that by the induction argument the intercept of both polynomials has the same sign such that

$$\text{sign}(P(0, \chi_k)) = \text{sign}(P(0, \chi_{k\backslash 0} \cup x_l)). \tag{A.2.27}$$

Using (A.2.27) together with standard arguments of polynomials and their sign as $x \to \pm\infty$, (A.2.26) is satisfied if $|P(0, \chi_k)| \leq |P(0, \chi_{k\backslash 0} \cup x_l)|$. Polynomials of order $S(k)$, that are different from $(x-t)^{S(k)}$, can have at most $(S(k)+1)$ intersections with the function $g(x) = \mathbf{1}\{x \geq t\}(x-t)^{S(k)}$ for $t > 0$. This reasoning implies that the polynomial $P(x, \chi_{k\backslash 0} \cup x_l)$ does not have any intersections with the function $g(x)$ for $x \leq x_0$, and in particular it does not have any root for $x \leq x_0$, so that it has the same sign for all $0 \leq x \leq x_0$. As $P(x_0, \chi_k) = 0$, we can conclude that $|P(x, \chi_k)| \leq |P(x, \chi_{k\backslash 0} \cup x_l)|$ for any $x \leq x_0$. This completes our proof of part (i).

To prove part (ii) of the lemma, note that it follows from textbook arguments that

$$\widehat{\beta}_{vp}(t, \chi) = \widehat{\beta}_{vp}(t, \chi_{-i}) + (1 - l_i)^{-1} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i,$$

where $\widehat{\epsilon}_i = \mathbf{1}\{x_i \geq t\}(x_i - t)^p - \sum_{v=0}^{p} \widehat{\beta}_{vp}(t, \chi) x_i^v$ is the ith regression residual and $l_i = \sum_{j=0}^{p} w_{jp,i}(\chi) x_i^j$ is the leverage of the ith observation. We now first consider the case that $\widehat{\beta}_{vp}(t, \chi_{-i}) \leq 0$ for all i, which implies that $\widehat{\beta}_{vp}(t, \chi) \leq (1 - l_i)^{-1} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i$. Since $\sum_{i=1}^{|\chi|} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i = 0$ and $0 \leq l_i < 1$ for all i by basic least squares algebra, we know that $(1 - l_i)^{-1} w_{vp,i,+}(h, \chi)\widehat{\epsilon}_i \leq 0$, for at least some i, which in turn means that $\widehat{\beta}_{vp}(t, \chi) \leq 0$. The same kind of argument applies to the case that $\widehat{\beta}_{vp}(t, \chi_{-i}) \geq 0$ for all i. □

2.E. ADDITIONAL MATERIALS FOR THE EMPIRICAL APPLICATION

In this appendix, we provide additional materials for the empirical application. Figure A.2.2 shows the fit of the polynomial regressions on which the two ROT values are based. The top four panel show the result for for the full data. Both the second and forth order polynomial specification provide a reasonable fit when log wage is the dependent variable, whereas both fits seem inadequate for the conditional treatment probabilities. The bottom four panels of Figure A.2.2 show the fits for the data excluding the 1947 cohort. Here both polynomials seem to provide good fit for outcomes and treatment probabilities.

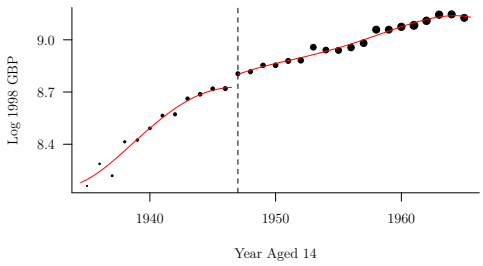
To further illustrate the order of magnitude of the implied smoothness bounds on the curvature of μ_Y, we plot examples of functions lying in the respective smoothness class

under consideration in Figure A.2.3. The resulting functions look very similar because of the scaling of the vertical axis, and hence we plot them again with a different scaling in Figure A.2.4. Figure A.2.5 shows analogous graphs with the data excluding the 1947 cohort. We perform the same exercise for the fraction of people staying in school beyond age of 14 in Figure A.2.6 with the full sample and in Figure A.2.7 with the data excluding the 1947 cohort. We want to emphasize again that the functions plotted in Figures A.2.3–A.2.7 are not meant to be estimates of the respective underlying conditional expectation functions. They are examples of elements of the respective smoothness classes, and plotted to help applied researchers understand the implications of choosing a particular smoothness bound.

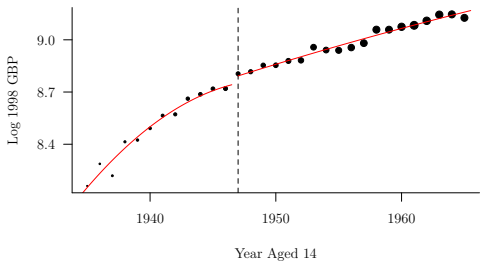
2.F. ADDITIONAL MATERIALS FOR THE SIMULATIONS

In this section, we report results from a variation of our main simulation study, in which the procedures using an estimate of the IK bandwidth were implemented with an estimate of "coverage error optimal" bandwidth proposed by Calonico et al. (2018). The latter was computed using the `R` package `rdrobust`. Table A.2.1 shows the coverage rates obtained in the main simulation (IK) for reference, and the ones newly obtained here. The values are overall very similar, suggesting that the results regarding robust bias correction in the main body of the paper are not driven by the details of the algorithm for bandwidth choice.

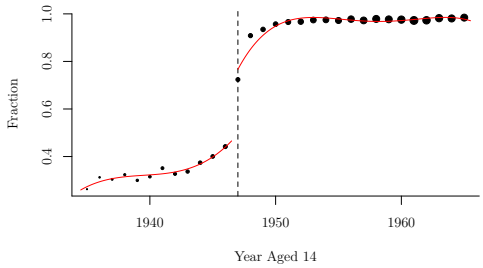
(a) ROT1, Wages, Full Data



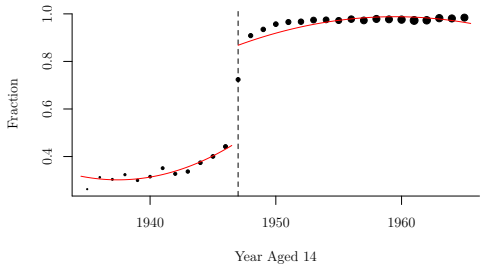
(b) ROT2, Wages, Full Data



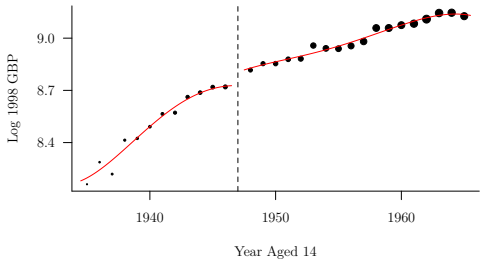
(c) ROT1, Treatment, Full Data



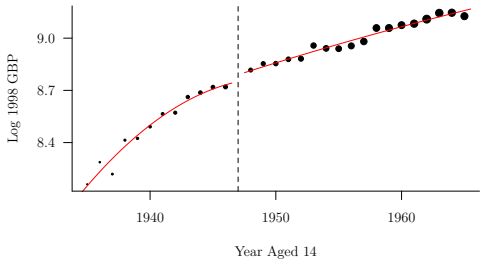
(d) ROT2, Treatment, Full Data



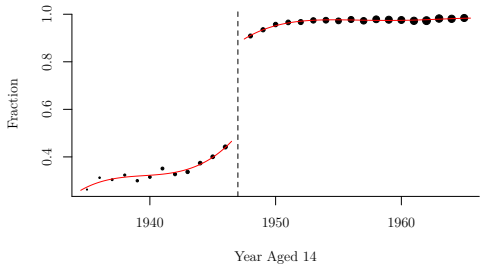
(e) ROT1, Wages, Excluding 1947 data



(f) ROT2, Wages, Excluding 1947 data



(g) ROT1, Treatment, Excluding 1947 data



(h) ROT2, Treatment, Excluding 1947 data

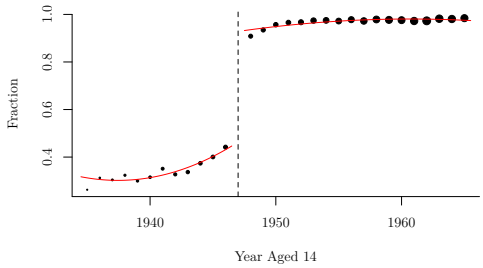
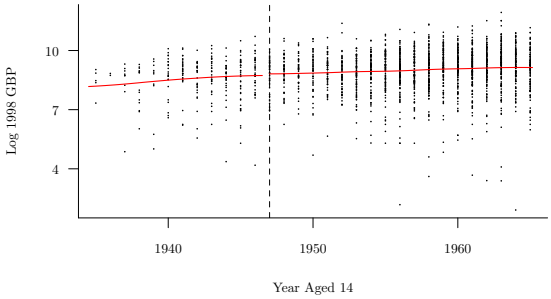


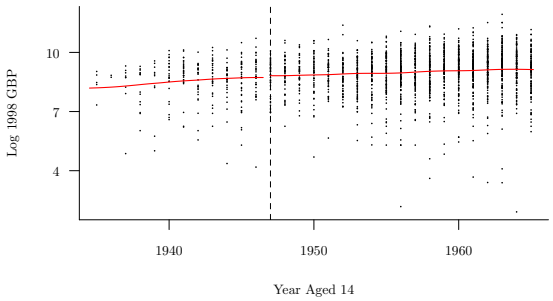
Figure A.2.2: Fits of polynomial specifications underlying ROT1 and ROT2.

(a) $B_Y = 0$

(c) $B_Y = .01$



(d) $B_Y = .02$



(a) $B_Y = .03$

(c)$B_Y = .04$

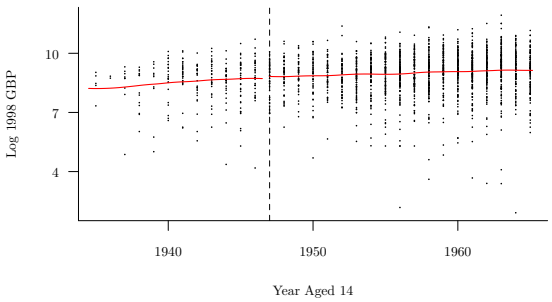
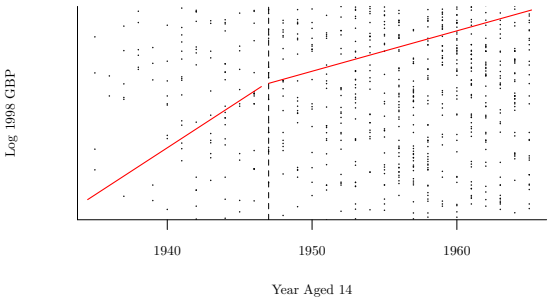
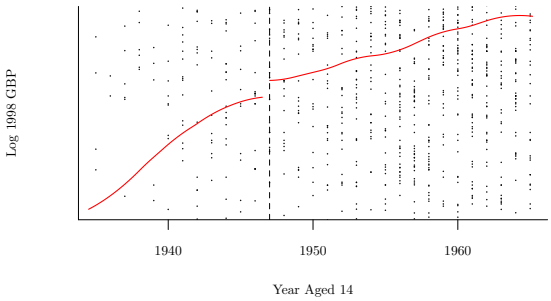


Figure A.2.3: Average Log Annual Earnings: Examples of elements of $\mathcal{F}_H(B_Y)$ for various values of B_Y based on the full data set. Figure also shows 2,000 random data points.

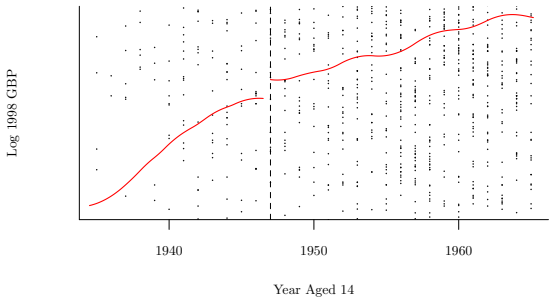
(a) $B_Y = 0$



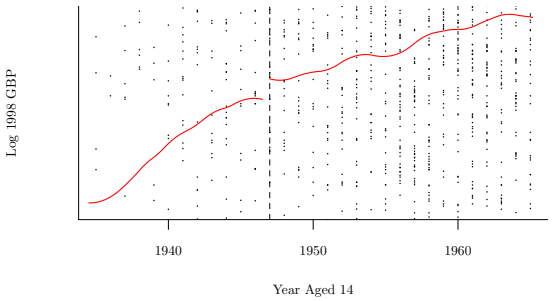
(c) $B_Y = .01$



(d) $B_Y = .02$



(a) $B_Y = .03$



(c)$B_Y = .04$

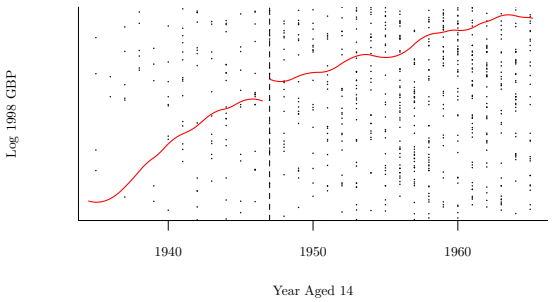
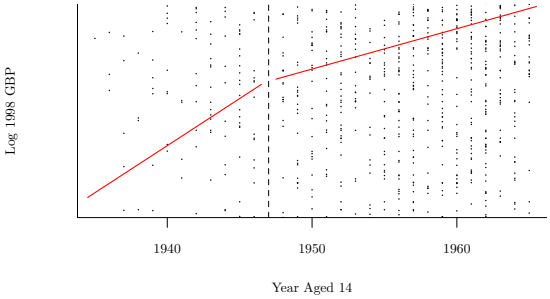
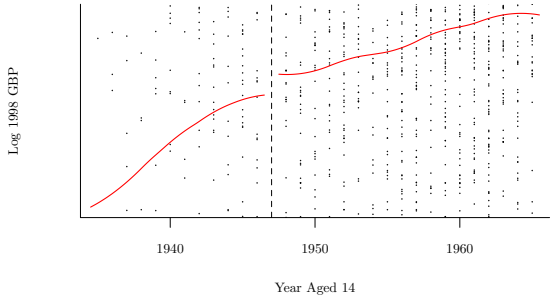


Figure A.2.4: Average Log Annual Earnings: Examples of elements of $\mathcal{F}_H(B_Y)$ for various values of B_Y based on the full data set. Figure also shows 2,000 random data points. Note that the functions in red are identical to those in Figure A.2.3. The scale of the vertical axis has been changed to better show their shape.

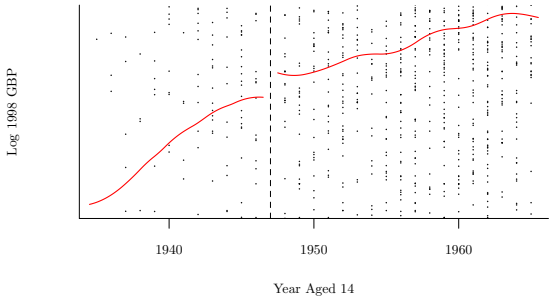
(a) $B_Y = 0$



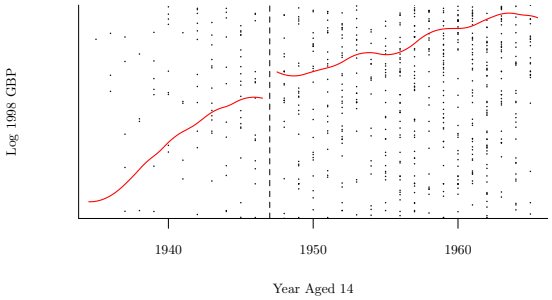
(c) $B_Y = .01$



(d) $B_Y = .02$



(a) $B_Y = .03$



(c)$B_Y = .04$

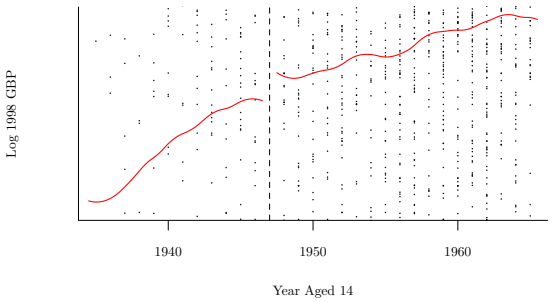
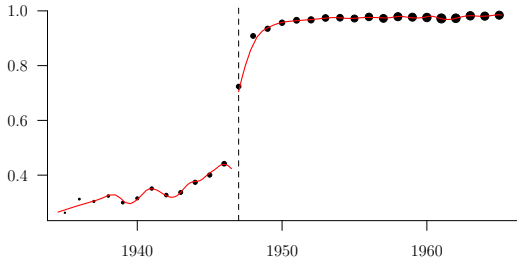
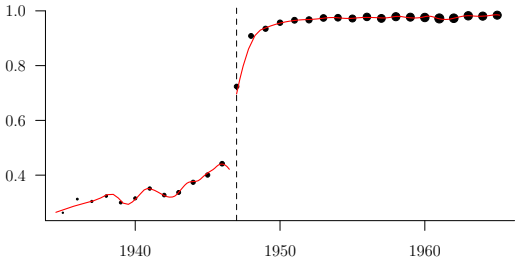


Figure A.2.5: Average Log Annual Earnings: Examples of elements of $\mathcal{F}_H(B_Y)$ for various values of B_Y based on data excluding the 1947 cohort. Figure also shows 2,000 random data points. The scale of the vertical axis is restricted to better show the shape of the candidate functions.

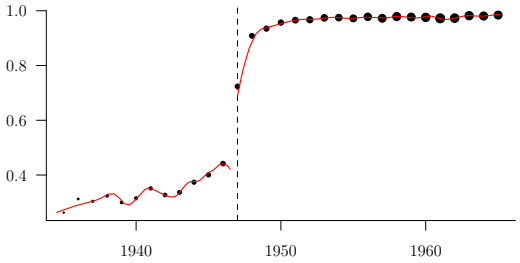
(a) $B_T = .12$



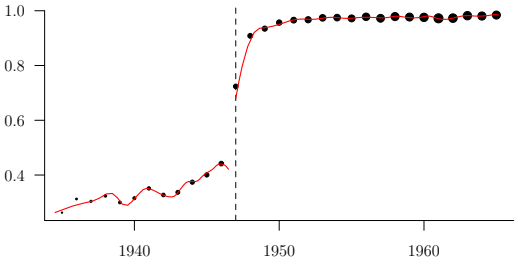
(c) $B_T = .14$



(d) $B_T = .16$



(a) $B_T = .18$



(c)$B_T = .2$

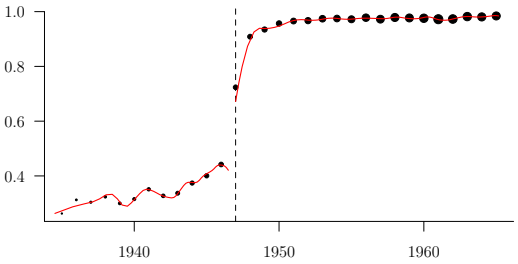
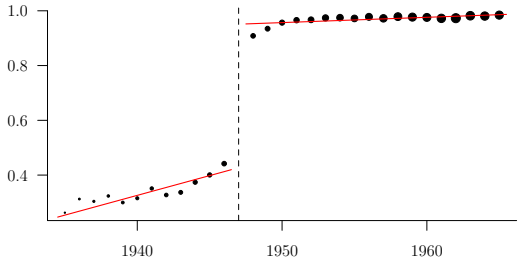
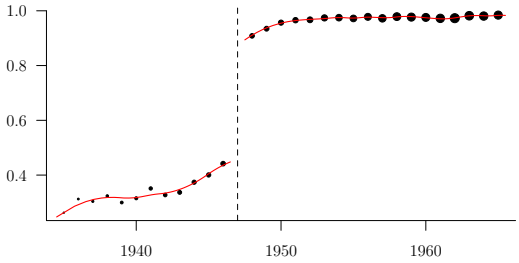


Figure A.2.6: School Attendance Beyond Age 14: Examples of elements of $\mathcal{F}_H(B_T)$ for various values of B_T based on the full data.

