ESSAYS IN ECONOMETRICS

By

Otávio Augusto Camargo Bartalotti

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Economics

2012

ABSTRACT

ESSAYS IN ECONOMETRICS

By

Otávio Augusto Camargo Bartalotti

This dissertation is divided in three self-contained chapters. The first extends the GMM redundancy results of Prokhorov and Schmidt (2009) for nonsmooth objective functions, giving sharp guidelines about how to obtain efficient estimates of parameters of interest (β_o) in the presence of nuisance parameters (γ_o). The use of one-step GMM estimators for both sets of parameters is asymptotically more efficient than two-step procedures. These results are applied to Wooldridge (2007)'s inverse probability weighted estimator (IPW), generalizing the framework to deal with missing-data in this context. Even though two-step estimation of β_o is more efficient than using known probabilities of selection, this is dominated by a one-step joint estimation procedure. Examples for quantile regression with missing data and instrumental variable quantile regression are provided.

The second chapter analyzes the asymptotic distribution of local polynomial estimators in the context of regression discontinuity designs. The standard "small-h" approach in the literature (Hahn et al., 2001; Porter, 2003; Imbens and Lemieux, 2008; Lee and Lemieux, 2009) is to assume the bandwidth, h, around the discontinuity shrinks towards zero as the sample size increases. However, in practice, the researcher has to choose an $h > 0$ to implement the estimator. This chapter derives the fixed-h asymptotic distribution that allows for the bandwidth to be positive, providing refined approximations for the estimator's behavior.

When $h > 0$, the small-h asymptotic variance is equivalent to assuming that the density of the running variable and the conditional variance of the dependent variable are constant around the cutoff. Simulations provide evidence that fixed-h asymptotic distributions better describe the behavior of both bias and variance of the estimator, leading to improved inference. Estimators for fixed-h standard errors are proposed and incorporate the theoretical gains of the improved approximations. The fixed-h variance estimators improve markedly over small-h estimators in the presence of some forms of heteroskedasticity. Interestingly, in the special case of homoskedastic errors using a local linear estimator, the variance estimators based on small-h asymptotics produce tests with similar size to the fixed-h variance estimators proposed in this chapter.

Chapter 3 develops the asymptotic properties of quantile regression estimators under standard stratification sampling, following Wooldridge (2001). Formulas for the asymptotic variance and feasible estimators are provided. Under exogenous stratification the usual quantile regression estimators and standard errors are still valid.

To Beatriz.

ACKNOWLEDGMENTS

"We have not journeyed all this way across the centuries, across the oceans, across the mountains, across the prairies, because we are made of sugar candy."
Winston Churchill

I would like acknowledge the great help and support that was given to this endeavor by Michigan State University faculty. Especially my advisor, Jeff Wooldridge, who supported me on my goals even before I started the Ph.D. program and kept me upbeat and motivated through the toughest moments, I would not have been able to accomplish this without his guidance, help and friendship. Also, Tim Vogelsang, who was fundamental in all the parts of this dissertation through his insights, truthfulness and great openness to my ideas. I cannot express enough thanks to Gary Solon, who always kept his door open and a great interest in my research, providing insight, guidance and great advice in the five years I spent at MSU's Economics Department. Peter Schmidt, always sharp on comments and an open mind provided formidable advice that greatly improved this dissertation. I am also thankful several other great faculty such as Todd Elder, Steve Haider, Matias Cattaneo (at University of Michigan) and many others. I cannot forget to thank Carlos Pereira for believing in me when few would. Finally, none of this would have been achieved without the support from Maria Carolina Leme who taught me through my first steps in academia and still teaches me so much to this day.

My dissertation and life have been made better by great friends, their inputs, ideas and long coffee breaks like Steve Dieterle, Quentin Brummet, Thomas Fujiwara, Ilya Rahkovsky,

Max Melstrom, Breno Braga, Valentin Verdier, Maksym Ivanyna, Elizabeth Quin, Ehren Schuttringer, Cuicui Lu, Stacy Miller, Iraj Rahmani, Paul Burkander, Hassan Enayati, Laura Rees, Jason Stornelli, Adithya Pattabhiramaiah among others.

I thank my parents Bete and Flávio and my stepmother Hilda for their love and support throughout this journey and for embedding in my formation the desire for knowledge that drove me to this career. Also my sister Laís who helps me keep things on perspective and I love. To my grandparents for the great life example they provide and especially to Antonio Bartalotti who did not live to see this day, but is in my thoughts and heart everyday.

Finally, I acknowledge the courage, help and love I have received from my wife, Beatriz, who shared with me every moment of the path, and was understanding and reassuring of my achievements and failures.

TABLE OF CONTENTS

LIST OF FIGURES . . . . . **ix**

1 GMM Efficiency and IPW for Nonsmooth Functions **1**
1.1 Introduction 1
1.2 General Estimation Problem 3
1.3 Estimation with missing data 15
1.3.1 Data Selection under Ignorability 16
1.3.2 Data Selection under Exogeneity of Selection 21
1.4 Examples 23
1.4.1 Quantile Regression under Ignorability of Selection 23
1.4.2 Instrumental Variable Quantile Regression 27
1.5 Conclusion 28

2 Fixed Bandwidth Asymptotics for Regression Discontinuity Designs **31**
2.1 Introduction 31
2.2 Model 34
2.2.1 Sharp Regression Discontinuity Design 34
2.2.2 Fuzzy Regression Discontinuity Design 35
2.3 Estimators 37
2.4 Assumptions 39
2.5 Asymptotic Distributions 40
2.5.1 Fuzzy Regression Discontinuity Design 45
2.6 Variance Estimators 47
2.7 Simulations 51
2.7.1 Simulations for Infeasible Inference 53
2.7.2 Simulations for Feasible Inference 58
2.8 Conclusion 62

3 Asymptotic Properties of Quantile Regression for Standard Stratified Samples **64**
3.1 Introduction 64
3.2 The Quantile Regression Population Problem 65
3.2.1 Quantile Regression under Stratified Sampling 66

3.2.2 Quantile Regression Estimation under Exogenous Stratification . . . 69
3.2.3 Sequence of Quantile Regressions . . . 71
3.3 Asymptotic Variance Estimation . . . 73
3.4 Conclusion . . . 76

APPENDICES . . . 79

A Proofs to "GMM Efficiency and IPW for Nonsmooth Functions" . . . 79

B Figures to "Fixed Bandwidth Asymptotics for Regression Discontinuity Designs" . . . 88
B.1 Simulations for Infeasible Inference . . . 88
B.1.1 Nadaraya-Watson Estimator . . . 88
B.1.2 Local Linear Estimator . . . 93
B.1.3 Heteroskedastic Errors . . . 97
B.2 Simulations for Feasible Inference . . . 101
B.2.1 Local Linear Estimator . . . 103
B.2.2 Heteroskedastic Errors . . . 105
B.2.3 Bandwidth Choice for $\widehat{f_O}(\overline{x})$ . . . 109

C Proofs to "Fixed Bandwidth Asymptotics for Regression Discontinuity Designs" . . . 111

D Proofs to "Asymptotic Properties of Quantile Regression for Standard Stratified Samples" . . . 123

BIBLIOGRAPHY . . . 130

LIST OF FIGURES

B.1 Nadaraya-Watson Estimator - DGP: No X - Homosk. Errors 89
B.2 Nadaraya-Watson Estimator - DGP: Linear - Homosk. Errors 90
B.3 Nadaraya-Watson Estimator - DGP: Linear - Bias Corrected - Homosk. Errors 91
B.4 Nadaraya-Watson Estimator - DGP: Linear - Comparison - Homosk. Errors 92
B.5 Local Linear Estimator - DGP: Linear - Homosk. Errors 93
B.6 Local Linear Estimator - DGP: Quadratic - Homosk. Errors 94
B.7 Local Linear Estimator - DGP: Quadratic - Bias Corrected - Homosk. Errors 95
B.8 Local Linear Estimator - DGP: Quadratic - Comparison - Homosk. Errors . 96
B.9 Local Linear Estimator - DGP: Linear - Heterosk. Errors Case 1 97
B.10 Local Linear Estimator - DGP: Quadratic - Heterosk. Errors Case 1 98
B.11 Local Linear Estimator - DGP: Linear - Heterosk. Errors Case 2 99
B.12 Local Linear Estimator - DGP: Quadratic - Heterosk. Errors Case 2 100
B.13 Nadaraya-Watson Estimator - DGP: No X - Feasible - Homosk. Errors . . . 101
B.14 Nadaraya-Watson Estimator - DGP: Linear - Feasible - Homosk. Errors . . . 102
B.15 Local Linear Estimator - DGP: Linear - Feasible - Homosk. Errors 103
B.16 Local Linear Estimator - DGP: Quadratic - Feasible - Homosk. Errors 104
B.17 Local Linear Estimator - DGP: Linear - Feasible - Heterosk. Errors Case 1 . 105
B.18 Local Linear Estimator - DGP: Quadratic - Feasible - Heterosk. Errors Case 1 106
B.19 Local Linear Estimator - DGP: Linear - Feasible - Heterosk. Errors Case 2 . 107
B.20 Local Linear Estimator - DGP: Quadratic - Feasible - Heterosk. Errors Case 2 108
B.21 Small-h Sensitivity to Density Bandwidth - DGP: Linear 109

B.22 Small-h Sensitivity to Density Bandwidth - DGP: Quadratic 110

CHAPTER 1

GMM Efficiency and IPW for Nonsmooth Functions

1.1 Introduction

This chapter extends Prokhorov and Schmidt (2009) analysis to the estimation of a general GMM problem with nonsmooth objective functions in which nuisance parameters are present. The framework developed encompasses several interesting problems in econometrics such as missing data, censored or truncated data, treatment effects, instrumental variables, etc. More importantly, by allowing nonsmooth objective functions, the analysis extends to models that have gained additional importance in recent years, e.g., least absolute deviations (LAD), quantile regression (QR), censored LAD, quantile treatment effects and instrumental variables quantile regression (IVQR).