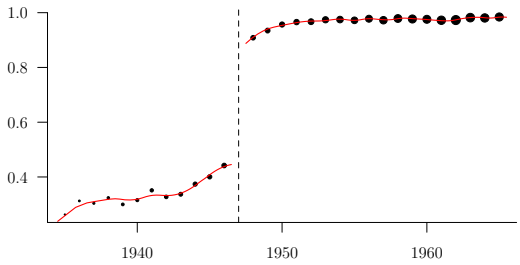
(a) $B_T = 0$



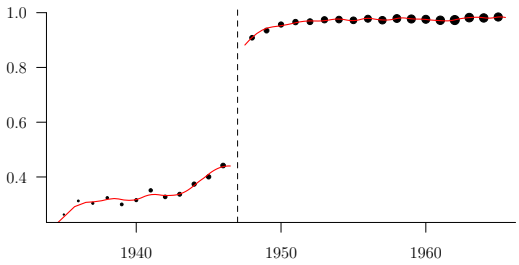
(c) $B_T = .01$



(d) $B_T = .02$



(c) $B_T = .01$



(d) $B_T = .02$

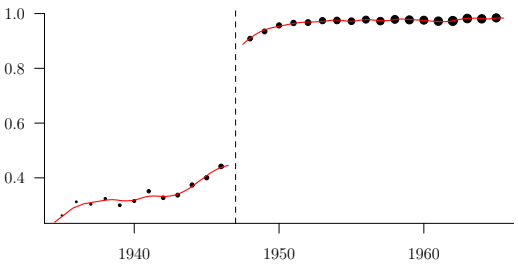


Figure A.2.7: Examples of elements of $\mathcal{F}_H(B_T)$ for various values of B_T based on data excluding the 1947 cohort.

Table A.2.1: Simulated coverage rate (in %) of true parameter for various types of confidence sets

			Anderson-Rubin						Delta Method					
			IK			CEO			IK			CEO		
τ_T	B_Y	B_T	Naive	US	RBC	Naive	US	RBC	Naive	US	RBC	Naive	US	RBC
Running Variable with Continuous Distribution														
0.5	1	0.2	93.1	93.4	93.4	93.4	93.1	93.5	90.8	90.4	91.3	90.4	89.4	90.7
0.5	1	1.0	93.0	93.3	93.3	93.3	93.3	93.5	90.4	89.9	91.0	89.9	89.0	90.2
0.5	10	0.2	92.4	93.0	92.5	93.0	92.9	93.0	88.3	88.3	88.7	88.3	87.8	88.4
0.5	10	1.0	92.2	92.8	92.2	92.8	92.7	92.8	87.9	88.2	88.4	88.2	87.4	88.3
0.5	100	0.2	78.5	87.9	74.7	87.9	90.7	86.7	72.8	80.8	72.2	80.8	82.9	80.5
0.5	100	1.0	78.2	88.0	74.4	88.0	90.7	86.8	72.6	80.9	72.1	80.9	83.2	80.6
0.1	1	0.2	93.7	94.0	94.0	94.0	94.0	94.2	76.6	74.0	79.2	74.0	70.8	75.5
0.1	1	1.0	93.4	93.8	93.8	93.8	93.9	94.0	76.5	73.8	79.0	73.8	71.0	75.2
0.1	10	0.2	93.4	94.0	93.7	94.0	94.0	94.1	71.0	69.6	73.3	69.6	66.7	71.0
0.1	10	1.0	93.2	93.9	93.5	93.9	94.0	94.0	70.3	69.2	72.7	69.2	66.5	70.5
0.1	100	0.2	83.2	91.0	79.0	91.0	93.2	89.9	36.7	47.4	37.1	47.4	53.5	47.6
0.1	100	1.0	83.2	91.0	79.1	91.0	93.2	90.0	36.5	47.4	37.0	47.4	53.2	47.7
Running Variable with Discrete Distribution														
0.5	1	0.2	94.3	94.6	94.6	94.6	93.8	95.0	89.9	88.9	91.0	88.9	88.1	89.5
0.5	1	1.0	94.0	94.2	94.4	94.2	93.6	94.7	89.6	88.5	90.5	88.5	87.6	89.0
0.5	10	0.2	93.6	93.9	93.7	93.9	93.1	94.2	85.3	84.8	85.4	84.8	86.1	85.7
0.5	10	1.0	93.6	93.7	93.6	93.7	93.1	94.1	84.6	84.4	84.6	84.4	85.5	85.3
0.5	100	0.2	67.9	60.4	57.8	60.4	60.4	65.1	26.8	27.9	17.1	27.9	28.5	22.3
0.5	100	1.0	67.2	59.5	57.2	59.5	59.5	64.6	26.5	27.6	16.7	27.6	28.1	22.0
0.1	1	0.2	94.7	94.9	95.1	94.9	94.3	95.2	71.2	64.8	75.6	64.8	62.6	67.4
0.1	1	1.0	94.5	94.7	94.7	94.7	94.2	95.0	70.5	64.5	74.8	64.5	62.0	67.0
0.1	10	0.2	94.5	94.5	94.6	94.5	94.1	94.9	55.9	52.6	59.9	52.6	59.1	57.0
0.1	10	1.0	94.5	94.5	94.6	94.5	94.1	94.8	55.8	51.8	59.6	51.8	58.6	56.2
0.1	100	0.2	73.7	66.9	63.9	66.9	66.9	69.4	68.6	69.5	65.3	69.5	65.8	69.4
0.1	100	1.0	73.2	66.4	63.0	66.4	66.3	68.7	68.1	69.0	64.8	69.0	65.2	68.8

Notes: Results based on 50,000 Monte Carlo draws for a nominal confidence level of 95%. Columns show simulated coverage rates of the true constants for confidence sets based on AR and DM. The bandwidth is chosen minimizing the MSE bandwidth and the CEO, see `rdrobust` for details. We consider confidence sets based on an approach ignoring the bias (Naive); undersmoothing (US); and robust bias correction (RBC). See main paper for details of the simulation design.

Chapter 3

Flexible Covariates Adjustments in Regression Discontinuity Designs

with Tomasz Olma and Christoph Rothe

3.1. INTRODUCTION

Regression discontinuity (RD) designs are widely used for estimating causal treatment effects from observational data in economics and other social sciences. In a sharp RD design, the treatment status is determined by whether the running variable exceeds a fixed cutoff value. Under standard assumptions, the average treatment effect at the cutoff is identified by the size of the jump in the conditional expectation of the outcome variable given the running variable at the cutoff. This parameter is typically estimated using local linear regression methods, and various inference procedures have been proposed in the literature; see, e.g., Imbens and Kalyanaraman (2012), Calonico et al. (2014), and Armstrong and Kolesár (2020).

The standard estimator of the average treatment effect in sharp RD designs is based solely on the outcome variable and the running variable, but in many empirical applications, researchers include additional, pretreatment covariates linearly in the RD regression to reduce the variance of the estimates (see Calonico et al., 2019). However, linear adjustments in general do not fully exploit the information contained in the covariates. The goal of this chapter is to improve upon these methods.

We propose a novel class of covariate-adjusted RD estimators. They are constructed in two stages. In the first stage, we obtain adjustment terms, which aim at capturing the variation in the outcome variable near the cutoff that can be explained by the additional covariates. The adjustment terms are estimated using cross-fitting, which allows us to use a wide range of methods in the first stage under weak conditions. We generate a covariate-adjusted outcome variable by subtracting the adjustment terms from the original outcomes. In the second stage, we estimate the RD parameter in a local linear regression with our generated outcome variable.

Our proposed approach is based on the premise that in a valid RD design, the conditional distribution of the additional covariates given the running variable should evolve

continuously through the cutoff. Such a condition is inherently related to the standard, behavioral identification arguments in RD designs, which postulate that the units just to the left and just to the right of the cutoff are very similar in all pretreatment characteristics.[1] Based on this feature, we can adjust our outcome variable by subtracting from it essentially any function of the additional covariates without changing the RD estimand. We can further choose the adjustment function that leads to the smallest variance of the RD estimator in the considered class of estimators. We find that the optimal adjustment function is given by the average of the conditional expectations of the outcome variable just to the left and just to the right of the cutoff given the additional covariates. This function is not known, and therefore we estimate it in the first stage.

An important feature of our proposed RD estimator is that it is very insensitive to the first-stage estimation error, which has the following important, practical and theoretical implications. First, we only require that the first-stage estimator concentrates, possibly very slowly, in a mean-squared-error-type sense around some deterministic sequence of functions. This condition is satisfied for a wide range of estimators, including parametric estimators, classic nonparametric methods, such as local linear and sieve estimators (Fan and Gijbels, 1996; Newey, 1997), as well as modern machine learning methods, such as lasso (Tibshirani, 1996), random forests (Breiman, 2001; Wager and Athey, 2018), and deep neural networks (Farrell et al., 2021). Importantly, our RD estimator is not very sensitive to the specific choice of the tuning parameters that are required for some of the above methods.

Second, in our asymptotic analysis, we can ignore the fact that the adjustment terms are estimated in the first stage. Our proposed RD estimator is asymptotically equivalent to an estimator employing the deterministic function around which the first-stage estimator concentrates. As a result, existing procedures for inference and bandwidth choice can be directly applied to the second-stage regression. Specifically, we obtain the standard error using the nearest-neighbors method. We also argue that one can choose the bandwidth and construct confidence intervals following the robust bias corrections approach of Calonico et al. (2014) or the bias-aware procedure of Armstrong and Kolesár (2020).

We further show that if the first-stage estimator consistently estimates the targeted conditional expectations, then our estimator is efficient in the considered class, but our asymptotic results remain valid whether or not this condition is satisfied. Our proposed covariate adjustments asymptotically lead to variance reductions compared to the standard RD estimator whenever the covariates have explanatory power for the outcome variable in a neighborhood of the cutoff.

[1]Indeed, in empirical applications, testing continuity of the distribution of baseline covariates at the cutoff has become a standard way of assessing the validity of an RD design (Cattaneo et al., 2019).

Our proposed procedure is related to covariate adjustments used in randomized experiments to improve efficiency of the average treatment effect estimator (see, e.g., Wager et al., 2016). This analogy occurs because RD designs are similar in nature to randomized experiments. In randomized experiments, comparability of the treated and untreated units is ensured by random assignment, whereas in RD designs, it is ensured for units close to the cutoff by continuity of potential outcomes' distributions. Our proposed RD estimator has a very similar structure as the augmented inverse probability weighted estimator, which is widely used in randomized experiments. Accordingly, the minimal variance that our estimator can achieve resembles the efficiency bound for estimation of the average treatment effect under unconfoundedness (Hahn, 1998).

Literature. There exists an extensive literature on estimation in RD designs; see, e.g., Imbens and Lemieux (2008) and Cattaneo et al. (2019) for a textbook treatment. In general, existing methods do not require covariate information, but it is standard practice to incorporate covariates in order to reduce the variance of the estimates (see, e.g., Lee and Lemieux, 2010, Section 3.2.3). We contrast our approach with two papers that are most closely related to our approach.

Calonico et al. (2019) employ a local linear regression in the running variable with additional covariates included in a linear fashion. We allow for linear adjustments as a special case, but we cover a wide range of other, more flexible adjustments that improve efficiency compared to simple linear adjustments. We discuss the relation of our approach to that of Calonico et al. (2019) in more detail in Section 3.6.1.

Frölich and Huber (2019) propose a procedure using first-stage nonparametric predictions of the treatment effect conditional on the additional covariates at the cutoff, which achieves approximately the same variance as our estimator in some settings. However, their results rely on strong assumptions about the number of covariates and/or smoothness of the conditional expectation of the outcome variable given the covariates, which are not needed for our method.[2]

Our chapter is also related to the literature on two-stage estimation with nuisance parameters (Andrews, 1994; Newey, 1994). The combination of locally-robust moment conditions and cross-fitting has been used, e.g., by Belloni et al. (2017); Chernozhukov et al. (2018). Estimation of conditional treatment effects with orthogonal moments have been studied, e.g., by Kennedy et al. (2017); Kennedy (2020); Fan et al. (2020).

[2]For example, Frölich and Huber (2019) allow for at most three continuous additional covariates and require that the bandwidth converges at a specific rate if the local linear estimator with a second-order kernel is used in the first stage.

Plan of the Chapter. The remainder of this chapter is organized as follows. In Section 3.2, we introduce the setup. In Section 3.3, we present our proposed covariate-adjusted estimator. In Section 3.4, we present our main theoretical results under general conditions on the covariate adjustments used. We discuss implementation details in Section 3.5. In Section 3.6, we consider specific examples of covariate adjustments. We present a simulation study in Section 3.7. Section 3.8 concludes.

Notation. Throughout the chapter, we use the following notation. For a generic function $f(x)$, we write $f(0^+) = \lim_{x \downarrow 0} f(x)$ and $f(0^-) = \lim_{x \uparrow 0} f(x)$ for the right and left limit of the function $f(x)$ at zero, respectively.

3.2. SETUP

In this section, we introduce the model and parameter of interest. Furthermore, we discuss estimation of the RD parameter based on local linear regression methods.

3.2.1. **Model and Parameter of Interest.** We consider a sharp RD design, in which the researcher investigates the causal effect of a binary treatment on some outcome variable of interest. The data $(W_i)_{i \in \{1,\dots,n\}}$ are an i.i.d. sample of size n from the distribution of $W_i = (Y_i, X_i, Z_i)$. Here, $Y_i \in \mathbb{R}$ is the outcome variable, $X_i \in \mathbb{R}$ is the running variable, and $Z_i \in \mathbb{R}^d$ is a vector of additional covariates. Units receive the treatment if and only if the running variable exceeds some known threshold, which we normalize to zero without loss of generality. We denote the treatment indicator by T_i, so that $T_i = \mathbf{1}\{X_i \geq 0\}$.

Throughout the chapter, we assume that the distribution of the running variable X_i is fixed, but we consider a sequence of conditional distributions of (Y_i, Z_i) given X_i that can change with n. In particular, we allow the dimension of Z_i to grow with n. For ease of notation, we leave the dependence on n implicit.

We denote the support of Z_i by $\mathcal{Z}$, and we let $\mathcal{X}$ be an open neighborhood of the cutoff that is contained in the support of the running variable. The density of the running variable is denoted by f_X, the conditional cumulative distribution function of Z_i given $X_i = x$ is denoted by $F_{Z|X}(z|x)$. If the corresponding conditional density exists, we denote it by $f_{Z|X}(z|x)$. Under standard assumptions (see, e.g., Lee and Lemieux, 2010) the average treatment effect at the cutoff is identified by the height of the jump in the conditional expectation of the observed outcome variable given the running variable at zero:

$$\tau = \mathbb{E}[Y_i | X_i = 0^+] - \mathbb{E}[Y_i | X_i = 0^-]. \tag{3.1}$$

We take this identification result as given and consider estimation of τ as defined above.

3.2.2. **Standard RD Estimator.** In RD designs, the parameter of interest is typically estimated using local linear regression (see, e.g., Fan and Gijbels, 1996). The standard estimator is given by:

$$\widehat{\tau}(h) = e_1^\top \arg\min_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K(X_i/h)(Y_i - \beta^\top (T_i, X_i, T_i X_i, 1))^2,$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$, $h > 0$ is a bandwidth, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector. Using simple algebra, this estimator can be expressed as a weighted sum of the outcome variable:

$$\widehat{\tau}(h) = \sum_{i=1}^{n} w_i(h) Y_i,$$

where the weights $w_i(h)$ depend only on the realizations of the running variable. We give the explicit expressions for the weights in Appendix 3.C.1.

Under standard assumptions, the estimator $\widehat{\tau}(h)$ is asymptotically normally distributed. Its leading bias term is proportional to $\partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^-}$, and it is of order h^2. The bias results from approximating the possibly non-linear conditional expectation function with a linear function. Its magnitude is determined by the degree of nonlinearity, measured by the value of the second derivative. The variance is of order $(nh)^{-1}$, and it is approximately proportional to $\mathbb{V}[Y_i|X_i = 0^+] + \mathbb{V}[Y_i|X_i = 0^-]$.

3.3. COVARIATE ADJUSTMENTS

In this section, we motivate our proposed estimation procedure, and we formally define the proposed covariate-adjusted RD estimator.

3.3.1. **Covariate-Adjusted Outcome Variable.** We now introduce the key object of this chapter. We consider a modified outcome variable of the following form:

$$M_i(\mu) = Y_i - \mu(Z_i), \tag{3.2}$$

where μ is a real-valued function of the additional covariates, which we refer to as the adjustment function.

For the further analysis, we impose a regularity condition on the admissible adjustment functions and require that $\mu(Z_i)$ is square integrable conditional on the running variable. We define the set of such functions as:

$$\mathcal{M}_n = \Big\{ \mu : \mathcal{Z} \to \mathbb{R} \text{ s.t. } \sup_{x \in \mathcal{X}} \mathbb{E}[\mu(Z_i)^2 | X_i = x] < \infty \Big\}.$$

The central premise for our proposed approach is that the conditional distribution

of the additional covariates given the running variable evolves continuously through the cutoff.

Assumption 3.1. *For all $n \in \mathbb{N}$ and $\mu \in \mathcal{M}_n$, $\mathbb{E}[\mu(Z_i)|X_i = x]$ is continuous in x on $\mathcal{X}$.*

Assumption 3.1 requires that the conditional distribution of Z_i given $X_i = x$ converges weakly to the distribution of Z_i given $X_i = 0$, as x converges to 0.[3] Under this assumption, we can replace the outcome variable Y_i in the definition of τ in (3.1) with $M_i(\mu)$ without affecting the value of the estimand, that is:

$$\tau = \mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-] \tag{3.3}$$

for all $\mu \in \mathcal{M}_n$.

Motivated by the above representation, for any fixed $\mu \in \mathcal{M}_n$, the RD parameter τ could be estimated using the local linear RD estimator with $M_i(\mu)$ as the outcome variable, which we denote by:

$$\widehat{\tau}(h; \mu) = \sum_{i=1}^{n} w_i(h) M_i(\mu). \tag{3.4}$$

In practice, the adjustment function might be estimated from the data. However, in a sense made precise in the next sections, we can replace the deterministic adjustment function with its estimate without affecting the first-order asymptotic properties of the final estimator of the RD parameter. We therefore first determine the adjustment function that minimizes the variance of the RD estimator $\widehat{\tau}(h; \mu)$.

3.3.2. **Optimal Adjustment Function.** The RD estimator $\widehat{\tau}(h; \mu)$ has variance that is approximately proportional to $\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]$. We find that the adjustment function that minimizes this expression is given by

$$\mu_n(z) = \frac{1}{2}\left(\mu_n^+(z) + \mu_n^-(z)\right), \tag{3.5}$$

where $\mu_n^+(z) = \mathbb{E}[Y_i|X_i = 0^+, Z_i = z]$ and $\mu_n^-(z) = \mathbb{E}[Y_i|X_i = 0^-, Z_i = z]$. This result follows from simple derivations, which we outline below to present the intuition behind this result.

Under Assumption 3.1, if μ_n^-, μ_n^+, $\mu \in \mathcal{M}_n$, then

$$\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-] = \mathbb{V}[M_i(\mu_n^+)|X_i = 0^+] + \mathbb{V}[M_i(\mu_n^-)|X_i = 0^-] + \mathcal{V}(\mu),$$

[3] This condition holds if $F_{Z|X}(z|x) \to F_{Z|X}(z|0)$, as $x \to 0$, for all continuity points of $F_{Z|X}(z|0)$.

where the first two terms on the right-hand side do not depend on μ, and

$$\mathcal{V}(\mu) = \mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0] + \mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0].$$

Our goal is therefore to minimize $\mathcal{V}(\mu)$. Each component of $\mathcal{V}(\mu)$ could be set to zero separately if μ was chosen as μ_n^+ or μ_n^-, respectively. It turns out that the whole expression $\mathcal{V}(\mu)$ is minimized by the function μ_n, which can be seen by noting that:

$$\mathcal{V}(\mu) = \mathcal{V}(\mu_n) + 2\mathbb{V}[\mu_n(Z_i) - \mu(Z_i)|X_i = 0] \geq \mathcal{V}(\mu_n).$$

This reasoning shows that indeed the expression $\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]$ achieves the smallest value if $\mu = \mu_n$. The function μ_n is essentially a unique minimizer up to shifts by a constant.

3.3.3. **Estimator.** We estimate τ in a two-stage procedure. In the first stage, we estimate the function μ_n defined in (3.5), which involves estimating the limits of the conditional expectation of the outcome variable given the additional covariates as the running variable approaches the cutoff from the left and from the right.

Conditional expectations at boundary points are often estimated using local linear methods because of their good bias properties. However, for our purposes, essentially any procedure can be adapted to estimate these limits by restricting the data to observations with the running variable close to the cutoff.[4] For example, we can use parametric estimators, classic nonparametric methods such as series and spline estimators (Masry, 1996; Newey, 1997), as well as modern machine learning methods including the lasso (Tibshirani, 1996), random forests (Breiman, 2001; Wager and Athey, 2018), and deep neural networks (Farrell et al., 2021).

In order to allow for a variety of, possibly highly complex, first-stage estimators, we use cross-fitting (see, e.g., Chernozhukov et al., 2018).[5] We split the data randomly into S disjoint folds denoted I_s for $s \in [S] = \{1, ..., S\}$, where all folds have the same number of observations to the left of the cutoff, and similarly to the right of the cutoff.[6] For $s \in [S]$, we define the complement of fold I_s as $I_s^c = [n] \setminus I_s$. Further, let $s(i)$ denote the index of the fold containing observation i, so that $i \in s(i)$. Given a selected estimation procedure, we define $\widehat{\mu}_{n,s}(z) = \widehat{\mu}_n(z; (W_i)_{i \in I_s^c})$, which is an estimator of $\mu_n(z)$ that uses

[4] In our asymptotic analysis, we require only that the first-stage estimator concentrates around some deterministic sequence.

[5] For simple first-stage estimators, such as linear adjustments, cross-fitting is not required, but it offers a unifying approach that is suitable for all considered types of adjustments.

[6] In simulations, we choose S to be a moderate number, e.g. 5. We assume that the number of observations both to the left and to the right of the cutoff is divisible by S in order to simplify the notation.

all observations except for the sth fold of the data.

In the second stage, we estimate the RD parameter using our covariate-adjusted outcome variable. For each observation, we generate the outcome using the first-stage estimate based on data from other folds. The final estimator is defined as:

$$\widehat{\tau}_{CF}(h;\widehat{\mu}_n) = \sum_{i=1}^{n} w_i(h) M_i(\widehat{\mu}_{n,s(i)}), \tag{3.6}$$

where the subscript CF refers to cross-fitting.

3.4. THEORETICAL RESULTS

In this section, we formally study the properties of the estimator $\widehat{\tau}_{CF}(h;\widehat{\mu}_n)$ under high-level conditions on the first-stage estimator. We also propose a method to estimate its variance.

3.4.1. **Assumptions.** The conditions we impose in this section consist of standard assumptions in RD designs without covariates as well as high-level assumptions on the first-stage estimator $\widehat{\mu}_n$. Low-level conditions, tailored to specific types of covariate adjustments, are discussed in Section 3.6. Throughout the chapter, we implicitly assume that if a real-valued function f is continuous on $\mathcal{X} \setminus \{0\}$, then also the limits $f(0^-)$ and $f(0^+)$ exist and are finite.

Assumption 3.2. *(i) X_i is continuously distributed with density f_X, which is continuous and bounded away from zero uniformly over $x \in \mathcal{X}$; (ii) The kernel function K is a bounded and symmetric density function that is continuous on its support and equals zero outside some compact set, say $[-1,1]$; (iii) As $n \to \infty$, $h \to 0$ and $nh \to \infty$.*

Assumption 3.2 contains basic conditions for our asymptotic analysis. The assumptions on the density of the running variable, kernel, and bandwidth are standard in the literature.

The next two assumptions concern the first-stage estimator. By construction, its properties are relevant only for observations that are used in the second-stage local linear regression, i.e. the observations with $|X_i| \leq h$. We define $\mathcal{X}_h = \mathcal{X} \cap [-h,h]$ and $\mathcal{Z}_h = \text{supp}(Z_i | X_i \in \mathcal{X}_h)$.

Assumption 3.3. *For all $n \in \mathbb{N}$, there exist a set $\mathcal{T}_n \subset \mathcal{M}_n$ and a function $\bar{\mu}_n \in \mathcal{T}_n$ such that: (i) $\widehat{\mu}_{n,s}$ belongs to $\mathcal{T}_n$ with probability approaching 1 for all $s \in [S]$; (ii) It holds that:*

$$\sup_{\mu \in \mathcal{T}_n} \sup_{x \in \mathcal{X}_h} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i = x\right] = o(1).$$

Assumption 3.3 specifies the required mode of convergence for the first-stage estimator. We require that it belongs with high probability to some realization set $\mathcal{T}_n \subset \mathcal{M}_n$, which contracts around a deterministic sequence of functions $(\bar{\mu}_n)_{n\in\mathbb{N}}$ in a mean-squared-error-type sense. This assumption is weak, as $\bar{\mu}_n$ can be any function, not necessarily the targeted, true function μ_n, and we do not require any specific rate at which $\mathcal{T}_n$ shrinks. In particular, we allow for $\widehat{\mu}_n$ to be based on a misspecified parametric model for the function μ_n, or to have an arbitrarily slowly vanishing bias, as long as the estimator concentrates around some deterministic sequence.

Assumption 3.3 can be ensured in various ways. If the adjustment function is linear, then it follows from convergence of the estimated coefficients if the additional covariates have bounded conditional second moments. Assumption 3.3 is also satisfied if the difference $\widehat{\mu}_{n,s} - \bar{\mu}_n$ converges to zero in the supremum norm on $\mathcal{Z}_h$. Such results are available for example for classic nonparametric estimators in settings with a fixed dimension of the additional covariates. Assumption 3.3 follows also from the unconditional convergence in mean square under mild conditions on the conditional distribution of the additional covariates given the running variable, which can be used to verify this assumption for machine learning methods; see Section 3.6.4 and Appendix 3.A.1.

Assumption 3.4. *For all $n \in \mathbb{N}$, it holds that:*

(i) $\mathbb{E}\left[\mu(Z_i)|X_i = x\right]$ *is twice continuously differentiable in x on $\mathcal{X} \setminus \{0\}$ for all $\mu \in \mathcal{M}_n$;*

(ii) $\sup_{\mu\in\mathcal{T}_n} \sup_{x\in\mathcal{X}_h\setminus\{0\}} \left|\partial_x^1 \mathbb{E}\left[\mu(Z_i) - \bar{\mu}_n(Z_i)|X_i = x\right]\right| = o(1/h)$;

(iii) $\sup_{\mu\in\mathcal{T}_n} \sup_{x\in\mathcal{X}_h\setminus\{0\}} \left|\partial_x^2 \mathbb{E}\left[\mu(Z_i) - \bar{\mu}_n(Z_i)|X_i = x\right]\right| = o(1)$.

Part (i) strengthens Assumption 3.1 and requires that $\mathbb{E}[\mu(Z_i)|X_i = x]$ is twice continuously differentiable to the left and to the right of the cutoff. We emphasize that we do not require continuity of the derivatives of $\mathbb{E}[\mu(Z_i)|X_i = x]$ at the cutoff. This assumption is analogous to the assumptions of Calonico et al. (2019), who assume that $\mathbb{E}[Z_i|X_i = x]$ is (thrice in their case) continuously differentiable to the left and to the right of the cutoff but not necessarily at the cutoff.[7] If, however, $\partial_x^2\mathbb{E}[\mu(Z_i)|X_i = x]$ is continuous at the cutoff, we can exploit this assumption to simplify our asymptotic results; see Corollary 3.1. Parts (ii) and (iii) impose high-level requirements on derivatives of $\mathbb{E}\left[\mu(Z_i) - \bar{\mu}_n(Z_i)|X_i = x\right]$ for $\mu \in \mathcal{T}_n$.

[7] In their main analysis, Calonico et al. (2019) assume only that $\mathbb{E}[Z_i|X_i = x]$ is continuous also at the cutoff, which ensures consistency of the RD estimator. The higher-order smoothness assumptions ensure that standard theory of local linear estimation can be applied to their RD estimator.

Assumption 3.4 follows from Assumption 3.3 under regularity conditions on the conditional distribution of the additional covariates given the running variable. Specific conditions may depend on the estimator used. If the adjustment function is linear, then it follows if each component of $\mathbb{E}[Z_i|X_i = x]$ is twice continuously differentiable on $\mathcal{X}\setminus\{0\}$. Assumption 3.4 also follows whenever the conditional density $f_{Z|X}(z|x)$ is bounded away from zero on its support and the partial derivatives $\partial_x^j f_{Z|X}(z|x)$ are L-Lipschitz continuous in x for all z and $j \in \{0, 1\}$. We discuss further, technical sufficient conditions for this assumption in Appendix 3.A.2.

Assumption 3.5. *There exist constants B and L such that the following conditions hold for all $n \in \mathbb{N}$. (i) $\mathbb{E}[M_i(\bar{\mu}_n)|X_i = x]$ is twice continuously differentiable on $\mathcal{X} \setminus \{0\}$ with L-Lipschitz continuous second derivative bounded by B; (ii) For all $x \in \mathcal{X}$ and some $q > 2$ $\mathbb{E}[(M_i(\bar{\mu}_n) - \mathbb{E}[M_i(\bar{\mu}_n)|X_i])^q|X_i = x]$ exists and is bounded by B; (iii) $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$ is L-Lipschitz continuous and bounded from below by $1/B$ for all $x \in \mathcal{X} \setminus \{0\}$.*

Assumption 3.5 is a translation of standard RD assumptions to the setting with $M_i(\bar{\mu}_n)$ as the outcome variable. We employ these conditions to show asymptotic normality of our proposed RD estimator and to characterize its bias. Part (i) requires that the conditional expectation of the outcome variable is twice continuously differentiable to the left and to the right of the cutoff. Parts (ii) and (iii) impose standard assumptions on conditional moments of the outcome variable.

3.4.2. **Main Asymptotic Results.** In this section, we study the asymptotic properties of our estimator. We define the following kernel constants: $\bar{\nu} = (\bar{\nu}_2^2 - \bar{\nu}_1\bar{\nu}_3)/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)$ and $\bar{\kappa} = \int_0^\infty (k(v)(\bar{\nu}_1 v - \bar{\nu}_2))^2 dv/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)^2$, where $\bar{\nu}_j = \int_0^\infty v^j k(v)dv$.

Theorem 3.1. *Suppose that Assumptions 3.1–3.4 hold.*

(i) It holds that

$$\widehat{\tau}_{CF}(h; \widehat{\mu}_n) = \widehat{\tau}(h; \bar{\mu}_n) + o_p(h^2 + (nh)^{-1/2}).$$

Suppose additionally that Assumption 3.5 holds.

(ii) It holds that

$$\sqrt{nh}\, V(\bar{\mu}_n)^{-1/2} \left(\widehat{\tau}_{CF}(h; \widehat{\mu}_n) - \tau - B(\bar{\mu}_n)h^2\right) \to \mathcal{N}(0, 1),$$

where for $\mu \in \mathcal{M}_n$

$$B(\mu) = \frac{1}{2}\bar{\nu}\left(\partial_x^2 \mathbb{E}[M_i(\mu)|X_i = x]\big|_{x=0^+} - \partial_x^2 \mathbb{E}[M_i(\mu)|X_i = x]\big|_{x=0^-}\right) + o_P(1),$$
$$V(\mu) = \frac{\bar{\kappa}}{f_X(0)}\left(\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]\right).$$

(iii) For all functions $\mu \in \mathcal{M}_n$, it holds that

$$V(\mu) \geq V(\mu_n) = \frac{\bar{\kappa}}{f_X(0)}\Big(\mathbb{E}\left[\mathbb{V}[Y_i|Z_i, X_i]|X_i = 0^+\right] + \mathbb{E}\left[\mathbb{V}[Y_i|Z_i, X_i]|X_i = 0^-\right] + \frac{1}{2}\mathbb{V}\left[\mu_n^+(Z_i) - \mu_n^-(Z_i)|X_i = 0\right]\Big).$$

Part (i) states the key technical result. It shows that the proposed estimator is asymptotically equivalent to its infeasible analog with the estimator $\widehat{\mu}_n$ replaced with the deterministic sequence $\bar{\mu}_n$. We emphasize that this equivalence holds even though the first-stage estimator can converge arbitrarily slowly. This high insensitivity is only possible because for all $k \in \mathbb{N}$

$$\partial_\mu^k \left(\mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-]\right)|_{\mu=\mu_n} = 0 \tag{3.7}$$

where ∂_μ^k is the k-th functional derivative with respect to the function μ. This property is in the spirit of Neyman orthogonality with respect to the adjustment function μ. We discuss it further in Appendix 3.B.3.

Based on the asymptotic equivalence result in part (i), the asymptotic normality shown in part (ii) follows from standard theory of local linear estimation. The approximate variance depends on the sequence $\bar{\mu}_n$ around which the first-stage estimator concentrates. If $\bar{\mu}_n = \mu_n$, then the variance expression is similar to the efficiency bound for estimation of the average treatment effect under unconfoundedness with a constant conditional probability of treatment equal to one half (Hahn, 1998). We discuss the analogy between the covariate adjustments used for randomized experiments and our approach in Appendix 3.B.2.

The proposed covariate adjustments lead to efficiency gains compared to the standard RD estimator in a very wide range of settings, even if $\bar{\mu}_n \neq \mu_n$. We show in Appendix 3.D that $V(\bar{\mu}_n) < V(0)$ if and only if $\mathbb{V}[\mu_n(Z_i) - \bar{\mu}_n(Z_i)|X_i = 0] < \mathbb{V}[\mu_n(Z_i)|X_i = 0]$, i.e. whenever $\bar{\mu}_n(Z_i)$ has some explanatory power for $\mu_n(Z_i)$. This condition is satisfied for example if $\bar{\mu}_n(Z_i)$ represents some nontrivial projection of Y_i on Z_i based on the data in a neighborhood of the cutoff.

The bias expression simplifies under an additional smoothness assumption. If the smoothness condition in Assumption 3.4(i) holds also at the cutoff, then the leading bias does not depend on the function $\bar{\mu}_n$. The simplified bias expression is convenient for conducting statistical inference based the bias-aware approach; see Section 3.5.2.

Corollary 3.1. *Suppose that Assumptions 3.1–3.5 hold and $\partial_x^2 \mathbb{E}[\bar{\mu}_n(Z_i)|X_i = x]$ is con-*

tinuous at the cutoff for all $n \in \mathbb{N}$. Then

$$B(\bar{\mu}_n) = \frac{1}{2}\bar{\nu}\left(\partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^-}\right) + o_P(1).$$

Remark 3.1. It follows from the proof of Theorem 3.1 that our proposed estimator is asymptotically equivalent to the average of RD estimators run on different folds of the data.[8] We prefer our version because existing estimation and inference routines as well as bandwidth selectors can be readily applied to the modified data $(M_i(\widehat{\mu}_{n,s(i)}), X_i)_{i\in[n]}$; see Section 3.5.

3.4.3. **Standard Error.** To estimate the variance of our estimator, we use a standard error of the form

$$\widehat{se}^2_{CF}(h; \widehat{\mu}_n) = \sum_{i=1}^{n} w_i^2(h)\widehat{\sigma}_i^2(\widehat{\mu}_{n,s(i)}),$$

where $\widehat{\sigma}_i^2(\widehat{\mu}_{n,s(i)})$ is an estimator of the variance $\sigma_i^2(\bar{\mu}_n) = \mathbb{V}[M_i(\bar{\mu}_n)|X_i]$. Following Noack and Rothe (2021), we consider a version of the nearest neighbor variance estimator of Abadie et al. (2014).[9] We choose some R, say $R = 5$, which determines the number of neighbors to be used in the variance estimation. Based on the realized running variable, for each unit i, we determine its R nearest neighbors that are on the same side of the cutoff and within the same fold as unit i. Our estimator $\widehat{\sigma}_i^2(\widehat{\mu}_{n,s(i)})$ is proportional to the squared difference between $M_i(\widehat{\mu}_{n,s(i)})$ and its best linear predictor given the running variable based on its R nearest neighbors. We give a formal definition of this estimator in Appendix 3.C.4.

Proposition 3.1. *Suppose that Assumptions 3.1–3.5 hold and that for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$, $\sup_{\mu\in\mathcal{T}_n} \mathbb{E}[(M_i(\mu) - \mathbb{E}[M_i(\mu)|X_i])^4|X_i = x]$ is bounded by B. Then*

$$nh\, \widehat{se}^2_{CF}(h; \widehat{\mu}_n) - V(\bar{\mu}_n) = o_P(1).$$

The additional assumption imposed in Proposition 3.1 strengthens Assumption 3.5(ii). Existence of conditional fourth moments of the outcome variable is often used for showing consistency of standard errors.

[8] A similar point is made by Chernozhukov et al. (2018) in the context of the (unconditional) average treatment effect estimation; cf. their methods DML1 and DML2. Fan et al. (2020) average local linear estimators run on different folds of the data in a conditional average treatment effect estimation problem.

[9] Alternatively, one can use the Eicker-Huber-White (EHW) standard error, but it might be conservative in finite samples; see the discussion by Abadie et al. (2014) in the standard nonparametric regression context.

In this section, we address point estimation and inference. We also discuss how to incorporate different bandwidths on different sides of the cutoff in the second stage.

3.5.1. **Bandwidth Choice.** One of the key steps to implement our estimation procedure is to choose the bandwidth h for the local linear regression in the second stage. We consider two approaches used in the RD literature.