The core results of this chapter extend Prokhorov and Schmidt (2009) results on GMM redundancy by allowing the use of nonsmooth objective functions. These results rely on Newey and McFadden (1994) to obtain the asymptotic variance of the GMM estimator under less restrictive assumptions on the smoothness of the objective functions. For that consider two sets of moment conditions, where the first includes both the parameters of interest (β_o) and

certain nuisance parameters (γ_o) while the second set includes only the nuisance parameters. By defining four competing estimators based on different assumptions regarding the information available about these nuisance parameters and the moment conditions utilized, results about the relative efficiency of each proposed estimator are derived. These results provide guidance to applied work in the presence of nuisance parameters.

As discussed by Prokhorov and Schmidt (2009), joint estimation of nuisance parameters and parameters of interest is more efficient than a two-step procedure or knowing the true nuisance parameters and disregarding the second set of moment conditions. This fact is due to the information contained in correlation between both sets of moment conditions, which is useful even when γ_o is known. Using only the first set of moment conditions and known values of γ_o in the estimation procedure does not use the additional information embedded in the second set of moment conditions. These results are shown to hold when the objective functions are nonsmooth.

The general results are directly applicable to missing data problems and encompass Wooldridge (2002b, 2007) analysis of inverse probability weighting (IPW) estimators, extending its use for nonsmooth objective functions under the usual "ignorability" assumptions about the selection process. The general estimation results described confirm the validity of the puzzle described by Wooldridge (2007), i.e., that it is better (in an efficiency sense) to estimate the selection probabilities, even if the latter are known. In other terms, we obtain more efficient estimates for β_o if we estimate γ_o than if we use the true γ_o. This result is "puzzling" because knowledge of γ_o, if properly exploited, cannot be harmful. Previous works discussed this result, such as Wooldridge (2002b, 2007) in the context of IPW. Hirano et al. (2003); Hitomi et al. (2008); Prokhorov and Schmidt (2009) addressed the problem for the smooth objective function case. Even though this issue has been considered by Chen, Hong, and Tarozzi (2008) in a semiparametric context with nonsmooth objective functions, the parametric approach proposed here provides, as a novelty, the conditions under which this puzzle is valid and, furthermore, shows that the two-step estimator is usually dominated

by a one-step joint estimation procedure that uses both the weighted moment conditions and the conditions associated with the selection model.

There have been several papers devoted to general theories of estimation in settings where nonsmooth objective functions are allowed, following Daniels (1961) and Huber (1967). Studies that allow for estimation of models based on nonsmooth objective functions include, among others, Pollard (1985); Pakes and Pollard (1989) and Newey and McFadden (1994, section 7). Recent studies have approached the problem of nonsmoothness with focus on semiparametric models, see Chen, Linton, and Van Keilegom (2003) for a general estimation approach; Chen et al. (2008) for an approach for missing data problems with nonparametric first stage; and Cattaneo (2010) for an approach on the estimation of multi-valued treatment effects on a semiparametric framework.

The remainder of the chapter is organized as follows. Section 1.2 sets up the general GMM framework used in the analysis and presents results regarding efficiency and redundancy of the estimators proposed, as well as estimators for the asymptotic variances of the parameters estimated. Section 1.3 studies the IPW approach to missing data problems proposed by Wooldridge (2002b, 2007), extending its scope to nonsmooth objective functions. Section 1.4 provides examples of the uses of the framework proposed here by, first, considering a model for the conditional quantile in a context with missing data; secondly I consider a simplified IVQR model as proposed by Chernozhukov and Hansen (2005, 2006). Section 1.5 concludes.

1.2 General Estimation Problem

Let $\boldsymbol{\omega}^* \in Q^* \subset R^{\dim(\omega^*)}$ be a random vector; $\theta \in \boldsymbol{\Theta} \subset \mathbb{R}^P$ be a parameter vector, $\boldsymbol{\Theta}$ is a compact set, and the population condition

$$g_o(\theta_o) = E[g(\boldsymbol{\omega}^*, \theta_o)] = 0 \tag{1.1}$$

where $g : Q^* \times \boldsymbol{\Theta} \to \mathbb{R}^m$ is a vector of known real-valued moment functions.

Newey and McFadden (1994) have shown consistency and asymptotic normality of the Generalized Method of Moments (GMM) estimator that minimizes a squared Euclidean distance of the random sample analogues of the population moments, *i.e.* $g_n(\theta) = n^{-1} \sum_{i=1}^{n} g(\boldsymbol{\omega}_i^*, \theta)$, from their population counterparts (which equal zero). I am interested in the case in which the moment functions, $g(\cdot)$, are allowed to be nonsmooth, so we can deal with a wider range of interesting problems. The GMM estimator minimizes the objective function

$$g_n(\theta)' \widehat{W} g_n(\theta) \tag{1.2}$$

where $\widehat{W}$ converges in probability to W, the appropriate positive semidefinite weighting matrix. Assume $\boldsymbol{\omega}_i^*, i = 1, ..., n$, are i.i.d. Two useful results from Newey and McFadden (1994) will be used to derive the asymptotic variance of the estimators. The first regards the consistency of the GMM estimator.

Theorem 1 (Newey and McFadden, 1994, Theorem 2.6) *Let* $\|\bullet\|$ *denote the Euclidean norm. Suppose that:*

(i) $\boldsymbol{\omega}_i^*$ *are i.i.d. for i=1,2,...;*

(ii) $\widehat{W} \xrightarrow{p} W$*;*

(iii) W *is positive semi-definite and* $W\ E[g(\boldsymbol{\omega}^*, \theta)] = 0$ *only if* $\theta = \theta_o$*;*

(iv) $\theta_o \in \boldsymbol{\Theta} \subset \mathbb{R}^P$*, and* $\boldsymbol{\Theta}$ *is compact;*

(v) $g(\boldsymbol{\omega}^*, \theta)$ *is continuous at each* θ *with probability one;*

(vi) $E\left[\sup_{\theta \in \boldsymbol{\Theta}} \|g(\boldsymbol{\omega}^*, \theta)\|\right] < \infty$*;*

Then $\widehat{\theta} \xrightarrow{p} \theta_o$*.*

This result relies on relatively weak conditions, and allow for discontinuities in the objective function.

The second theorem demonstrates the asymptotic normality of the GMM estimator under a certain form of nonsmoothness of the objective function.

Theorem 2 (Newey and McFadden, 1994, Theorem 7.2) *Suppose that:*

(i) $g_n(\widehat{\theta})'\widehat{W}g_n(\widehat{\theta}) \leq \inf_{\boldsymbol{\theta}\in\Theta} g_n(\theta)'\widehat{W}g_n(\theta) + o_p(n^{-1})$*;*

(ii) $\widehat{\theta} \xrightarrow{p} \boldsymbol{\theta}_o$*;*

(iii) $\widehat{W} \xrightarrow{p} W$*, and* W *is positive semi-definite;*

(iv) $g_o(\boldsymbol{\theta}_o) = 0$*;*

(v) $g_o(\boldsymbol{\theta})$ *is differentiable at* $\boldsymbol{\theta}_o$ *with derivative* G *such that* $G'WG$ *is nonsingular;*

(vi) $\boldsymbol{\theta}_o$ *is an interior point in* $\boldsymbol{\Theta}$*;*

(vii) $\sqrt{n}g_n(\boldsymbol{\theta}_o) \xrightarrow{d} N(\mathbf{0}, \Sigma)$*;*

(viii) for any $\delta_n \xrightarrow{p} 0$*,*

$$\sup_{\|(\boldsymbol{\theta}-\boldsymbol{\theta}_o)\|\leq\delta_n} \frac{\sqrt{n}\,\|g_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta}_o) - g_o(\boldsymbol{\theta})\|}{[1+\sqrt{n}\,\|(\boldsymbol{\theta}-\boldsymbol{\theta}_o)\|]} \xrightarrow{p} 0$$

Then, $\sqrt{n}(\widehat{\theta}-\boldsymbol{\theta}_o) \xrightarrow{d} N\left[0, (G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1}\right]$.

As shown by Pollard (1985) the differentiability of the objective function $g(\boldsymbol{\omega}_i^*, \theta)$ can be replaced by the differentiability of $g_o(\boldsymbol{\theta})$ for the purpose of obtaining the asymptotic normality of these estimators. As emphasized by Newey and McFadden (1994) the key condition to allow for nonsmooth objective functions is condition (viii), which is a "stochastic equicontinuity" assumption that guarantees uniform convergence in probability of the linear approximation of $g_o(\boldsymbol{\theta})$ by $g(\boldsymbol{\omega}_i^*, \theta)$ in a shrinking neighborhood of $\boldsymbol{\theta}_o$. This is similar to the stochastic differentiability condition in Pollard (1985) and primitive conditions are available in Pollard (1985), Andrews (1994) and Chen et al. (2003). Those simplify the task of checking its validity to specific moment functions, however this is beyond the scope of this work and I refer the reader to those papers.

Suppose that $\boldsymbol{\theta}$ can be partitioned into subsets of parameters $(\boldsymbol{\beta}', \boldsymbol{\gamma}')' \in \mathbf{B}\times\boldsymbol{\Gamma} \subset \mathbb{R}^{p1}\times\mathbb{R}^{p2}$ and that $g(\cdot)$ can be partitioned into subsets of functions $(g_1(\cdot)', g_2(\cdot)')'$ as defined below. For notational convenience, $\boldsymbol{\omega}^*$ is suppressed in the following discussion, then