First, we can select the bandwidth that minimizes the asymptotic mean squared error (AMSE), which is defined as:

$$AMSE_n(h) = B(\bar{\mu}_n)^2 h^4 + \frac{1}{nh} V(\bar{\mu}_n).$$

The optimal bandwidth is then given by $h_{\text{opt}} = \left(V(\bar{\mu}_n)/(4B(\bar{\mu}_n)^2)\right)^{1/5} n^{-1/5}$. It can be estimated following the procedures proposed by Imbens and Kalyanaraman (2012) and Calonico et al. (2014). These procedures require estimating $\partial_x^2 \mathbb{E}[M_i(\bar{\mu}_n)|X_i = x]$ to the left and to the right of the cutoff, which can be done using our generated outcome variable under additional smoothness assumptions.

Second, we can adapt the 'bias-aware' approach of Armstrong and Kolesár (2020). They select the bandwidth that minimizes the worst-case mean squared error over a function class formed by imposing a bound on the second derivatives of the considered function. Suppose that $|\partial_x^2 \mathbb{E}[M_i(\bar{\mu}_n)|X_i = x]|$ is bounded by constants B_{M-} and B_{M+} to the left and to the right of the cutoff, respectively, and let $B_M = B_{M-} + B_{M+}$. The leading bias term of our estimator is then bounded in absolute value by $\frac{1}{2}|\bar{\nu}|B_M h^2$. The bandwidth minimizing the corresponding worst-case asymptotic mean squared error is given by $h_{\text{opt}}^{BA} = \left(V(\bar{\mu}_n)/(\bar{\nu}B_M)^2\right)^{1/5} n^{-1/5}$. Implementation of this bandwidth selector requires choosing the smoothness constants B_{M-} and B_{M+}. See Armstrong and Kolesár (2020) and Noack and Rothe (2021) for discussions of the choice of smoothness constants. We note that under the smoothness assumption in Corollary 3.1, it suffices if the smoothness constants B_{M-} and B_{M+} are chosen so as to bound $|\partial_x^2 \mathbb{E}[Y_i|X_i = x]|$ to the left and to the right of the cutoff, respectively.

3.5.2. **Confidence Intervals.** We construct confidence intervals (CIs) for τ based on the asymptotic distribution obtained in part (ii) of Theorem 3.1. The variance $V_n(\bar{\mu}_n)$ can be estimated using the standard error $\widehat{\text{se}}_{CF}(h; \widehat{\mu}_n)$ proposed in Section 3.4.3. To account for the asymptotic bias, we can adapt standard methods available in the nonparametric literature.

First, we consider undersmoothing (US), which relies on selecting a bandwidth of order smaller than $n^{-1/5}$. In this case, the bias is asymptotically negligible, and an

asymptotically valid $1-\alpha$ CI can be formed as:

$$CI_\alpha^{US} = [\widehat{\tau}_{CF}(h;\widehat{\mu}_n) \pm z_{1-\alpha/2} \cdot \widehat{\text{se}}_{CF}(h;\widehat{\mu}_n)], \tag{3.8}$$

where z_u is the u-quantile of the standard normal distribution. The two further approaches allow for the optimal bandwidths discussed in the previous section, which are of order $n^{-1/5}$.

Second, the robust bias corrections (RBC) proposed by Calonico et al. (2014) can be easily adapted to our setting. In this approach, we subtract an estimate of the leading bias term and account for the additional variation in the bias-corrected estimator when forming a CI. These additional steps can be conducted using our generated outcome variable $M_i(\widehat{\mu}_{n,s(i)})$ instead of the original outcome Y_i under further regularity conditions. Let $\widehat{\tau}_{CF}^{RBC}(h;\widehat{\mu}_n)$ be the bias-corrected estimator and $\widehat{\text{se}}_{CF}^{RBC}(h;\widehat{\mu}_n)$ the corresponding standard error. The proposed CI is given by:

$$CI_\alpha^{RBC} = [\widehat{\tau}_{CF}^{RBC}(h;\widehat{\mu}_n) \pm z_{1-\alpha/2} \cdot \widehat{\text{se}}_{CF}^{RBC}(h;\widehat{\mu}_n)], \tag{3.9}$$

The third approach adapts the 'bias-aware' approach of Armstrong and Kolesár (2020). Under the assumption of bounded second derivatives discussed in the previous section, it follows that an asymptotically valid $1-\alpha$ confidence interval can be formed as:

$$CI_\alpha^{BA} = [\widehat{\tau}_{CF}(h;\widehat{\mu}_n) \pm \text{cv}_{1-\alpha}(\widehat{r}(h)) \cdot \widehat{\text{se}}_{CF}(h;\widehat{\mu}_n)],$$

where $\widehat{r}(h) = \frac{1}{2}|\bar{\nu}|B_M h^2/\widehat{\text{se}}_{CF}(h)$ and $\text{cv}_{1-\alpha}(t)$ is the $1-\alpha$ quantile of the folded normal distribution $|\mathcal{N}(t,1)|$. One can also account for the maximal bias of the infeasible estimator $\widehat{\tau}(h;\bar{\mu}_n)$ conditional on $\mathcal{X}_n$ instead of bounding only the leading bias term.

3.5.3. **Different Bandwidths.** Our estimation procedure introduced in Section 3.3.3 employs a single bandwidth in the second-stage local linear regression. In some empirical settings, however, it might be desirable to run two separate local linear regressions using different bandwidths on different sides of the cutoff. The reason for that might be, for example, that the curvature of the conditional expectation of the outcome variable or its conditional variance are different to the left and to the right of the cutoff. Another reason for choosing different bandwidths might be that the density of the running variable is very steep at the cutoff, so that the numbers of observations with the running variable in $(-h_{\text{opt}}, 0)$ and $(0, h_{\text{opt}})$ are substantially different.

It is straightforward to account for different bandwidths in the asymptotic distribution of our estimator, but the adjustment term based on μ_n is no longer optimal in such a case. We therefore generalize the optimality result in part (iii) of Theorem 3.1. When

bandwidths h_- and h_+ are used to the left and to the right of the cutoff, respectively, then the variance of our estimator in large samples is approximately equal to:

$$\widetilde{V}(\bar{\mu}_n) = \omega_+ \mathbb{V}[M_i(\bar{\mu}_n)|X_i = 0^+] + \omega_- \mathbb{V}[M_i(\bar{\mu}_n)|X_i = 0^-],$$

where $\omega_- = \sum_{i=1}^n w_{i,-}(h_-)^2$ and $\omega_+ = \sum_{i=1}^n w_{i,+}(h_+)^2$ and the weights $w_{i,-}$ and $w_{i,+}$ correspond to the local linear estimators run using the data to the left and to the right of the cutoff, respectively. The explicit expressions are given in Appendix 3.C.1.[10] The weights ω_- and ω_+ capture the inverse of the effective sample size to the left and to the right of the cutoff, respectively.

We show in Appendix 3.D that $\widetilde{V}(\mu)$ is minimized by the function

$$\mu_n^*(z) = \frac{\omega_-}{\omega_- + \omega_+}\mu_n^-(z) + \frac{\omega_+}{\omega_- + \omega_+}\mu_n^+(z) \tag{3.10}$$

in the sense that $\widetilde{V}(\mu_n^*) \leq \widetilde{V}(\mu)$ for all $\mu \in \mathcal{M}_n$. This result is consistent with Theorem 3.1 because $\omega_-/(\omega_- + \omega_+) \to 1/2$ under our assumptions if $h_- = h_+$.

We remark that for any given bandwidths the above weighting scheme puts more weight to the side of the cutoff where the effective sample size is smaller. The reason for that is apparent in the proof given in Appendix 3.D, where we show that minimization of $\widetilde{V}(\mu)$ is equivalent to minimization of $\widetilde{\mathcal{V}}(\mu) = \omega_+ \mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0] + \omega_- \mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0]$. If, for example, ω_+ is large compared to ω_-, then choosing μ so as to make $\mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0]$ small is relatively more important than decreasing the value of $\mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0]$. Accordingly, μ_n^+ receives a higher weight in (3.10) in such a case.

3.6. EXAMPLES OF COVARIATE ADJUSTMENTS

In this section, we give primitive conditions for our high-level Assumptions 3.3 and 3.4, which concern the properties of the first-stage estimator. We consider in turn: linear, non-linear parametric, local linear, and generic machine learning adjustments. In Sections 3.6.1–3.6.3, where we consider methods suitable for settings with a low-dimensional covariate, we assume that the distribution of W_i does not change with n.

[10] Apart from allowing for different bandwidths, $\widetilde{V}(\mu)$ differs from $V(\mu)$ in Theorem 3.1 in that it does not rely on kernel-weighted sums of X_i to converge to their limits. As such, $\widetilde{\mathcal{V}}(\bar{\mu}_n)$ may capture the finite-sample variance of our estimator more accurately. Still, this expression remains valid only asymptotic as we use $\mathbb{V}[M_i(\mu)|X_i = x]$ evaluated to the left and to the right of the cutoff, rather than for each X_i separately.

3.6.1. **Linear Adjustments.** We define a linear estimator using observations close to the cutoff:

$$\widehat{\beta}_s = \arg\min_\beta \sum_{s \in I_s^c} K(X_i/h)(Y_i - \beta^\top (Z_i^\top T_i, Z_i^\top (1-T_i), X_i, T_i X_i, T_i, 1)^\top)^2. \tag{3.11}$$

Let $\widehat{\beta}^+_{s,Z}$ denote the first d components of $\widehat{\beta}_s$ and let $\widehat{\beta}^-_{s,Z}$ be the next d components of $\widehat{\beta}_s$. We define $\widehat{\mu}_{n,s}(z) = z^\top \widehat{\beta}_{s,Z}$, where $\widehat{\beta}_{s,Z} = \frac{1}{2}(\widehat{\beta}^+_{s,Z} + \widehat{\beta}^-_{s,Z})$.[11] Let $\bar{Z}_i = (1, Z_i^\top, X_i/h_1)^\top$. Assumptions 3.3 and 3.4 hold if we can ensure that the estimated slope coefficients concentrate around some deterministic sequence and the conditional expectation $\mathbb{E}[Z_i|X_i = x]$ is sufficiently smooth.

Assumption 3.6. *(i) Each component of $\mathbb{E}[Z_i|X_i = x]$ is twice differentiable on $\mathcal{X} \setminus \{0\}$ with L-Lipschitz continuous second derivative for some constant L; (ii) The limit as $n \to \infty$ of $\mathbb{E}[K_{h_1}(X_i)\bar{Z}_i\bar{Z}_i^\top]$ is non-singular; (iii) $\mathbb{E}[Z_i^\top Z_i|X_i = x]$ is bounded uniformly over $x \in \mathcal{X}$.*

Proposition 3.2. *Suppose that Assumption 3.6 holds and either (i) $h_1 \to 0$ and $nh_1 \to \infty$ or (ii) $h_1 \to c > 0$. Then Assumptions 3.3 and 3.4 are satisfied.*

This type of adjustments bears a resemblance to the procedure of Calonico et al. (2019). Specifically, they obtain their estimator from a regression as in (3.11) but with two main differences. First, they using the whole sample. With these simple adjustments, cross-fitting is not necessary in our procedure, but it does not have any adverse effects. Second, they impose the restriction that $\widehat{\beta}^+_Z = \widehat{\beta}^-_Z$. Doing so implies by standard OLS algebra that $\widehat{\mu}_n(z)$ puts more weight to the side of the cutoff with the larger effective sample size. As can be seen in Section 3.5.3, this type of weighting is not optimal.[12]

3.6.2. **Non-linear Parametric Adjustments.** Suppose that the researcher uses some parametric specification $m_\beta(z) = \frac{1}{2}(m^-_\beta(z) + m^+_\beta(z))$ for the function μ_n, which can be based, e.g., on the logit or probit model. This specification might be correct or incorrect. The function m_β is known up to a finite-dimensional parameter $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$. We assume that there is an estimator $\widehat{\beta}$ converging to some nonrandom probability limit $\bar{\beta}$. Classic conditions for consistency in M-estimation problems are given, e.g., by Newey and McFadden (1994).

Assumption 3.7. *(i) For some $\bar{\beta}$ and $r_n \to 0$, $\|\widehat{\beta} - \bar{\beta}\|_\infty = O_p(r_n)$; (ii) For all $\beta_1, \beta_2 \in \mathcal{B}$, $z \in \mathcal{Z}$, and some constant G, $|m_{\beta_1}(z) - m_{\beta_2}(z)| \leq G||\beta_1 - \beta_2||_\infty$.*

[11] As discussed in Section 3.3.2, it suffices to estimate μ_n up to a constant.

[12] A similar point is made by Negi and Wooldridge (2020) in the context of randomized experiments.

Assumption 3.7 guarantees that the first-stages estimator converges in the supremum norm to some limiting function. With this mode of convergence, Assumption 3.3 follows trivially, and Assumption 3.4 also holds under regularity conditions on the conditional distribution of the additional covariates given the running variable. For concreteness, we assume that Z_i is continuously distributed given X_i, but analogous results can be derived if the additional covariates have a discrete distribution.

Proposition 3.3. *Suppose that Assumptions 3.1, 3.2, and 3.7 hold. Moreover, Z_i has bounded support and $\partial_x^j f_{Z|X}(z|x)$ is L-Lipschitz continuous in x for all z and $j \in \{0,1\}$. Then Assumptions 3.3 and 3.4 are satisfied.*

3.6.3. **Nonparametric Adjustments.** We consider covariate adjustments based on classic nonparametric methods, which are suitable if the number of additional covariates is not too large. To fix ideas, we focus on local linear estimation (Fan and Gijbels, 1996), but similar results can be obtained for example for sieve estimation (Newey, 1997).

For $z \in \mathbb{R}^d$, we define the multivariate kernel as the product of univariate kernels, $\mathcal{K}_h(z) = \prod_{i=1}^d K_h(z_i)$, where $K_h(v) = \frac{1}{h}K(v/h)$.[13] We define estimators of $\mu_n^+(z)$ and $\mu_n^-(z)$ using data in the complement of the sth fold as:

$$\widehat{\mu}_{n,s}^+(z) = e_1^\top \arg\min_\beta \sum_{i \in I_s^c} T_i K_{h_X}(X_i)\mathcal{K}_{h_Z}(Z_i - z)(Y_i - \beta^\top(1, (Z_i - z)^\top, X_i))^2,$$
$$\widehat{\mu}_{n,s}^-(z) = e_1^\top \arg\min_\beta \sum_{i \in I_s^c} (1 - T_i) K_{h_X}(X_i)\mathcal{K}_{h_Z}(Z_i - z)(Y_i - \beta^\top(1, (Z_i - z)^\top, X_i))^2.$$

In Assumption A.3.8 in Appendix 3.C.6, we impose standard assumptions on the data generating process for the local linear estimator.

Proposition 3.4. *Suppose that Assumptions 3.1, 3.2, and A.3.8 hold. Further, assume that $h_X \to 0$, $h_Z \to 0$, $\log(n)/(nh_X h_Z^d) \to 0$, and $\partial_x^2 f_{Z|X}(z|x)$ is L-Lipschitz continuous in x for all z. Then Assumptions 3.3 and 3.4 are satisfied.*

Under Assumption A.3.8 and the bandwidth conditions of Proposition 3.4, Masry (1996) shows that the local linear estimator is uniformly consistent. Using this result, Assumption 3.1 follows trivially. Assumption 3.4 also follows under the additional smoothness conditions; see the discussion in Appendix 3.A.2.

We emphasize that the bandwidth conditions are very mild, and they can be chosen, e.g., via cross-validation under further, standard regularity conditions. With a moderate number of covariates, it is optimal to choose a relatively large bandwidth, but this is

[13] The kernel chosen for the local linear first-stage estimator can be also different from the kernel used in the second stage.

allowed as long as they converge to zero. In general, with our method is advisable to oversmooth, rather than undersmooth when choosing the bandwidths in order to guarantee that the estimator is not too volatile. Oversmoothing comes at the cost of a possible increase in the variance of the final estimator, but it renders the normal approximation of the asymptotic distribution more reliable in finite samples.

3.6.4. **Adjustments Based on Machine Learning Methods.** We outline a general approach to ensuring that our high-level assumptions hold for many machine learning methods. Results about estimation of conditional expectations using machine learning methods typically concern convergence in mean square. We can make use of these results by estimating the functions μ_n^- and μ_n^+ based on narrow, fixed 'slices' of the data to the left and to the right of the cutoff, respectively.[14] Specifically, for any fixed h_1, we can readily obtain the result that the selected estimator belongs to some realization set $\mathcal{T}_n$ with probability approaching one, and

$$\sup_{\mu \in \mathcal{T}_n} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i \in \mathcal{X}_{h_1}\right] = o(1), \tag{3.12}$$

where $\bar{\mu}_n(z) = \frac{1}{2}(\bar{\mu}_n^+(z) + \bar{\mu}_n^-(z))$ with $\bar{\mu}_n^+(z) = \mathbb{E}[Y_i | Z_i = z, X_i \in (0, h_1)]$ and $\bar{\mu}_n^-(z) = \mathbb{E}[Y_i | Z_i = z, X_i \in (-h_1, 0)]$. If the conditional distribution of the additional covariates given the running variable is sufficiently smooth on the interval $(-h_1, h_1)$, then the above property implies that Assumptions 3.3 and 3.4 hold; see Appendix 3.A for more details.

Primitive conditions for (3.12) are available for a variety of machine learning techniques, e.g. post-lasso (Belloni et al., 2012), random forests (Breiman, 2001; Wager and Athey, 2018), and deep neural networks (Farrell et al., 2021). Hence, we can flexibly choose a method that is best-suited for a given dataset under the assumptions imposed.

With fixed h_1, $\bar{\mu}_n$ might be different from μ_n. Our theory allows for that, but this procedure in general does not achieve the optimal variance $V(\mu_n)$. In the previous section, we show that for the local linear estimator, the optimal variance can be achieved by choosing h_1 that converges to zero. It would be interesting to formally study the setting with $h_1 \to 0$ for other methods. We leave this for future research.

[14] Restricting the sample corresponds to weighting the observations based on a uniform kernel. Our reasoning applies also to any other kernel weighting scheme, e.g. using the triangular kernel.

We compare the finite sample performance of our proposed estimator for different first-stage estimation methods in a Monte Carlo study.

3.7.1. **Setup.** We consider four models, which differ in the number of covariates entering the outcome equation, which we denote by $L \in \{0, 4, 10, 25\}$. The running variable X_i follows the uniform distribution over $[-1, 1]$. There are four independent, baseline covariates, denoted by Z_i^b, which are distributed uniformly over $[-1 + x^2, 1 + x^2]^4$ conditional on $X_i = x$. We generate further covariates based on the baseline covariates using Hermit polynomials. Let $b_l(Z_i^b)$ denote the l-th covariate. The outcome is generated according to the following model:

$$Y_i = D_i + \mu_L(X_i, Z_i) + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 0.25)$ and

$$\mu_L(X_i, Z_i) = \text{sign}(X_i) \cdot (X_i + X_i^2 - 2\,(X_i - 0.1)_+^2) + \bar{\iota}_L(\rho) \sum_{l=1}^{L} b_l(Z_i^b).$$

For positive L and ρ, we chose the coefficient $\bar{\iota}_L(\rho)$ so that $\mathbb{V}[\mu_L(0, Z_i)|X_i = 0] = \rho^2 \mathbb{V}[\varepsilon_i]$. In this definition, ρ represents the signal to noise ratio at the cutoff given the treatment status. It determines the scope for improvements from using covariates, but it does not affect the relative performance of different covariate adjustments. For concreteness, in the main text, we consider $\rho = 3$. We report simulation results for further values of ρ in Appendix 3.E. The results are based on $5,000$ simulation draws. The sample size is $2,000$ for the main results.