$$E[g_1(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o)] = 0 \tag{1.3}$$

$$E[g_2(\boldsymbol{\gamma}_o)] = 0 \tag{1.4}$$

where $\boldsymbol{\beta} \in \mathbf{B}$, $\boldsymbol{\gamma} \in \mathbf{\Gamma}$, $g_1(\cdot)$ and $g_2(\cdot)$ are m_1 and m_2 vectors of known functions, respectively ($m = m_1 + m_2$). Note that the second set of moment conditions does not depend on $\boldsymbol{\beta}$ while the first set of moment conditions depend on the full parameter set $\boldsymbol{\theta}$. Let $g_{n1}(\theta) = n^{-1} \sum_{i=1}^{n} g_1(\boldsymbol{\omega}_i^*, \theta)$ and $g_{n2}(\gamma) = n^{-1} \sum_{i=1}^{n} g_2(\boldsymbol{\omega}_i^*, \boldsymbol{\gamma})$. The framework developed here is valid for the general case of overidentification, *i.e.*, $m_1 \geqslant p_1$ and $m_2 \geqslant p_2$. This guarantees that $\boldsymbol{\gamma}_o$ is identified by 1.4 alone, and that for a given $\boldsymbol{\gamma}$, β_o can be identified by 1.3 alone, hence, two step estimation is possible. Let the asymptotic covariance matrix for the moment functions, Σ, be defined as

$$\Sigma = V\left[g(\boldsymbol{\theta}_o)\right] \equiv \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

where we assume Σ is finite and nonsingular so its inverse exists:

$$\Sigma^{-1} \equiv \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix} = \begin{bmatrix} C_{11}^{-1}(I + C_{12}E^{-1}C_{21}C_{11}^{-1}) & -C_{11}^{-1}C_{12}E^{-1} \\ -E^{-1}C_{21}C_{11}^{-1} & E^{-1} \end{bmatrix}$$

since Σ (and Σ^{-1}) is symmetric $C_{12} = C_{21}'$ and the second equality holds (see White, 1984, p. 80) for $E \equiv C_{22} - C_{21}C_{11}^{-1}C_{12}$.

Define the matrix of derivatives as

$$\begin{aligned} G &\equiv \nabla_\theta g_o(\theta_o) = \nabla_\theta E[g(\theta_o)] \equiv \begin{bmatrix} G_{11} & G_{12} \\ 0 & G_{22} \end{bmatrix} \\ G_{11} &\equiv \nabla_\beta E[g_1(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o)] \\ G_{12} &\equiv \nabla_\gamma E[g_1(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o)] \\ G_{22} &\equiv \nabla_\gamma E[g_2(\boldsymbol{\gamma}_o)] \end{aligned}$$

where the lower off-diagonal matrix equals zero since the second set of moment conditions does not depend on $\boldsymbol{\beta}$.

Following Prokhorov and Schmidt (2009), define four different possible GMM estimators that differ in which moment conditions are used and/or whether γ is treated as known.

Definition 1 *Call the estimator of θ_o that minimizes*

$$g_n(\theta)' \widehat{W} g_n(\theta) \tag{1.5}$$

with the weighting matrix $\widehat{W} = \Sigma^{-1}$ *the ONE-STEP estimator.*

This is the usual GMM estimator that uses all the available orthogonality conditions jointly to estimate $\boldsymbol{\beta_o}$ and $\boldsymbol{\gamma}_o$.

Definition 2 *Call the estimator of* β_o *that minimizes*

$$g_{n1}(\boldsymbol{\beta}, \boldsymbol{\gamma}_o)' C_{11}^{-1} g_{n1}(\boldsymbol{\beta}, \boldsymbol{\gamma}_o) \tag{1.6}$$

and γ_o *is treated as known the KNOW-*γ *estimator.*

This estimator ignores the second set of orthogonality conditions 1.4, treating γ_o as a known vector of parameters and estimating β_o using only the information available in the first set of moment assumptions. This could arise if one has information about the true values of γ_o or if he disregards the fact that γ_o was estimated in the first stage and, hence its variability, in what could be called a "naive" estimator.

Definition 3 *Call the estimator of* β_o *that minimizes*

$$g_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_o)' \Sigma^{-1} g_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_o) \tag{1.7}$$

and γ_o *is treated as known the KNOW-*γ*-JOINT estimator.*

This is the GMM estimator for β_o in the form considered by Qian and Schmidt (1999). In this case, one has information about the true values of γ_o but still uses both set of moments conditions in obtaining an estimate for β_o.

Definition 4 *Call the estimator of* θ_o *obtained in the following fashion, the TWO-STEP estimator:*

(i) the estimator $\widehat{\gamma}$ *is obtained by minimizing*

$$g_{n2}(\boldsymbol{\gamma})' C_{22}^{-1} g_{n2}(\boldsymbol{\gamma}) \tag{1.8}$$

(ii) the estimator $\widehat{\beta}$ is obtained by minimizing

$$g_{n1}(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}})' C_{11}^{-1} g_{n1}(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}) \tag{1.9}$$

and $\widehat{\boldsymbol{\gamma}}$ is treated as given.

This is the sequential estimator that uses only the second set of moment conditions 1.4 to obtain a consistent estimator of the unknown parameter vector γ_o and then uses only the first set of moment conditions 1.3 to obtain the estimator of β_o. This estimator is widely used in the applied economics literature and encompasses several common problems.

The estimators defined above depend on a known Σ. In practice, Σ is not known and has to be replaced by an initial consistent estimate.

To compare the properties of these different estimators we need to obtain their asymptotic variances. Those are derived directly by Theorem 2.