We consider in total seven implementations of the first-stage estimator: (i) the standard RD estimator with no covariate adjustments; (ii) the infeasible, optimal RD estimator with covariate adjustments based on the true conditional expectation function; (iii) the infeasible RD estimator with adjustments based on the best linear prediction on the population level of the true conditional expectation function given the four baseline covariates.[15] We consider four feasible adjustment functions based on:[16] (iv) a linear regression given the four baseline covariates; (v) a local linear regression given the four

[15] We obtain the population projection coefficients based on 10^7 draws with $X_i = 0$ and $\varepsilon_i = 0$. We fix this estimate through all simulations for each data generating process.

[16] In the first-stage, the observations are weighted using kernel weights with the bandwidth selected for the standard RD estimator.

baseline covariates; (vi) a post-lasso regression given 200 covariates; and (vii) a random forest with the four baseline covariates.

To keep the exposition simple, in the main text, we consider only the bias-aware approach for the implementation of the second stage. Our procedure is based on the true bound on the second derivative of the conditional expectation of the outcome variable. The bandwidth is chosen to be optimal in terms of the estimated worst-case mean squared error. The main qualitative conclusions of our simulation study hold also for robust bias corrections and undersmoothing. We present these results in Appendix 3.E. There we also compare our estimators to the linear covariates adjustment method proposed by Calonico et al. (2019).[17]

Table 3.1: Simulation Results

	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
	Model 1: L=0					**Model 2: L=4**				
Standard	97.0	-1.4	7.4	32.4	43.2	96.1	-7.1	18.6	81.8	68.8
Optimal Inf	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2
Linear Inf	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2
Linear	97.0	-1.4	7.4	32.7	43.3	96.7	-1.5	7.5	32.6	43.3
Local Linear	97.0	-1.4	7.4	32.7	43.3	96.8	-1.4	7.5	32.7	43.3
Lasso	96.7	-1.4	7.6	33.1	43.6	96.6	-2.1	8.8	38.3	46.6
Forest	96.8	-1.5	7.6	33.1	43.6	96.7	-2.1	8.7	37.9	46.5
	Model 3: L=10					**Model 4: L=25**				
Standard	96.4	-9.5	19.1	87.6	79.3	95.9	-6.3	18.5	81.0	68.5
Optimal Inf	96.5	-1.3	7.6	32.5	43.2	96.9	-1.3	7.4	32.4	43.2
Linear Inf	96.7	-4.8	12.7	56.2	61.8	96.8	-4.3	10.3	47.2	59.0
Linear	95.9	-4.0	13.7	59.1	59.7	96.5	-4.3	10.8	49.2	58.8
Local Linear	96.3	-1.6	8.3	35.6	45.2	96.8	-1.6	8.2	35.9	45.6
Lasso	96.2	-2.0	9.2	39.1	46.7	96.8	-1.4	7.7	34.0	44.3
Forest	96.6	-1.9	8.5	37.2	46.9	97.1	-2.2	9.3	41.3	49.0

Notes: Results based on 5000 Monte Carlo draws for the bias-aware approach. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h).

[17] All computations are carried out with the statistical software `R`. The Hermit-polynomials are computed using the package `calculus`. To implement the first-stage estimators, we use the following packages: `np` for local polynomial regressions; `glmnet` for lasso regressions; `grf` for random forests, where predictions are based on 200 trees. In the second stage, a triangular kernel is used and EHW standard errors are computed. The bias-aware approach is based on the package `RDHonest`, and the other two approaches are implemented using the package `rdrobust`.

3.7.2. **Simulations Results.** Table 3.1 reports estimation and inference results for different types of adjustments. The CIs for all estimators have simulated coverage rates close to their nominal ones.[18] First, we compare the standard RD estimator and the infeasible estimators. In Model 1, these estimators are numerically equal. In Models 2–4, where the covariates have some explanatory power for the outcome, the infeasible estimators have a substantially lower standard deviation than the standard estimator has. If the linear model is misspecified, the standard deviation of the optimal infeasible estimator is much smaller than that of the infeasible estimator with linear adjustments. We now turn to the feasible covariate-adjusted RD estimators. As predicted by Theorem 3.1, their mean standard deviations are close to those of their respective infeasible estimator, with only a slight increase due to the first-stage estimation.

In Figures 3.1 and 3.2, we compare the difference between the optimal infeasible RD estimator and two feasible ones: with adjustments based on local linear regression and post-lasso regression for several choices of the tuning parameters. In each simulation draw, we find the MSE-optimal tuning parameters via cross-validation, and then scale it down or up by different factors.[19] We consider two sample sizes, $n = 2000$ and $n = 10000$. We normalize the difference by the standard error of the optimal infeasible RD estimator.

In Figure 3.1, we observe that the normalized difference between the estimators is relatively small for a wide range of bandwidths around the optimal one. By comparing panels (a) and (b), we can see that these normalized differences become smaller as the sample size increases, which illustrates the asymptotic equivalence result in part (i) of Theorem 3.1. For a given sample size, the average absolute value of the normalized differences is U-shaped as a function of the bandwidth. If the bandwidth chosen in the first stage is too small, then the local linear estimator is very unstable. In this case, the property in Assumption 3.3 is not a good description of its finite-sample behavior, and the equivalence result in Theorem 3.1 fails. If the bandwidth is chosen to be too large, the local linear estimator has a relatively small variance, but it might be heavily biased, and it is effectively very similar to the linear estimator. In this case, the equivalence to an infeasible estimator holds with a different limiting sequence $(\bar{\mu}_n)_{n\in\mathbb{N}}$. We expect the estimator to be less efficient, but we emphasize that our inference procedure remain valid in this case.

Figure 3.2 shows a very similar pattern as Figure 3.1. If the penalty parameter in the lasso regression is chosen to be too small, effectively all covariates are classified as

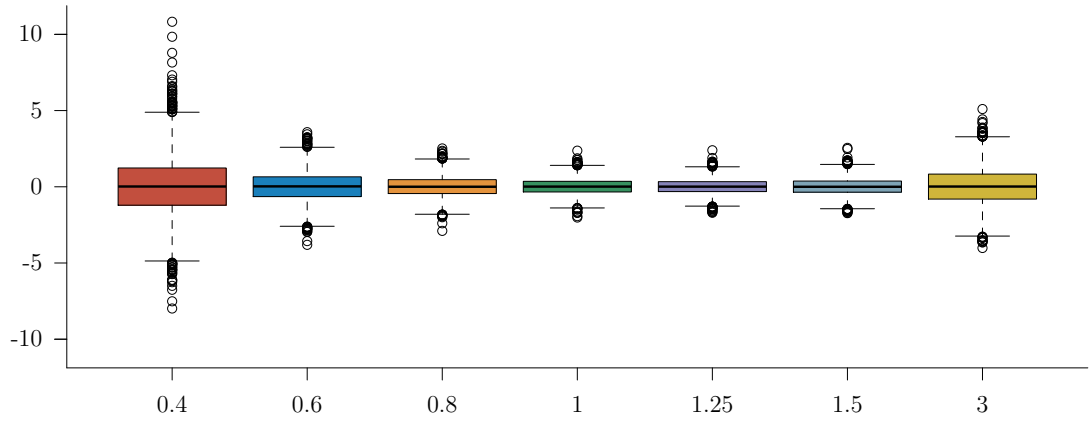
[18] In the considered models, the maximal bias is not achieved, so that the bias-aware CIs are conservative.

[19] To facilitate comparisons of different covariate adjustments, in each simulation draw, we use the bandwidth selected for the standard RD estimator in the second stage across all different methods.

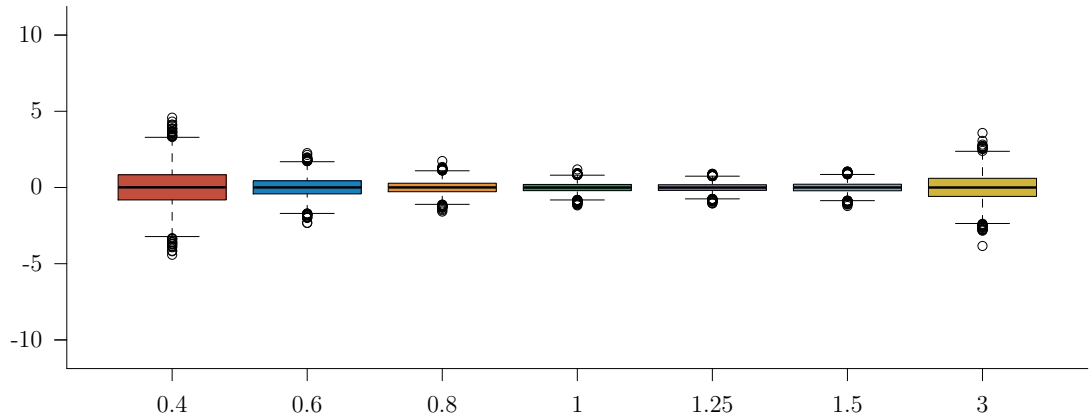
relevant, and the first-stage estimator has a high variance. In contrast, if the penalty parameter is chosen to be too large, very few covariates are classified as relevant. In this case, the RD estimator behaves similarly to the standard RD estimator.

3.8. CONCLUSIONS

Linear covariate adjustments are commonly used in RD designs to improve efficiency of the standard RD estimator. In this chapter, we propose a class of RD estimators that allow for nonparametric covariate adjustments, which can reduce the variance of the RD estimator even further. We allow for a wide range of covariate adjustments under mild conditions. Despite using possibly highly complex covariate adjustments, inference on the RD parameter can be conducted using standard methods available in the literature. We illustrate our results in a simulation study.

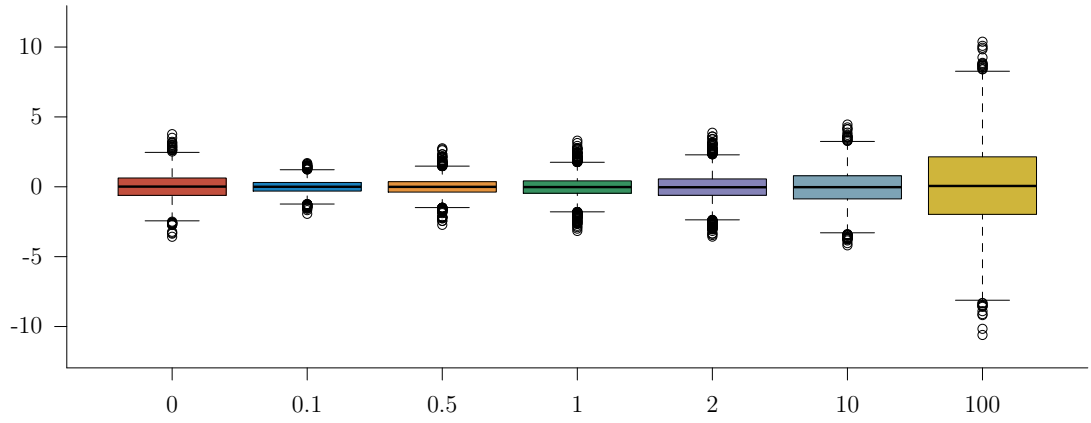


(a) Sample size $n = 2,000$.

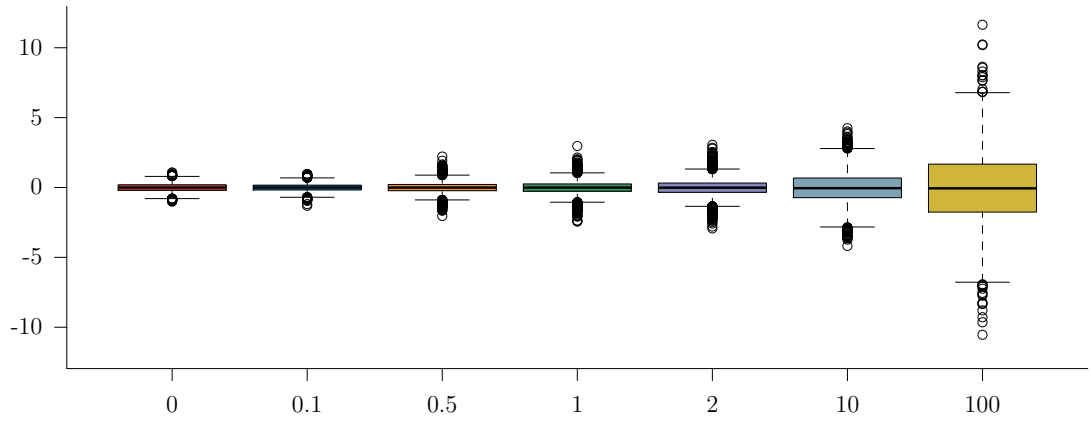


(b) Sample size $n = 10,000$.

Figure 3.1: Normalized difference of RD estimates with local linear adjustments.
Notes: Difference between optimal infeasible and feasible RD estimate normalized by standard deviation of infeasible estimator. We consider various scaling factors for the cross-validated MSE-optimal first-stage bandwidth. Simulations are based on Model 3. Panel (a) shows simulation results for $n = 2,000$, and Panel (b) for $n = 10,000$.



(a) Sample size $n = 2,000$.



(b) Sample size $n = 10,000$.

Figure 3.2: Normalized difference of RD estimates with post-lasso regression adjustments. *Notes:* Difference between optimal infeasible and feasible RD estimate normalized by standard deviation of infeasible estimator. We consider various scaling factors for the cross-validated MSE-optimal first-stage penalty parameter. Simulations are based on Model 3. Panel (a) shows simulation results for $n = 2,000$, and Panel (b) for $n = 10,000$.

3.A. FURTHER SUFFICIENT CONDITIONS FOR MAIN ASSUMPTIONS

In this section, we discuss sufficient conditions for our high-level Assumptions 3.3 and 3.4.

3.A.1. **Sufficient Conditions for Assumption 3.3.** We outline a generic way of ensuring that Assumption 3.3 holds, which can be employed for a wide range of estimators. For concreteness, we assume that the additional covariates are continuously distributed conditional on the running variable, but similar results can be derived for discrete distributions or intermediate cases.

Many results in the machine learning literature concern convergence in mean square, which means that we can obtain the following property:

$$\sup_{\mu\in\mathcal{T}_n} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i \in \mathcal{X}_h\right] = o(1). \tag{A.3.13}$$

We can infer our assumption from the above condition if the conditional distribution of the additional covariates does not change abruptly around the cutoff. Specifically, suppose that

$$\sup_{x\in\mathcal{X}_h} \sup_{z\in\mathcal{Z}_h} \frac{f_{Z|X}(z|x)}{f_{Z|X\in\mathcal{X}_h}(z)} < B, \tag{A.3.14}$$

for some constant B and h small enough. If the conditions in (A.3.13) and (A.3.14) hold, then Assumption 3.3 is satisfied because::

$$\begin{aligned}
&\sup_{\mu\in\mathcal{T}_n} \sup_{x\in\mathcal{X}_h} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i = x\right] \\
&\qquad = \sup_{\mu\in\mathcal{T}_n} \sup_{x\in\mathcal{X}_h} \int_{\mathcal{Z}_h} (\mu(Z_i) - \bar{\mu}_n(Z_i))^2 f_{Z|X\in\mathcal{X}_h}(z) \frac{f_{Z|X}(z|x)}{f_{Z|X\in\mathcal{X}_h}(z)} dz \\
&\qquad \leq B \sup_{\mu\in\mathcal{T}_n} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}_n(Z_i))^2 | X_i \in \mathcal{X}_h\right] = o(1).
\end{aligned}$$

3.A.2. **Sufficient Conditions for Assumption 3.4.** We show that Assumption 3.4 can be inferred from the convergence imposed in Assumption 3.3 under mild additional smoothness conditions on the conditional distribution of the additional covariates given the running variable. This can be most intuitively seen when the support $\mathcal{Z}$ is discrete. In the continuous case some additional integrability conditions are needed.

3.A.2.1. *Discrete Additional Covariates.* Suppose that the support of the additional covariates, $\mathcal{Z}$, is finite. In this case, Assumption 3.3 implies that $\sup_{\mu\in\mathcal{T}_n} \sup_{z\in\mathcal{Z}_h} |\mu(z) - \bar{\mu}(z)| = o(1)$. Then for $j \in \{1, 2\}$,

$$\partial_x^j \mathbb{E}[\mu(Z_i) - \bar{\mu}_n(Z_i) | X_i = x] = \sum_{z\in\mathcal{Z}} (\mu(z) - \bar{\mu}(z)) \partial_x^j \mathbb{P}[Z_i = z | X_i = x].$$

Given Assumption 3.3, Assumption 3.4 holds if $\sup_{x\in\mathcal{X}_h\setminus\{0\}}\sup_{z\in\mathcal{Z}_h}\partial_x^1\mathbb{P}[Z_i=z|X_i=x]=O(1/h)$ and $\sup_{x\in\mathcal{X}_h\setminus\{0\}}\sup_{z\in\mathcal{Z}_h}\partial_x^2\mathbb{P}[Z_i=z|X_i=x]=O(1)$.

3.A.2.2. *Continuous Additional Covariates.* Suppose that the additional covariates are continuously distributed given the running variable, and that the conditional density $f_{Z|X}(z|x)$ is twice differentiable with respect to x on $\mathcal{X}\setminus\{0\}$ for all z. Further, assume that for $j\in\{0,1\}$, there exists a function $H_j(z)$ integrable over $\mathcal{Z}$ such that for all x_1,$x_2\in(0,h)$,

$$\left|\partial_x^j f_{Z|X}(z|x_1)-\partial_x^j f_{Z|X}(z|x_2)\right|+\left|\partial_x^j f_{Z|X}(z|{-x_1})-\partial_x^j f_{Z|X}(z|{-x_2})\right|\leq H_j(z)|x_1-x_2|.$$

In this setting, Assumption 3.4 holds if in addition to Assumption 3.3 for $j\in\{0,1\}$ either

(i) $\sup\limits_{\mu\in\mathcal{T}_n}\sup\limits_{z\in\mathcal{Z}_h}|\mu(z)-\bar{\mu}_n(z)|\to 0$, or

(ii) $\sup\limits_{x\in\mathcal{X}_h\setminus\{0\}}\mathbb{E}\left[\left(H_j(Z_i)/f_{Z|X}(Z_i|x)\right)^2\big|X_i=x\right]<\infty$.

The first condition requires that the first-stage estimator converges in the supremum norm. This condition is satisfied for classic nonparametric estimators such as kernel and sieve estimators, see, e.g., Masry (1996); Newey (1997).

The second condition ensures that Assumption 3.4 holds in combination with L_2-convergence assumed in Assumption 3.3. The additional integrability condition holds for example if the conditional density $f_{Z|X}(z|x)$ is bounded away from zero and $\partial_x^j f_{Z|X}(z|x)$ is bounded for $j\in\{1,2\}$ uniformly in x and z.

3.B. RELATION TO THE LITERATURE

In this section, we compare our asymptotic results with those of Frölich and Huber (2019) and draw an analogy between our approach and double-robust estimation of the average treatment effect in randomized experiments. We also discuss the relation to estimation based on Neyman-orthogonal moments.

3.B.1. **Comparison with Frölich and Huber (2019).** Our procedure with the local linear estimator in the first stage is related to that proposed by Frölich and Huber (2019). Under our assumptions, for sharp designs with the same kernels of order $\lambda=2$ used in

both stages, their bias expression simplifies to:

$$\begin{aligned} bias^{FH} = \frac{\bar{\nu}}{2} \int (\mu_n^+(z) - \mu_n^-(z) - \tau) \frac{\partial_x^2 f(x,z)}{f_X(0)} dz h^2 \\ + \frac{\bar{\nu}}{2} \int (\partial_x^2 \mu_n(x,z)|_{x=0^+} - \partial_x^2 \mu_n(x,z)|_{x=0^-}) f_{Z|X}(z|0) dz h_x^2 \\ + \frac{\nu_2}{2} \sum_{l=1}^{L} \int (\partial_{z_l}^2 \mu_n^+(z) - \partial_{z_l}^2 \mu_n^-(z)) f_{Z|X}(z|0) dz h_z^2, \end{aligned}$$

where $\mu_n(x,z) = \mathbb{E}[Y_i|X_i = x, Z_i = z]$, $\bar{\nu}$ is the "boundary bias kernel constant" defined before Theorem 3.1, and $\nu_2 = \int v^2 k(v) dv$. This expression has a more complicated than the bias in Theorem 3.1, and it does not simplify further under the additional smoothness assumption in Corollary 3.1.

The asymptotic variance equals the variance of our proposed estimator when the first-stage estimator is consistent, $\mathcal{V}^{FH} = V(\mu_n)$. The procedure of Frölich and Huber (2019), however, allows for at most three continuous additional covariates if a second-order kernel is used in the first-stage local linear regression.

3.B.2. **Analogy with ATE estimation.** RD designs are very similar in nature to randomized controlled trials. Conditional on the running variable being close to the cutoff, if the distribution of the covariates evolves continuously through the cutoff, the probability of observing a unit with any given value of the additional covariate is approximately the same to the left and to the right of the cutoff. Hence, the treatment is as if randomly assigned and the propensity score is constant.

In an experiment where the treatment probability is constant across covariates, the augmented inverse probability weighted estimator of the average treatment effect is given by:

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{m}_1(Z_i) - \widehat{m}_0(Z_i) + \frac{T_i(Y_i - \widehat{m}_1(Z_i))}{\widehat{p}} - \frac{(1-T_i)(Y_i - \widehat{m}_0(Z_i))}{1-\widehat{p}} \right), \quad \text{(A.3.15)}$$

where, $\widehat{m}_t(z)$ is an estimator of $\mathbb{E}[Y_i|Z_i = z, T_i = t]$ for $t \in \{0,1\}$, and $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} T_i$ is the proportion of treated units.

This estimator can be also represented as the difference in means in the treatment and control group of a modified outcome variable:

$$\widehat{\tau} = \frac{\sum_{i=1}^{n} T_i(Y_i - \widehat{m}(Z_i;\widehat{p}))}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n}(1-T_i)(Y_i - \widehat{m}(Z_i;\widehat{p}))}{\sum_{i=1}^{n}(1-T_i)}, \quad \text{(A.3.16)}$$

where $\widehat{m}(z;\widehat{p}) = (1-\widehat{p})\widehat{m}_1(z) + \widehat{p}\,\widehat{m}_0(z)$. Our proposed estimator is analogous to the expression in (A.3.16) in the sense that it is the difference between estimates from the

treatment and control group, except that we replace the estimated propensity score $\widehat{p}$ with the known one, which equals one half.

3.B.3. **Insensitivity to the First Stage.** In two-stage estimation procedures, the first stage generally affects the properties of the final estimator. This complication, however, can be avoided using estimators based on so-called Neyman-orthogonal moments (Neyman, 1959, 1979), whose derivative with respect to the nuisance parameter estimated in the first stage is zero. This method has been recently used in the semiparametric literature in settings where a, possibly high-dimensional, nuisance parameter is estimated using machine learning methods; see, e.g., Belloni et al. (2017); Chernozhukov et al. (2018). In our context, Neyman-orthogonality means that

$$\partial_\mu^1 \left(\mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-] \right) \big|_{\mu=\mu_n} = 0, \tag{A.3.17}$$

where ∂_μ^k denotes the k-th functional derivative in all possible directions.

Our setting is related to estimation problems with Neyman-orthogonal moments but it differs in two main aspects. The property in Equation (3.7) is much stronger than (A.3.17) because functional derivatives of all orders evaluated at any function $\mu \in \mathcal{M}_n$ vanish. However, this property holds only conditional on the running variable been at the cutoff, whereas any estimation procedure has to rely on the data in some neighborhood of the cutoff.

3.C. PROOFS OF MAIN RESULTS

3.C.1. **Additional Notation.** We use the following notation throughout the proofs. For $s \in [S]$, $i \in I_{s(i)}$, and $j \in \{0, 1\}$, we define the local linear weights as

$$\begin{aligned}
w_{i,s}^{(j)}(h) &= w_{i,s,+}^{(j)}(h) - w_{i,s,-}^{(j)}(h), \\
w_{i,s,+}^{(j)}(h) &= e_{j+1}^\top Q_{s,+}^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_{s,+} = \sum_{i \in I_s} K(X_i/h) \widetilde{X}_i, \widetilde{X}_i^\top \mathbf{1}\{X_i \geq 0\}, \\
w_{i,s,-}^{(j)}(h) &= e_{j+1}^\top Q_{s,-}^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_{s,-} = \sum_{i \in I_s} K(X_i/h) \widetilde{X}_i \widetilde{X}_i^\top \mathbf{1}\{X_i < 0\},
\end{aligned}$$

with $\widetilde{X}_i = (1, X_i)^\top$. We omit the index s if the sum is taken over the whole sample and we omit the superscript (j) if $j = 0$.

Further, for $\mu \in \mathcal{M}_n$, we let

$$\begin{aligned}
T_{s,+}(\mu) &= \sum_{i \in I_s} K(X_i/h) \widetilde{X}_i \mu(Z_i) \mathbf{1}\{X_i \geq 0\} \\
T_{s,-}(\mu) &= \sum_{i \in I_s} K(X_i/h) \widetilde{X}_i \mu(Z_i) \mathbf{1}\{X_i < 0\}.
\end{aligned}$$

Let $m(x;\mu) = \mathbb{E}[\bar{\mu}(Z_i) - \mu(Z_i)|X_i = x]$. We define $\beta_0(\mu) = m(0;\mu)$, $\beta_1^+(\mu) = \partial_x m(x;\mu)|_{x=0^+}$, and $\beta_1^-(\mu) = \partial_x m(x;\mu)|_{x=0^-}$, and further $\beta^+(\mu) = (\beta_0(\mu), \beta_1^+(\mu))$ and $\beta^-(\mu) = (\beta_0(\mu), \beta_1^-(\mu))$. Let $H = \text{diag}(1,h)$ and $\mathbb{I}_2 = \text{diag}(1,1)$.

3.C.2. **Proof of Theorem 3.1.** The proof of Theorem 3.1 is preceded by two lemmas.

Lemma A.3.1. *Suppose that Assumption 3.2 holds. Then for all $s \in [S]$ it holds that:*

(i) For all $j \in \mathbb{N}$,

$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j T_i = \bar{\nu}_j f_X(0^+) + o_P(1),$$
$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j (1-T_i) = \bar{\nu}_j f_X(0^-) + o_P(1),$$
$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j T_i = \frac{S}{nh}\sum_{i\in I_s} K(X_i/h)(X_i/h)^j T_i + O_P((nh)^{-1/2}),$$
$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j (1-T_i) = \frac{S}{nh}\sum_{i\in I_s} K(X_i/h)(X_i/h)^j (1-T_i) + O_P((nh)^{-1/2}).$$

(ii) For $j \in \{0,1\}$, $h^{2j}\sum_{i\in I_s} w_{i,s}^{(j)}(h)^2 = O_P((nh)^{-1})$ and $h^j \sum_{i\in I_s} |w_{i,s}^{(j)}(h) X_i^2| = O_P(h^2)$.

Proof. Standard kernel calculations. □

Lemma A.3.2. *Suppose that Assumptions 3.1–3.4 hold. Then*

$$G_{s,\star}^{(j)} \equiv e_{j+1}^\top H(Q_{s,\star}^{-1} T_{s,\star}(\bar{\mu}_n - \widehat{\mu}_{n,s}) - \beta^\star(\bar{\mu}_n - \widehat{\mu}_{n,s})) = o_p(h^2 + (nh)^{-1/2})$$

for all $s \in [S]$, $\star \in \{+,-\}$, and $j \in \{0,1\}$.

Proof. We analyze the expectation and variance of $G_{s,\star}^{(j)}$ conditional on $\mathcal{X}_n$ and $(W_j)_{j\in I_s^c}$.

First, we consider the expectation. It holds with probability approaching one that

$$\begin{aligned}|\mathbb{E}[G_{s,\star}^{(j)}|\mathcal{X}_n, (W_j)_{j\in I_s^c}]| &= \left|\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\mathbb{E}[\bar{\mu}_n(Z_i) - \widehat{\mu}_{n,s}(Z_i)|X_i, (W_j)_{j\in I_s^c}]\right| \\ &\leq \sup_{\mu\in\mathcal{T}_n}\left|\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\mathbb{E}[\bar{\mu}_n(Z_i) - \mu(Z_i)|X_i]\right|\end{aligned}$$

By Taylor's theorem with the mean-value form of the remainder, it holds that

$$m(X_i;\mu) = m(0;\mu) + \partial_x m(x;\mu)|_{x=0^\star} X_i + \frac{1}{2}\partial_x^2 m(\widetilde{x}_i;\mu)X_i^2,$$

for some $\widetilde{x}_i$ between 0 and X_i. Using standard local linear algebra and the triangle inequality, we obtain that

$$\begin{aligned}|\mathbb{E}[G_{s,\star}^{(j)}|\mathcal{X}_n,(W_j)_{j\in I_s^c}]| &\leq \sup_{\mu\in\mathcal{T}_n}\left|\frac{1}{2}\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\partial_x^2 m(\widetilde{x}_i;\mu)X_i^2\right| \\ &\leq \sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h\setminus\{0\}}\frac{1}{2}|\partial_x^2 m(x;\mu)|\sum_{i\in I_s}\left|w_{i,s,\star}^{(j)}(h)X_i^2\right| = o_p(h^2),\end{aligned}$$

where we use Lemma A.3.1 and Assumption 3.4 in the last step.

Second, we consider the conditional variance. It holds with probability approaching one that

$$\begin{aligned}\mathbb{V}\left[G_{s,\star}^{(j)}|\mathcal{X}_n,(W_j)_{j\in I_s^c}\right] &= \sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2\mathbb{V}\left[\bar{\mu}_n(Z_i)-\widehat{\mu}_{n,s}(Z_i)|\mathcal{X}_n,(W_j)_{j\in I_s^c}\right] \\ &\leq \sup_{\mu\in\mathcal{T}_n}\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2\mathbb{E}[(\bar{\mu}_n(Z_i)-\mu(Z_i))^2|X_i] \\ &\leq \sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}[(\bar{\mu}_n(Z_i)-\mu(Z_i))^2|X_i=x]\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2 \\ &= o_p((nh)^{-1}).\end{aligned}$$

where we use Lemma A.3.1 and Assumption 3.3 in the last step. The conditional convergence implies the unconditional one (see Chernozhukov et al., 2018, Lemma 6.1), which concludes the proof. □

Proof of Theorem 3.1. We prove the three parts separately.
Part (i) It holds that:

$$\begin{aligned}&\widehat{\tau}_{CF}(h;\widehat{\mu}_n)-\widehat{\tau}(h;\bar{\mu}_n) \\ &= e_1^\top\sum_{s=1}^S\left\{Q_+^{-1}T_{s,+}(\bar{\mu}_n-\widehat{\mu}_{n,s})-Q_-^{-1}T_{s,-}(\bar{\mu}_n-\widehat{\mu}_{n,s})\right\} \\ &= e_1^\top\sum_{s=1}^S Q_+^{-1}Q_{s,+}(Q_{s,+}^{-1}T_{s,+}(\bar{\mu}_n-\widehat{\mu}_{n,s})-\beta^+(\bar{\mu}_n-\widehat{\mu}_{n,s}))+e_1^\top\sum_{s=1}^S Q_+^{-1}Q_{s,+}\,\beta^+(\bar{\mu}_n-\widehat{\mu}_{n,s}) \\ &- e_1^\top\sum_{s=1}^S Q_-^{-1}Q_{s,-}(Q_{-,s}^{-1}T_{s,-}(\bar{\mu}_n-\widehat{\mu}_{n,s})-\beta^-(\bar{\mu}_n-\widehat{\mu}_{n,s}))-e_1^\top\sum_{s=1}^S Q_-^{-1}Q_{s,-}\,\beta^-(\bar{\mu}_n-\widehat{\mu}_{n,s}). \\ &\equiv A_1+A_2-A_3-A_4.\end{aligned}$$

In the following, we consider each of the four terms separately. First, note that

$$A_1 = e_1^\top H^{-1} \sum_{s=1}^{S} H Q_+^{-1} H H^{-1} Q_{s,+} H^{-1} H (Q_{s,+}^{-1} T_{s,+}(\bar{\mu}_n - \widehat{\mu}_{n,s}) - \beta^+(\bar{\mu}_n - \widehat{\mu}_{n,s}))$$

By Lemma A.3.1, for all $s \in [S]$, it holds that

$$H Q_+^{-1} H H^{-1} Q_{s,+} H^{-1} = \frac{1}{S}\mathbb{I}_2 + O_P((nh)^{-1/2}), \tag{A.3.18}$$

where throughout the proof we assume that the term $O_P((nh)^{-1/2})$ has conformable dimensions. Using Lemma A.3.2 and noting that $e_1^\top H^{-1} = e_1^\top$, we obtain that $A_1 = o_p(h^2 + (nh)^{-1/2})$.

Second, it holds that

$$A_2 = e_1^\top H^{-1} \sum_{s=1}^{S} H Q_+^{-1} H H^{-1} Q_{s,+} H^{-1} H\, \beta^+(\bar{\mu}_n - \widehat{\mu}_{n,s}).$$

Using equation (A.3.18), we obtain that

$$\begin{aligned} A_2 &= \frac{1}{S}\sum_{s=1}^{S}(e_1^\top + O_p((nh)^{-1/2}))H\, \beta^+(\bar{\mu}_n - \widehat{\mu}_{n,s}) \\ &= \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s})(1 + O_p((nh)^{-1/2})) + h\beta_1^+(\bar{\mu}_n - \widehat{\mu}_{n,s})O_p((nh)^{-1/2}) \\ &= \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s}) + o_p((nh)^{-1/2}), \end{aligned}$$

where we use the fact $\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s}) = o_p(1)$ by Assumption 3.3 and $h\beta_1^+(\bar{\mu}_n - \widehat{\mu}_{n,s}) = o_p(1)$ by Assumption 3.4 for all $s \in [S]$.

Using analogous calculations, we can show that $A_3 = o_P(h^2 + (nh)^{-1/2})$ and $A_4 = \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu}_n - \widehat{\mu}_{n,s}) + o_P((nh)^{-1/2})$, which concludes the proof of part (i).

Part (ii). By the conditional version of Lyapunov CLT, we obtain that

$$\mathrm{se}(h;\bar{\mu}_n)^{-1}(\widehat{\tau}(h;\bar{\mu}_n) - \mathbb{E}[\widehat{\tau}(h;\bar{\mu}_n)|\mathcal{X}_n]) \to \mathcal{N}(0,1).$$

where $\mathrm{se}^2(h;\bar{\mu}_n) = \sum_{i=1}^{n} w_i(h)^2 \mathbb{V}[M_i(\bar{\mu}_n)|X_i = X_i]$. Further, using L-Lipschitz continuity

of $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$, we obtain that

$$\begin{aligned}&\mathrm{se}^2(h;\bar{\mu}_n)\\&\quad=\sum_{i=1}^{n} w_{i,-}(h)^2\mathbb{V}[M_i(\bar{\mu}_n)|X_i=0^-]+\sum_{i=1}^{n} w_{i,+}(h)^2\mathbb{V}[M_i(\bar{\mu}_n)|X_i=0^+]+o_p((nh)^{-1/2}).\end{aligned}$$

It then follows from standard local linear arguments, that $nh\,\mathrm{se}^2(h;\bar{\mu}_n) - V(\bar{\mu}_n) = o_P(1)$ and $\mathbb{E}[\widehat{\tau}(h;\bar{\mu}_n)|\mathcal{X}_n] - \tau = B(\bar{\mu}_n)h^2 + o_p(h^2)$.

Part (iii). The proof is discussed in Section 3.3.2. It also follows from Proposition A.3.5. □

3.C.3. **Proof of Corollary 3.1.** This result follows directly from linearity of the second derivative operator.

3.C.4. **Definition of Standard Error.** We first introduce the notation. Let $\mu \in \mathcal{M}_n$. We denote the standard error by $\widehat{\mathrm{se}}^2(h;\mu) = \sum_{i=1}^n w_i^2(h)\widehat{\sigma}_i^2(\mu)$, where

$$\widehat{\sigma}_i^2(\mu) = \frac{1}{1+H_i}\left(M_i(\mu) - \sum_{j\in\mathcal{R}_i} v_{j,i}M_j(\mu)\right)^2,$$
$$v_{j,i} = \widetilde{X}_i\left(\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top\widetilde{X}_j\right)^{-1}\widetilde{X}_j^\top, \quad H_i = \widetilde{X}_i\left(\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top\widetilde{X}_j\right)^{-1}\widetilde{X_i}$$

Here $\widetilde{X}_i = (1, X_i)$ and $\mathcal{R}_i$ is the set of the R nearest neighbors of unit i based on the running variable and within the same fold and on the same side of the cutoff as unit i. We note that by basic OLS algebra, the weights $v_{j,i}$ satisfy: $\sum_{j\in\mathcal{R}_i} v_{j,i} = 1$, $\sum_{j\in\mathcal{R}_i} v_{j,i}(X_j - X_i) = 0$, and $\sum_{j\in\mathcal{R}_i} v_{j,i}^2 = H_i$.

We further let $\widehat{\mathrm{se}}_s^2(h;\mu) = \sum_{i\in I_s} w_i^2(h)\widehat{\sigma}_i^2(\mu)$, so that $\widehat{\mathrm{se}}^2(h;\mu) = \sum_{s=1}^S \widehat{\mathrm{se}}_s^2(h;\mu)$. Similarly, we define $\mathrm{se}_s^2(h;\mu) = \sum_{i\in I_s} w_i^2(h)\sigma_i^2(\mu)$ and $\mathrm{se}^2(h;\mu) = \sum_{s=1}^S \mathrm{se}_s^2(h;\mu)$.