Theorem 3 *Let $V_{ONE-STEP}$, $V_{KNOW-\gamma}$,$V_{KNOW-\gamma-JOINT}$ and $V_{TWO-STEP}$ denote the asymptotic variance of ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP respectively. Then, under the conditions described in Theorem 1 and 2.*

$$\begin{aligned}
V_{ONE-STEP} &= \left(G'\Sigma^{-1}G\right)^{-1} && (1.10)\\
V_{KNOW-\gamma} &= \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} && (1.11)\\
V_{KNOW-\gamma-JOINT} &= \left(G_{11}'C^{11}G_{11}\right)^{-1} && (1.12)\\
V_{TWO-STEP} &= B\Sigma B' && (1.13)
\end{aligned}$$

where,

$$B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

with

$$\begin{aligned}
B_{11} &= -\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} G_{11}'C_{11}^{-1}\\
B_{12} &= \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} G_{11}'C_{11}^{-1}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1} G_{22}'C_{22}^{-1}\\
B_{22} &= -\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1} G_{22}'C_{22}^{-1}
\end{aligned}$$

Proof. All proofs are provided in the appendix. ■

It is possible to analyze the relative asymptotic efficiency of these estimators.[1]

Theorem 4 *For the estimators defined above as the ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP with asymptotic variances given by 1.10, 1.11, 1.12 and 1.13, respectively, the following statements hold:*

1. KNOW-γ-JOINT is no less efficient than ONE-STEP, KNOW-γ and TWO-STEP for β_o.

2. If $C_{12} = 0$ then KNOW-γ-JOINT and KNOW-γ are equally efficient for β_o.

3. If $G_{12} = 0$ then TWO-STEP and KNOW-γ are equally efficient for β_o.

4. If $C_{12} = 0$ and $G_{12} = 0$, then ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP are equally efficient for β_o, and ONE-STEP and TWO-STEP are equally efficient for γ_o.

5. ONE-STEP is no less efficient than TWO-STEP.

6. If $m_1 = p_1$ then the ONE-STEP and TWO-STEP estimates of γ_o are equal.

7. If $m_1 = p_1$ and $m_2 = p_2$ then the ONE-STEP and TWO-STEP estimates are equal for both β_o and γ_o.

8. If $m_1 = p_1$ and $C_{12} = 0$ then the ONE-STEP and TWO-STEP estimates are equally efficient for both β_o and γ_o.

9. If $G_{12} = C_{12}C_{22}^{-1}G_{22}$, then KNOW-$\gamma$-JOINT and ONE-STEP are equally efficient for β_o.

10. If $G_{12} = C_{12}C_{22}^{-1}G_{22}$, then ONE-STEP, KNOW-$\gamma$-JOINT and TWO-STEP are no less efficient for β_o than KNOW-γ.

[1] I denote the asymptotic variance of $\widehat{\theta}$ as V meaning that $\sqrt{n}(\widehat{\theta} - \theta_o)$ converges in distribution to $N(0, V)$.

The statements that form Theorem 4 are direct extensions of Prokhorov and Schmidt (2009) for the case in which nonsmooth objective functions are allowed.

Statement 1 shows, as expected, that KNOW-γ-JOINT dominates the other estimators. This is an intuitive result since the known value of γ_o is at least as efficient as any estimate of γ_o, and KNOW-γ-JOINT uses the full set of relevant moment conditions.

Statement 2 is the result Qian and Schmidt (1999), where it is shown that using additional moment conditions that include no unknown parameters (as is the case for KNOW-γ-JOINT) improves efficiency except in the special case in which $C_{12} = 0$. In other words, the second set of moments is redundant in the estimation of β_o, Prokhorov and Schmidt (2009) call this M-redundancy.

Statement 3 gives the condition under which the first stage estimation of the nuisance parameter γ_o does not affect the asymptotic behavior of the second stage estimate of β_o. This result is similar to the one shown in Wooldridge (2002a), however in this case we are dealing with a nonsmooth objective function and, therefore, the restriction $G_{12} \equiv \nabla_\gamma E[g_1(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o)] = 0$ differs from the one proposed by Wooldridge since the derivative of $g_1(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o)$ is not necessarily available.

Statement 4 provides conditions under which the ONE-STEP, KNOW-γ, KNOW-γ-JOINT and TWO-STEP estimators are equally efficient for β_o, hence the use of the additional moment conditions in 1.4 by the ONE-STEP, KNOW-γ-JOINT and TWO-STEP estimators does not improve the precision of the estimated parameters of interest as in the previous statement; and the knowledge of γ_o does not help in estimating β_o. This would hold if two sets of moment conditions are asymptotically uncorrelated ($C_{12} = 0$) and γ is not present in the first set of moment conditions ($G_{12} = 0$).

Statement 5 is the usual result that in general, sequential estimation procedures are less efficient than joint (one step) estimation.