3.C.5. **Proof of Proposition 3.1.** Using the triangular inequality, we first note that

$$\begin{aligned}|nh\,\widehat{\mathrm{se}}_{CF}^2(h;\widehat{\mu}_n) - V(\bar{\mu}_n)| &\le nh|\widehat{\mathrm{se}}_{CF}^2(h;\widehat{\mu}_n) - \mathrm{se}^2(h;\bar{\mu}_n)| + |nh\,\mathrm{se}^2(h;\bar{\mu}_n) - V(\bar{\mu}_n)|\\&\le S\max_{s\in[S]} nh|\widehat{\mathrm{se}}_s^2(h;\widehat{\mu}_{n,s}) - \mathrm{se}_s^2(h;\bar{\mu}_n)| + o_p(1),\end{aligned}$$

where the second inequality follows from the proof of Theorem 3.1. The main step in this proof is to show that for any $s \in [S]$ and conditional on $\mathcal{X}_n$ and $(W_j)_{j\in I_s^c}$, it holds that

$$nh|\widehat{\mathrm{se}}_s^2(h;\widehat{\mu}_{n,s}) - \mathrm{se}_s^2(h;\bar{\mu}_n)| = o_P(1). \tag{A.3.19}$$

We remark that the condition in (A.3.19) would essentially follow from the results of Noack and Rothe (2021) if $\mathbb{V}[M_i(\mu)|X_i = x]$ was L-Lipschitz continuous for all $\mu \in \mathcal{T}_n$. Our setting is different as we impose L-Lipschitz continuity only for the function $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$. Still, some steps of our proof follow from the proof of Theorem 4 of Noack and Rothe (2021). We note that

$$\begin{aligned}
&\widehat{\text{se}}_s^2(h;\widehat{\mu}_{n,s}) - \text{se}_s^2(h;\bar{\mu}_n) \\
&= (\mathbb{E}[\widehat{\text{se}}_s^2(h;\bar{\mu}_n)|\mathcal{X}_n] - \text{se}_s^2(h;\bar{\mu}_n)) + (\widehat{\text{se}}_s^2(h;\widehat{\mu}_{n,s}) - \mathbb{E}[\widehat{\text{se}}_s^2(h;\widehat{\mu}_{n,s})|\mathcal{X}_n, (W_j)_{j\in I_s^c}]) \\
&\quad + (\mathbb{E}[\widehat{\text{se}}_s^2(h;\widehat{\mu}_{n,s}) - \widehat{\text{se}}_s^2(h;\bar{\mu}_n)|\mathcal{X}_n, (W_j)_{j\in I_s^c}]) \\
&\equiv G_1 + G_2 + G_3.
\end{aligned}$$

In the following, we show that each of the three terms is of order $o_P((nh)^{-1})$. First, it follows from the proof of Theorem 4 of Noack and Rothe (2021) that $G_1 = o_P((nh)^{-1})$ as $\mathbb{V}[M_i(\bar{\mu}_n)|X_i = x]$ is L-Lipschitz continuous by Assumption 3.5.

Second, it is clear that $\mathbb{E}[G_2|\mathcal{X}_n, (W_j)_{j\in I_s^c}] = 0$. Further, it follows that with probability approaching one,

$$\mathbb{E}[G_2^2|\mathcal{X}_n, (W_j)_{j\in I_s^c}] \leq \sup_{\mu\in\mathcal{T}_n} \mathbb{E}\left[\left(\widehat{\text{se}}_s^2(h;\mu) - \mathbb{E}[\widehat{\text{se}}_s^2(h;\mu)|\mathcal{X}_n]\right)^2\right] = o_p((nh)^{-2}),$$

where the last equality follows from the proof of Theorem 4 of Noack and Rothe (2021) using boundedness of the fourth conditional moment assumed in the proposition.

We now consider G_3. We note that with probability approaching one

$$\begin{aligned}
|G_3| &= |\sum_{i\in I_s} w_i^2(h)\mathbb{E}[\widehat{\sigma}_i^2(\widehat{\mu}_{n,s}) - \widehat{\sigma}_i^2(\bar{\mu}_n)|\mathcal{X}_n, (W_j)_{j\in I_s^c}]| \\
&\leq \sup_{j\in I_s:\, X_j\in\mathcal{X}_h} \sup_{\mu\in\mathcal{T}_n} \left|\mathbb{E}[\widehat{\sigma}_j^2(\mu) - \widehat{\sigma}_j^2(\bar{\mu}_n)|\mathcal{X}_n]\right| \sum_{i\in I_s} w_i(h)^2.
\end{aligned}$$

Following Noack and Rothe (2021), we note that for any $\mu \in \mathcal{T}_n$ and any $i \in I_s$

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma}_i(\mu)|\mathcal{X}_n] = \sigma_i^2(\mu) &+ \frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i} v_{j,i}^2(\sigma_j^2(\mu) - \sigma_i^2(\mu))\right) \\
&+ \frac{1}{1+H_i}\left(\mathbb{E}[M_i(\mu)|X_i] - \sum_{j\in\mathcal{R}_i} v_{j,i}\mathbb{E}[M_j(\mu)|X_j]\right)^2.
\end{aligned} \tag{A.3.20}$$

In the following, we denote by C a positive constant, which might be different from line to line. By a second-order Taylor-expansion and by a simple OLS-algebra, it holds for

the last term in the above expression that

$$\sup_{i\in I_s:\, X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}\frac{1}{1+H_i}\left(\mathbb{E}[M_i(\mu)|X_i]-\sum_{j\in\mathcal{R}_i}v_{j,i}\mathbb{E}[M_j(\mu)|X_j]\right)^2 \tag{A.3.21}$$
$$\leq C\sup_{i\in I_s:\, X_i\in\mathcal{X}_h}\sup_{j\in\mathcal{R}_i}|X_i-X_j|^4\sup_{x\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}(\partial_x^2\mathbb{E}[M_i(\mu)|X_i=x])^2=o_p(1),$$

where we used that $\frac{1}{1+H_i}\sum_{j\in\mathcal{R}_i}v_{j,i}^2\leq 1$ and $\sup_{x\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}\partial_x^2\mathbb{E}[M_i(\mu)|X_i=x]=O(1)$ by Assumptions 3.4 and 3.5.

Using (A.3.20) and (A.3.21), we obtain that

$$\begin{aligned}
&\sup_{i\in I_s:\, X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}|\mathbb{E}[\widehat{\sigma}_i^2(\mu)-\widehat{\sigma}_i^2(\bar{\mu}_n)|\mathcal{X}_n]|\\
&\leq\sup_{i\in I_s:\, X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}\left|\sigma_i^2(\mu)-\sigma_i^2(\bar{\mu}_n)+\frac{1}{1+H_i}\left(\sum_{j\in\mathcal{R}_i}v_{j,i}^2(\sigma_j^2(\mu)-\sigma_j^2(\bar{\mu}_n)+\sigma_i^2(\bar{\mu}_n)-\sigma_i^2(\mu))\right)\right|\\
&\quad+o_p(1)\\
&\leq C\sup_{i\in I_s:\, X_i\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}|\sigma_i^2(\mu)-\sigma_i^2(\bar{\mu}_n)|+o_p(1)\\
&\leq C\sup_{x\in\mathcal{X}_h}\sup_{\mu\in\mathcal{T}_n}|\mathbb{V}[M_i(\mu)|X_i=x]-\mathbb{V}[M_i(\bar{\mu}_n)|X_i=x]|+o_p(1)\\
&=o_p(1),
\end{aligned}$$

where we used that $\frac{1}{1+H_i}\sum_{j\in\mathcal{R}_i}v_{j,i}^2\leq 1$ and Assumption 3.3. Since $\sum_{i\in I_s}w_i(h)^2=O_p((nh)^{-1})$, we conclude that $G_3=o_P((nh)^{-1})$.

3.C.6. **Proofs for sufficient conditions in Section 3.6.**

Proof of Proposition 3.2. We start by showing that Assumption 3.3 holds. It follows from basic OLS algebra that there exists $\bar{\beta}_Z$ such that for all $s\in[S]$ it holds that $\|\widehat{\beta}_{s,Z}-\bar{\beta}_Z\|_\infty=O_P((nh_1)^{-1/2})$. This implies that $\widehat{\beta}_{s,Z}\in[\bar{\beta}_{s,Z}\pm(nh_1)^{-1/2}v_n]$ w.p.a. 1. Let $v_n\to\infty$ be a sequence s.t. $(nh_1)^{-1/2}v_n\to 0$. We define

$$\mathcal{T}_n=\{\mu:\mu(z)=\beta^\top z,\text{ where }\beta\in\mathcal{B}_n=[\bar{\beta}_Z\pm(nh_1)^{-1/2}v_n]\}.$$

By construction, $\bar{\mu}\in\mathcal{T}_n$ and $\mathbb{P}[\widehat{\mu}_{n,s}\in\mathcal{T}_n]=1+o(1)$ for all $s\in[S]$. Assumption 3.3 follows by noting that

$$\sup_{\beta\in\mathcal{B}_n}\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[(\beta^\top Z_i-\bar{\beta}_Z^\top Z_i)^2|X_i=x\right]\leq d\sup_{\beta\in\mathcal{B}_n}\|\beta-\bar{\beta}_Z\|_\infty^2\sup_{x\in\mathcal{X}_h}\mathbb{E}\left[Z_i^\top Z_i|X_i=x\right]=o(1).$$

We now consider Assumption 3.4. For $j\in\{1,2\}$, all $\beta\in\mathcal{B}_n$ and $x\in\mathcal{X}\setminus\{0\}$, we

have that

$$\partial_x^j \mathbb{E}\left[\beta^\top Z_i - \bar{\beta}_Z^\top Z_i | X_i = x\right] = (\beta_Z - \bar{\beta}_Z)^\top \partial_x^j \mathbb{E}\left[Z_i | X_i = x\right],$$

which concludes this proof. □

Proof of Proposition 3.3. We start by showing that Assumption 3.3 holds. Let v_n be a sequence such that $v_n \to \infty$ and $r_n v_n \to 0$. We define

$$\mathcal{T}_n = \{\mu : \mu(z) = m_\beta(z), \text{ where } \beta \in \mathcal{B}_n = [\bar{\beta} \pm r_n v_n]\}.$$

By construction, $\bar{\mu} \in \mathcal{T}_n$ and $\mathbb{P}[\widehat{\mu}_{n,s} \in \mathcal{T}_n] = 1 + o(1)$ for all $s \in [S]$. Assumption 3.3 follows by noting that

$$\sup_{\beta \in \mathcal{B}_n} \sup_{x \in \mathcal{X}_h} \mathbb{E}[(m_\beta(Z_i) - m_{\bar{\beta}}(Z_i))^2 | X_i = x] \leq \sup_{\beta \in \mathcal{B}_n} \|\beta - \bar{\beta}\|_\infty^2 G^2 = o(1).$$

We now consider Assumption 3.4. Under the assumptions made, for $j \in \{1, 2\}$, all $\beta \in \mathcal{B}_n$, and $x \in \mathcal{X} \setminus \{0\}$, we have that

$$\partial_x^j \mathbb{E}[m_\beta(Z_i) - m_{\bar{\beta}}(Z_i) | X_i = x] = \int (m_\beta(z) - m_{\bar{\beta}}(z)) \partial_x^j f_{Z|X}(z|x) dz.$$

It then follows that for $j \in \{1, 2\}$

$$\begin{aligned} \sup_{\beta \in \mathcal{B}_n} \sup_{x \in \mathcal{X}_h \setminus \{0\}} & \left|\partial_x^j \mathbb{E}[m_\beta(Z_i) - m_{\bar{\beta}}(Z_i) | X_i = x]\right| \\ & \leq G \sup_{\beta \in \mathcal{B}_n} \|\beta - \bar{\beta}\|_\infty \int_{\mathcal{Z}} |\partial_x^j f_{Z|X}(z|x)| dz = o_P(1), \end{aligned}$$

which concludes the proof. □

For completeness, we restate the classic assumptions for uniform convergence of the local linear estimator used by Masry (1996).

Assumption A.3.8. *(i) (X_i, Z_i) are continuously distributed, and $\mathcal{X}$ and $\mathcal{Z}$ are compact and convex; (ii) The joint density $f(x, z)$ is bounded, has bounded first-order derivatives, and is bounded away from zero for all $(x, z) \in \mathcal{X} \times \mathcal{Z}$; (iii) $\mathbb{E}[Y_i | X_i = x, Z_i = z]$ is twice continuously differentiable w.r.t. x and z and the second derivatives are Lipschitz continuous; (iv) $\sup_{x,z} \mathbb{E}[|Y_i|^{2+\delta} | X_i = x, Z_i = z] < \infty$ for some constant $\delta > 0$; (v) For $j \in \{0, ..., 3\}$, $H_j(u) \equiv u^j K(u)$ is Lipschitz continuous;*

Proof of Proposition 3.4. By Theorem 6 of Masry (1996), $\sup_{z \in \mathcal{Z}_h} \|\widehat{\mu}_n(z) - \mu_n(z)\| = O_P(r_n)$, where $r_n = o(1)$. Hence, the set $\mathcal{T}_n$ can be chosen s.t. $\sup_{\mu \in \mathcal{T}_n} \|\mu(Z_i) - \mu_n(Z_i)\|_\infty = o(1)$. Assumption 3.3 follows trivially. Assumption 3.4 is also satisfied, as discussed in Section 3.A.2. □

3.D. VARIANCE CALCULATIONS

In this section, we provide formal derivations for the optimality result discussed in Section 3.5.3 and for the discussion of variance reductions in comparison to the standard RD estimator discussed in Section 3.4. Recall that

$$\widetilde{V}(\mu) = \omega_+ \mathbb{V}[M_i(\mu)|X_i = 0^+] + \omega_- \mathbb{V}[M_i(\mu)|X_i = 0^-],$$
$$\mu_n^*(z) = \frac{\omega_-}{\omega_- + \omega_+}\mu_n^-(z) + \frac{\omega_+}{\omega_- + \omega_+}\mu_n^+(z).$$

We obtain the variance $V(\mu)$ and the function μ_n as a special case when $\omega_+ = \omega_- = 1$.

Proposition A.3.5. *Suppose that Assumptions 3.1–3.5 hold. Then for all $\mu \in \mathcal{M}_n$, it holds that:*

(i) $\widetilde{V}(\mu_n^*) \leq \widetilde{V}(\mu)$ with $\widetilde{V}(\mu_n^*) = \widetilde{V}(\mu)$ if and only if $\mathbb{V}[\mu(Z_i) - \mu_n^*(Z_i)|X_i = 0] = 0$;

(ii) $\widetilde{V}(\mu) < \widetilde{V}(0)$ if and only if $\mathbb{V}[\mu_n(Z_i) - \mu(Z_i)|X_i = 0] < \mathbb{V}[\mu_n(Z_i)|X_i = 0]$.

Proof. Fix μ, $\widetilde{\mu} \in \mathcal{M}_n$. By basic properties of the conditional expectation, we have that

$$\widetilde{V}(\mu) = \omega_+ \mathbb{V}[Y_i - \mu_n^+(Z_i)|X_i = 0^+] + \omega_- \mathbb{V}[Y_i - \mu_n^-(Z_i)|X_i = 0^-] + \widetilde{\mathcal{V}}(\mu),$$

where the first two terms on the right-hand side do not depend on μ, and

$$\widetilde{\mathcal{V}}(\mu) = \omega_+ \mathbb{V}[\mu_n^+(Z_i) - \mu(Z_i)|X_i = 0] + \omega_- \mathbb{V}[\mu_n^-(Z_i) - \mu(Z_i)|X_i = 0].$$

Further, it holds that

$$\begin{aligned}\widetilde{\mathcal{V}}(\mu) = \widetilde{\mathcal{V}}(\mu_n^* + \mu - \mu_n^*) &= \omega_+ \mathbb{V}\left[\frac{\omega_-}{\omega_+ + \omega_-}(\mu_n^+(Z_i) - \mu_n^-(Z_i)) - (\mu(Z_i) - \mu_n^*(Z_i))|X_i = 0\right] \\ &\quad + \omega_- \mathbb{V}\left[\frac{-\omega_+}{\omega_+ + \omega_-}(\mu_n^+(Z_i) - \mu_n^-(Z_i)) - (\mu(Z_i) - \mu_n^*(Z_i))|X_i = 0\right] \\ &= \widetilde{V}(\mu_n^*) + (\omega_+ + \omega_-)\mathbb{V}[\mu(Z_i) - \mu_n^*(Z_i)|X_i = 0].\end{aligned}$$

Hence, $\widetilde{V}(\mu) < \widetilde{V}(\widetilde{\mu})$ if and only if $\mathbb{V}[\mu(Z_i) - \mu_n^*(Z_i)|X_i = 0] < \mathbb{V}[\widetilde{\mu}(Z_i) - \mu_n^*(Z_i)|X_i = 0]$, and similarly with equalities instead of inequalities. Both parts of the lemma follow. □

3.E. ADDITIONAL SIMULATION RESULTS

In this section, we present further simulation results. Table A.3.1 extends the results in Table 3.1. Apart from the bias-aware approach discussed in the main text, we consider bandwidth choices and confidence intervals based on robust bias corrections and

undersmoothing.[20] The qualitative conclusions about the relative performance of different first-stage estimators in different models remain the same as discussed in the main text.

The simulated mean bandwidth of robust bias corrections is on average smaller than that of the bias-aware approach, and the confidence intervals are larger. This feature is known in the nonparametric literature. In the last two rows of Table A.3.1 we report the results using the procedure of Calonico et al. (2019). In this simulation setting, they are essentially the same as the results for our procedure with a linear adjustment function.

In Table A.3.2, we report simulation results for Model 3 for different values of the signal-to-noise ratio. This illustrates that the potential gains from covariate adjustments are large if the covariates explain a large portion of variation in the outcome variable.

[20]The bandwidth for undersmoothing is chosen as $n^{-1/20}$ times the MSE-optimal bandwidth estimated using the `rdrobust` package.

Table A.3.1: Full simulation results for different numbers of relevant covariates

		Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
		Model 1: L=0					**Model 2: L=4**					**Model 3: L=10**					**Model 4: L=25**				
Standard	BA	97.0	-1.4	7.4	32.4	43.2	96.1	-7.1	18.6	81.8	68.8	96.4	-9.5	19.1	87.6	79.3	95.9	-6.3	18.5	81.0	68.5
	RBC	94.8	1.5	11.0	41.5	29.9	94.7	0.0	35.1	130.9	30.5	94.6	1.1	37.1	140.0	26.9	94.2	1.6	39.3	145.7	24.5
	US	94.9	0.6	11.3	42.5	20.5	94.5	-1.1	36.0	133.4	20.9	94.7	0.1	38.0	142.4	18.4	94.3	0.8	40.5	148.4	16.7
Optimal Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2	96.5	-1.3	7.6	32.5	43.2	96.9	-1.3	7.4	32.4	43.2
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.6	1.5	11.3	41.5	29.9	94.2	1.5	11.0	41.4	30.0
	US	94.9	0.6	11.3	42.5	20.5	94.5	0.3	11.3	42.5	20.4	93.7	0.5	11.6	42.6	20.4	94.4	0.5	11.3	42.5	20.5
Linear Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.5	32.5	43.2	96.7	-4.8	12.7	56.2	61.8	96.8	-4.3	10.3	47.2	59.0
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.7	1.3	23.4	85.9	26.3	94.6	0.7	19.9	75.4	19.7
	US	94.9	0.6	11.3	42.5	20.5	94.5	0.3	11.3	42.5	20.4	94.1	0.3	23.9	87.6	18.0	94.2	0.2	20.5	76.6	13.4
Linear	BA	97.0	-1.4	7.4	32.7	43.3	96.7	-1.5	7.5	32.6	43.3	95.9	-4.0	13.7	59.1	59.7	96.5	-4.3	10.8	49.2	58.8
	RBC	94.8	1.5	11.0	41.8	30.0	94.3	1.4	11.1	41.8	29.9	94.0	1.6	25.0	91.8	27.9	94.3	0.7	21.6	81.0	20.3
	US	95.1	0.6	11.3	42.9	20.5	94.6	0.3	11.4	42.8	20.5	94.2	0.6	25.6	93.7	19.1	94.4	0.2	22.2	82.4	13.9
Local Linear	BA	97.0	-1.4	7.4	32.7	43.3	96.8	-1.4	7.5	32.7	43.3	96.3	-1.6	8.3	35.6	45.2	96.8	-1.6	8.2	35.9	45.6
	RBC	94.5	1.5	11.1	41.9	30.0	94.5	1.4	11.1	41.9	29.9	94.3	1.4	12.7	47.1	29.3	94.2	1.5	13.0	49.0	27.8
	US	94.9	0.6	11.4	42.9	20.5	94.7	0.4	11.4	43.0	20.5	94.3	0.5	13.1	48.2	20.0	94.3	0.5	13.5	50.1	19.0
Lasso	BA	96.7	-1.4	7.6	33.1	43.6	96.6	-2.1	8.8	38.3	46.6	96.2	-2.0	9.2	39.1	46.7	96.8	-1.4	7.7	34.0	44.3
	RBC	94.4	1.5	11.6	43.5	28.8	95.0	1.2	13.8	52.1	29.1	93.9	1.3	14.6	53.0	29.5	94.3	1.1	13.2	49.0	24.3
	US	95.1	0.7	11.8	44.5	19.7	94.7	0.2	14.2	53.2	19.9	94.1	0.4	15.0	54.2	20.2	94.2	0.5	13.5	50.0	16.6
Forest	BA	96.8	-1.5	7.6	33.1	43.6	96.7	-2.1	8.7	37.9	46.5	96.6	-1.9	8.5	37.2	46.9	97.1	-2.2	9.3	41.3	49.0
	RBC	94.6	1.5	11.3	42.5	29.9	94.9	1.0	13.4	50.7	29.7	94.0	1.0	15.2	56.0	23.3	94.0	0.8	18.6	68.8	19.8
	US	94.6	0.6	11.6	43.6	20.5	94.8	0.0	13.8	51.8	20.3	94.3	0.3	15.5	57.0	15.9	94.3	0.4	19.1	70.1	13.6
CCFT	RBC	94.5	1.4	11.0	41.3	29.7	93.9	1.3	11.1	41.3	29.7	93.4	1.3	23.5	85.1	26.3	94.3	0.7	20.1	74.6	19.6
	US	94.4	0.6	11.4	42.2	20.3	94.0	0.3	11.4	42.2	20.3	93.4	0.3	24.1	86.3	18.0	93.5	0.2	20.8	75.2	13.4

Notes: Results based on 5000 Monte Carlo draws based on Model 3 explained in the main text. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h). Bandwidth and confidence intervals are constructed based on the bias-aware approach (BA), robust bias correction (RBC), and undersmoothing (US).

Table A.3.2: Simulation results for different signal-to-noise ratios

		Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
		$\rho = 3$					$\rho = 1$					$\rho = 5$					$\rho = 10$				
Standard	BA	96.2	-8.9	19.6	87.5	79.3	96.4	-3.1	9.8	43.3	52.0	96.0	-14.7	29.6	134.7	95.5	95.5	-16.8	58.6	241.5	99.9
	RBC	94.7	1.1	37.2	139.6	26.9	94.1	0.8	16.4	61.2	27.9	94.6	0.9	59.8	226.5	26.7	94.7	-0.1	118.3	446.9	26.7
	US	94.3	0.4	38.2	142.1	18.4	93.8	-0.2	16.9	62.4	19.1	94.5	-0.2	61.4	230.3	18.3	94.8	-0.5	120.5	454.4	18.3
Optimal Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.4	32.5	43.2	96.5	-1.3	7.6	32.4	43.2	96.9	-1.3	7.3	32.5	43.2
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.5	1.5	11.3	41.5	29.8	94.2	1.5	11.0	41.4	30.0
	US	94.9	0.6	11.3	42.5	20.5	94.5	0.3	11.3	42.5	20.4	93.6	0.6	11.6	42.6	20.4	94.5	0.5	11.3	42.5	20.5
Linear Inf	BA	96.0	-4.8	12.9	56.2	61.9	96.2	-1.9	8.3	36.1	46.1	96.4	-8.5	18.1	81.5	76.6	95.9	-14.1	33.0	146.9	96.3
	RBC	94.2	1.2	23.1	85.8	26.4	94.0	1.2	13.1	48.9	28.3	93.8	1.2	35.8	131.6	25.8	94.8	1.2	66.2	252.9	25.6
	US	93.9	0.6	23.8	87.5	18.0	94.4	0.2	13.4	49.9	19.4	94.1	0.1	36.6	134.0	17.7	95.0	0.6	67.2	257.4	17.5
Linear	BA	96.1	-4.0	13.8	59.1	59.8	96.1	-1.9	8.4	36.5	45.9	95.6	-7.2	21.2	90.7	74.0	95.8	-14.1	37.0	161.5	95.4
	RBC	94.0	1.7	24.9	91.9	27.9	94.0	1.2	13.2	49.3	28.6	94.0	1.7	42.2	155.3	28.8	94.6	2.3	83.4	312.8	29.0
	US	93.9	1.0	25.6	93.8	19.1	94.3	0.3	13.5	50.4	19.5	94.2	0.7	43.3	158.4	19.7	94.8	1.0	85.5	318.8	19.8
Local Linear	BA	96.7	-1.7	8.2	35.6	45.1	96.5	-1.5	7.8	33.9	44.1	96.5	-1.7	8.7	37.3	46.3	97.0	-2.6	10.2	45.1	52.3
	RBC	94.4	1.5	12.5	47.1	29.3	94.2	1.3	11.7	43.9	29.7	94.0	1.5	13.7	50.5	28.8	94.6	1.6	17.4	65.1	27.6
	US	94.7	0.7	12.9	48.2	20.1	94.3	0.4	12.0	45.0	20.3	94.0	0.6	14.1	51.7	19.7	94.6	0.7	18.0	66.4	18.9
Lasso	BA	96.8	-2.0	9.1	39.3	46.9	96.8	-1.6	7.7	34.0	44.5	96.1	-2.7	11.5	48.4	51.0	96.2	-4.9	18.1	75.8	61.6
	RBC	93.8	1.5	14.4	53.3	29.6	94.3	1.0	12.4	46.9	26.5	93.9	1.5	18.5	67.7	31.1	94.1	1.9	32.6	117.7	32.2
	US	94.3	0.6	14.9	54.6	20.2	94.4	0.3	12.7	47.9	18.1	94.2	0.6	19.1	69.2	21.3	94.2	0.8	33.3	120.1	22.0
Forest	BA	96.6	-1.9	8.5	37.2	46.9	96.5	-1.6	7.7	33.7	44.3	96.7	-2.6	10.0	43.8	51.1	96.3	-5.8	14.7	64.5	64.5
	RBC	94.1	1.1	15.1	56.1	23.1	94.1	1.1	12.1	45.3	27.7	94.5	0.8	18.9	70.6	21.8	93.9	0.5	31.7	116.1	20.7
	US	94.3	0.6	15.5	57.2	15.8	94.5	0.2	12.4	46.2	19.0	95.0	0.2	19.3	71.8	14.9	94.2	0.1	32.5	118.1	14.2
CCFT	RBC	93.8	1.2	23.3	85.0	26.3	93.5	1.1	13.2	48.6	28.1	93.4	1.2	36.0	130.3	25.8	94.6	1.3	66.2	250.4	25.6
	US	93.3	0.6	24.0	86.3	18.0	93.9	0.2	13.5	49.4	19.2	93.3	0.2	36.9	132.0	17.7	94.7	0.5	67.4	253.4	17.5

Notes: Results based on 5000 Monte Carlo draws based on Model 3 explained in the main text. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h). Bandwidth and confidence intervals are constructed based on the bias-aware approach (BA), robust bias correction (RBC), and undersmoothing (US).

References

Abadie, A. and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for misspecified models with fixed regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Almond, D. and J. Doyle (2011): "After midnight: A regression discontinuity design in length of postpartum hospital stays," *American Economic Journal: Economic Policy*, 3, 1–34.

Anderson, G., O. B. Linton, and Y.-J. Whang (2012): "Nonparametric estimation and inference about the overlap of two distributions," *Journal of Econometrics*, 171, 1 – 23.

Anderson, T. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62, 43–72.

Andrews, D. W. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the sensitivity of parameter estimates to estimation moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

——— (2020a): "On the informativeness of descriptive statistics for structural estimates," *Econometrica*, 88, 2231–2258.

——— (2020b): "Transparency in structural research," *Journal of Business & Economic Statistics*, 38, 711–722.

Andrews, I. and J. M. Shapiro (2020): "A Model of Scientific Communication," *NBER Working Paper*.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. and W. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88, 450–77.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

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

Armstrong, T. B. and M. Kolesár (2018): "Optimal inference in a class of regression models," *Econometrica*, 86, 655–683.

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 11, 1–39.

——— (2021a): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 89, 1141–1177.

——— (2021b): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

Armstrong, T. B., M. Kolesár, and S. Kwon (2020): "Bias-Aware Inference in Regularized Regression Models," *arXiv preprint arXiv:2012.14823*.

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1177.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

Bertanha, M. and M. J. Moreira (2020): "Impossible inference in econometrics: Theory and applications," *Journal of Econometrics*, 218, 247–270.

Bonhomme, S. and M. Weidner (2018): "Minimizing sensitivity to model misspecification," *arXiv preprint arXiv:1807.02161*.

——— (2019): "Posterior average effects," *arXiv preprint arXiv:1906.06360*.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Bugni, F. A. (2010): "Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set," *Econometrica*, 78, 735–753.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the effect of bias estimation on coverage accuracy in nonparametric inference," *Journal of the American Statistical Association*, 113, 767–779.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression Discontinuity Designs Using Covariates," *The Review of Economics and Statistics*, 101, 442–451.

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

Card, D. and L. Giuliano (2016): "Can tracking raise the test scores of high-ability minority students?" *American Economic Review*, 106, 2783–2816.

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

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A practical introduction to regression discontinuity designs: Foundations*, Cambridge University Press.

Chen, X., E. T. Tamer, and A. Torgovitsky (2011): "Sensitivity analysis in semiparametric likelihood models," *Cowles foundation discussion paper*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., I. Férnandez--Val, and A. Galichon (2010): "Quantile and Probability Curves Without Crossing," *Econometrica*, 78, 1093–1125.

Christensen, T. and B. Connault (2019): "Counterfactual sensitivity and robustness," *arXiv preprint arXiv:1904.00989*.

Clark, D. and P. Martorell (2014): "The signaling value of a high school diploma," *Journal of Political Economy*, 122, 282–318.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly exogenous," *Review of Economics and Statistics*, 94, 260–272.

Coviello, D., A. Guglielmo, and G. Spagnolo (2018): "The effect of discretion on procurement performance," *Management Science*, 64, 715–738.

Dahl, C. M., M. Huber, and G. Mellace (2017): "It's never too LATE: A new look at local average treatment effects with or without defiers," *Discussion Papers on Business and Economics, University of Southern Denmark*, 2.

Dahl, G. B., K. V. Løken, and M. Mogstad (2014): "Peer effects in program participation," *American Economic Review*, 104, 2049–74.

De Chaisemartin, C. (2017): "Tolerating defiance? Local average treatment effects without monotonicity," *Quantitative Economics*, 8, 367–396.

Dong, Y. (2018): "Alternative assumptions to identify LATE in fuzzy regression discontinuity designs," *Oxford Bulletin of Economics and Statistics*, 80, 1020–1027.

Donoho, D. L. (1994): "Statistical estimation and optimal recovery," *Annals of Statistics*, 22, 238–270.

Dube, A., L. Giuliano, and J. Leonard (2019): "Fairness and frictions: The impact of unequal raises on quit behavior," *American Economic Review*, 109, 620–63.

Dümbgen, L. (1993): "On nondifferentiable functions and the bootstrap," *Probability Theory and Related Fields*, 95, 125–140.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2020): "Estimation of Conditional Average Treatment Effects With High-Dimensional Data," *Journal of Business & Economic Statistics*, 1–15.

Fang, Z. and A. Santos (2018): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412.

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep neural networks for estimation and inference," *Econometrica*, 89, 181–213.

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

Fieller, E. C. (1954): "Some problems in interval estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, 16, 175–185.

Fiorini, M., K. Stevens, et al. (2014): "Assessing the monotonicity assumption in IV and fuzzy RD designs," *Economics Working Paper Series-University of Sidney*, 13.

Frandsen, B. R., L. J. Lefgren, and E. C. Leslie (2019): "Judging judge fixed effects," Tech. rep., National Bureau of Economic Research.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013): "Long-term effects of class size," *The Quarterly Journal of Economics*, 128, 249–285.

Freyberger, J. and Y. Rai (2018): "Uniform confidence bands: Characterization and optimality," *Journal of Econometrics*, 204, 119–130.

Frölich, M. and M. Huber (2019): "Including Covariates in the Regression Discontinuity Design," *Journal of Business & Economic Statistics*, 37, 736–748.

Giné, E. and A. Guillou (2002): "Rates of strong uniform consistency for multivariate kernel density estimators," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, Elsevier, vol. 38, 907–921.

Giné, E. and R. Nickl (2008): "Uniform central limit theorems for kernel density estimators," *Probability Theory and Related Fields*, 141, 333–387.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Haile, P. A. and E. Tamer (2003): "Inference with an incomplete model of English auctions," *Journal of Political Economy*, 111, 1–51.

Hinnerich, B. T. and P. Pettersson-Lidbom (2014): "Democracy, redistribution, and political participation: Evidence from Sweden 1919–1938," *Econometrica*, 82, 961–993.

Hong, H. and J. Li (2018): " "The numerical delta method"," *Journal of Econometrics*, 206, 379 – 394.

Horowitz, J. L. and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

Huang, X. and Z. Zhan (2020): "Does health behavior change after diagnosis? Evidence from a reliable fuzzy regression discontinuity approach," *Working Paper*.

Huber, M. (2014): "Sensitivity checks for the local average treatment effect," *Economics Letters*, 123, 220 – 223.

——— (2015): "Testing the Validity of the Sibling Sex Ratio Instrument," *Labour*, 29, 1–14.

Huber, M., L. Laffers, and G. Mellace (2017): "Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations Under Endogeneity and Noncompliance," *Journal of Applied Econometrics*, 32, 56–79.

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

Ignatiadis, N. and S. Wager (2020): "Bias-aware confidence intervals for empirical Bayes analysis," *Working Paper*.

Imbens, G. W. (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132.

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

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

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

Imbens, G. W. and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

Imbens, G. W. and S. Wager (2019): "Optimized regression discontinuity designs," *Review of Economics and Statistics*, 101.

Jepsen, C., P. Mueser, and K. Troske (2016): "Labor market returns to the GED using regression discontinuity analysis," *Journal of Political Economy*, 124, 621–649.

Kamat, V. (2018): "On nonparametric inference in the regression discontinuity design," *Econometric Theory*, 34, 694–703.

Kennedy, E. H. (2020): "Optimal doubly robust estimation of heterogeneous causal effects," *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 1229.

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

——— (2021): "The identification region of the potential outcome distributions under instrument independence," *Journal of Econometrics*.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, infinitesimal neighborhoods, and moment restrictions," *Econometrica*, 81, 1185–1201.

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

Kline, P. and A. Santos (2013): "Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure," *Quantitative Economics*, 4, 231–267.

Kolesár, M. and C. Rothe (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 108, 2277—2304.

Le Barbanchon, T., R. Rathelot, and A. Roulet (2019): "Unemployment insurance and reservation wages: Evidence from administrative data," *Journal of Public Economics*, 171, 1–17.

Lee, D. S. (2009): "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," *The Review of Economic Studies*, 76, 1071–1102.

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

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

Li, K.-C. (1989): "Honest confidence regions for nonparametric regression," *Annals of Statistics*, 17, 1001–1008.

Low, M. (1997): "On nonparametric confidence intervals," *Annals of Statistics*, 25, 2547–2554.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212, 522–555.

Malenko, N. and Y. Shen (2016): "The role of proxy advisory firms: Evidence from a regression-discontinuity design," *The Review of Financial Studies*, 29, 3394–3427.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323.

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

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

Masten, M. A. and A. Poirier (2020): "Inference on Breakdown Frontiers," *Quantitative Economics*, 11, 41–111.

——— (2021): "Salvaging falsified instrumental variable models," *Econometrica*, 89, 1449–1469.

Mogstad, M., A. Torgovitsky, and C. Walters (2019): "Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions," *NBER Working Paper*.

Mukhin, Y. (2018): "Sensitivity of regular estimators," *arXiv preprint arXiv:1805.08883*.

Negi, A. and J. M. Wooldridge (2020): "Revisiting regression adjustment in experiments with heterogeneous treatment effects," *Econometric Reviews*, 40, 504–534.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Neyman, J. (1959): "Optimal asymptotic tests of composite hypotheses," *Probability and Statsitics*, 213–234.

——— (1979): "C (α) tests and their use," *Sankhyā: The Indian Journal of Statistics, Series A*, 1–21.

Noack, C. and C. Rothe (2021): "Bias-aware inference in fuzzy regression discontinuity designs," *arXiv preprint arXiv:1906.04631*.

Norris, S., M. Pecenco, and J. Weaver (2020): "The effects of parental and sibling incarceration: Evidence from Ohio," *Available at SSRN 3590735*.

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

——— (2008): "Corrigendum: Estimating average and local average treatment effects of education when compulsory schooling laws really matter," *Internet-Only Corrigendum; American Economic Review*.

Roth, J. and A. Rambachan (2019): "An Honest Approach to Parallel Trends," .

Sacks, J. and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *Annals of Statistics*, 6, 1122–1137.

Schennach, S. M. (2020): "A bias bound approach to non-parametric inference," *The Review of Economic Studies*, 87, 2439–2472.

Scott-Clayton, J. and B. Zafar (2019): "Financial aid, debt management, and socioeconomic outcomes: Post-college effects of merit-based aid," *Journal of Public Economics*, 170, 68–82.

Shapiro, A. (1991): "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, 30, 169–186.

Small, D. S., Z. Tan, R. R. Ramsahai, S. A. Lorch, M. A. Brookhart, et al. (2017): "Instrumental variable estimation with a stochastic monotonicity assumption," *Statistical Science*, 32, 561–579.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 557–586.

Stoye, J. (2005): "Essays on partial identification and satistical decisions," *Ph.D. thesis, Northewstern University*.

——— (2010): "Partial identification of spread parameters," *Quantitative Economics*, 1, 323–357.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Urquiola, M. and E. Verhoogen (2009): "Class-size caps, sorting, and the regression-discontinuity design," *American Economic Review*, 99, 179–215.

van der Vaart, A. (1998a): *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

van der Vaart, A. W. (1998b): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341.

Wager, S. and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016): "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.

Curriculum Vitæ

Claudia Luise Charlotte Noack

Education

2015–2021	University of Mannheim (Germany) *Ph.D. Student*
2019	Yale University (USA) *Visiting Ph.D. Student*
2018	University of Mannheim (Germany) *Master of Science in Economics*
2016–2017	University of California, Berkeley (USA) *Visiting Ph.D. Student*
2012–2015	University of Mannheim (Germany) *Bachelor of Science in Economics*
2014	Toulouse School of Economics (France) *Visiting Undergraduate Student*
2012	Erich Kästner Gymnasium, Laatzen (Germany) *Abitur*

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig angefertigt und die benutzten Hilfsmittel vollständig und deutlich angegeben habe. Hiermit erkläre ich mich damit einverstanden, dass die Universität meine Dissertation zum Zwecke des Plagiatsabgleichs in elektronischer Form speichert, an Dritte versendet, und Dritte die Dissertation zu diesem Zwecke verarbeiten.

Claudia Noack