Statement 6, 7 and 8 follow directly from Ahn and Schmidt (1995) and show that the GMM separability holds in the framework that allows non-smooth objective functions. The GMM

estimates for γ_o are not improved by the inclusion of an equal number of additional moment conditions and parameters. It can be shown that if G_{11} is nonsingular, the ONE-STEP estimator for β_o can be written in terms of the ONE-STEP estimator of γ_o using the equation $g_{n1}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = C_{12}C_{22}^{-1}g_{n2}(\widehat{\boldsymbol{\gamma}})$ (see appendix for details). Thus, as described by Prokhorov and Schmidt (2009) the ONE-STEP and TWO-STEP estimators for β_o will be derived from the same equation as long as $g_{n2}(\widehat{\boldsymbol{\gamma}}) = 0$, which will be true under exact identification of γ_o, and asymptotically equally efficient if $C_{12} = 0$, since the moment conditions will be asymptotically uncorrelated, not adding to the information set exploited by ONE-STEP relatively to TWO-STEP.

Statement 9 and 10 are direct extensions of Prokhorov and Schmidt (2009). Statement 9 says that KNOW-γ-JOINT and ONE-STEP are equally efficient for the estimation of β_o, which means that knowledge of γ_o is not useful in terms of the efficiency of the estimates for β_o if we are using the full set of moment conditions and $G_{12} = C_{12}C_{22}^{-1}G_{22}$.

Statement 10 shows that under the same condition about G_{12}, KNOW-γ is dominated by ONE-STEP, KNOW-γ-JOINT and TWO-STEP. This happens because knowledge of γ_o is not useful in the estimation of β_o in this case, and the KNOW-γ estimator does not use the information in the second set of moment conditions, which is useful unless $C_{12} = 0$.

The statements presented in theorem 4 show that the results for GMM redundancy presented by Prokhorov and Schmidt (2009) extend to GMM estimation procedures based on nonsmooth objective functions.

Under the conditions of parts 9 and 10 of theorem 4, the following corollary can be obtained.

Corollary 1 *If $G_{12} = C_{12}C_{22}^{-1}G_{22}$ and G_{22} is invertible, then*

$$V(\widehat{\beta}_{TWO-STEP}) = \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} G_{11}'C_{11}^{-1}D_oC_{11}^{-1\prime}G_{11}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1} \quad (1.14)$$

Additionally, if G_{11} is invertible, then

$$V(\widehat{\beta}_{TWO-STEP}) = G_{11}^{-1}D_oG_{11}^{-1\prime} \quad (1.15)$$

where

$$\begin{aligned} D_o &= E\left[e_i e_i'\right] \\ e_i &= \left[g_1(\boldsymbol{\omega}_i^*, \theta) - C_{12} C_{22}^{-1} g_2(\boldsymbol{\omega}_i^*, \gamma)\right] \end{aligned}$$

Note that e_i is the residual of the linear projection of the first set of moments conditions on the second set of moment conditions. This result is useful in the estimation of the asymptotic variance of the estimators, as I discuss below. Unfortunately, this applies only if the second set of moment conditions is exactly identified for formula 1.14 and if both sets of moment conditions are exactly identified for formula 1.15.

An arresting issue is to obtain estimates of the variance matrices described in theorem 3. The nonsmoothness of the objective function creates some obstacles to the usual estimations procedures. As described by Lee (2008) the fact that the estimates for the variances depend on the derivative of the expectation of the estimating function in the nonsmooth case warrants a more careful approach in estimating the variances used for inference.

A general approach that work in most cases is offered in Newey and McFadden (1994), and consists on obtaining consistent estimators for the separate components of the variance matrix. For estimating Σ or its relevant components a standard estimator is available. This procedure can be used in a first-step to obtain consistent estimates of the appropriate

weighting matrix for the desired estimation procedure.

$$\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n} g(\omega_i^*,\hat{\theta})g(\omega_i^*,\hat{\theta}) \tag{1.16}$$

$$\widehat{C}_{11} = n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\theta})g_1(\omega_i^*,\hat{\theta}) \tag{1.17}$$

$$\widehat{C}_{12} = n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\theta})g_2(\omega_i^*,\hat{\gamma}) \tag{1.18}$$

$$\widehat{C}_{22} = n^{-1}\sum_{i=1}^{n} g_2(\omega_i^*,\hat{\gamma})g_2(\omega_i^*,\hat{\gamma}) \tag{1.19}$$

$$\widehat{C}_{21} = \widehat{C'}_{12} \tag{1.20}$$

To be able to plug this estimates on the equations derived in Theorem 3 we need to obtain estimates of G, which can be difficult to obtain due to the nonsmoothness of the objective function. In this approach an estimate of G is obtained by numerical derivatives. Following Newey and McFadden (1994) let e_i denote the i^{th} unit vector, ϵ_n denote a small positive constant that depends on the sample size. Define the estimators for G and its components as

$$\widehat{G_j} = \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g(\omega_i^*,\hat{\theta}+e_j\epsilon_n) - g(\omega_i^*,\hat{\theta}-e_j\epsilon_n)\right]$$

$$\widehat{G}_{11_j} = \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\beta}+e_j\epsilon_n,\hat{\gamma}) - g_1(\omega_i^*,\hat{\beta}-e_j\epsilon_n,\hat{\gamma})\right]$$

$$\widehat{G}_{12_j} = \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g_1(\omega_i^*,\hat{\beta},\hat{\gamma}+e_j\epsilon_n) - g_1(\omega_i^*,\hat{\beta},\hat{\gamma}-e_j\epsilon_n)\right]$$

$$\widehat{G}_{22_j} = \frac{1}{2\epsilon_n}\left[n^{-1}\sum_{i=1}^{n} g_2(\omega_i^*,\hat{\gamma}+e_j\epsilon_n) - g_2(\omega_i^*,\hat{\gamma}-e_j\epsilon_n)\right]$$

Where the subscript j denotes the j^{th} column of the matrix being estimated. Newey and McFadden (1994, Theorem 7.4) shows that if ϵ_n converges to zero and $\sqrt{n}\epsilon_n$ converges to infinity as n gets larger, these estimators will be consistent for the terms of the variances

presented in theorem 3, and plugging them in the formulas provide consistent estimators for the variances of the parameters being estimated.

However, these estimators are cumbersome and not practical. As emphasized by Newey and McFadden, the choice of ϵ_n is a difficult problem and the formulation described above, using a unique value for ϵ_n would be good only if the estimated parameters had been scaled to have similar magnitudes. If that is not done, we would have to pick different ϵ_n for different components.

On specific cases, other estimators are available. As discussed in Newey and McFadden (1994) if $g(\omega^*, \hat{\theta})$ is differentiable with probability one, with $\nabla_\theta g(\omega^*, \hat{\theta})$ that is continuous at θ_o with probability one and dominated by an integrable function in a neighborhood of θ_o, then $\widehat{G} = n^{-1} \sum_{i=1}^{n} \nabla_\theta g(\omega_i^*, \hat{\theta})$ is a consistent estimator for G. Hence, the more standard estimator is available and would be easier to implement.

Clearly, alternatives could be available for specific moment conditions. Section 1.4 provides the example for the leading case of IPW for linear quantile regression.

Even in this case, the calculation of the matrix B that is present in the asymptotic variance of the TWO-STEP estimator could be cumbersome. For the cases in which the conditions from part 9 and 10 of theorem 4 hold, namely $G_{12} = C_{12}C_{22}^{-1}G_{22}$, corollary 1 offers a different approach to the problem of estimating the asymptotic variance in those cases (even though we still need to resort to one of the estimators above to obtain $\widehat{G}_{11}$). We can obtain an estimate of the matrix $E\left[e_i e_i'\right]$ by regressing the first set of moment conditions on the second set of moment conditions in the sample to obtain the residuals $\widehat{e}_i = g_1(\omega_i^*, \hat{\beta}, \hat{\gamma}) - \left[n^{-1} \sum_{i=1}^{n} g_2(\omega_i^*, \hat{\gamma}) g_1(\omega_i^*, \hat{\beta}, \hat{\gamma})\right] \left[n^{-1} \sum_{i=1}^{n} g_2(\omega_i^*, \hat{\gamma}) g_2(\omega_i^*, \hat{\gamma})\right]^{-1} g_2(\omega_i^*, \hat{\gamma})$, and calculating the sample analogue of the desired matrix $\widehat{D} = n^{-1} \sum_{i=1}^{n} \widehat{e}_i \widehat{e}_i'$. Unfortunately, this simple procedure is valid only for the asymptotic variance of the TWO-STEP estimator under the condition above and under exact identification of at least the second set of moment conditions.

For most of the relevant problems, we could use a bootstrap procedure to obtain consistent estimates of the variance of $\hat{\theta}$ directly, but these could be computationally demanding for models in which the solution of the optimization problem for both sets of moment conditions require numerical optimization of the objective function.

1.3 Estimation with missing data

This section specializes the results of the section 1.2 to a model in which missing data is allowed in a framework that expands that proposed by Wooldridge (2002b, 2007) to allow nonsmooth objective functions.

Consider $\boldsymbol{\omega} \in Q \subset R^{\dim(\omega)}$ a random vector with density $f(\boldsymbol{\omega})$; $\beta \in \mathbf{B} \subset \mathbb{R}^{p1}$ a parameter vector, where $\mathbf{B}$ is a compact set. Suppose there is the population moment equation

$$g_o(\beta_o) = E[g(\boldsymbol{\omega}, \beta_o)] = 0 \tag{1.21}$$

where $g : Q \times \mathbf{B} \to \mathbb{R}^{m1}$ is a vector of known real-valued moment functions with $m_1 \geq p_1$, so β_o could be overidentified. Assume β_o is the unique solution to 1.21. I am interested in estimating β_o.

Note that the moment conditions presented above hold in the unselected population. Assume nonrandom sampling occurs and it is characterized by a selection indicator, $s \in \{0, 1\}$, such that $\boldsymbol{\omega}_i$ is observed if and only if $s_i = 1$. Keep in mind that all or part of $\boldsymbol{\omega}_i$ is not observed when $s_i = 0$.

The GMM estimator based on 1.21 using the selected sample, in effect makes the empirical moments $n^{-1} \sum_{i=1}^{n} s_i g(\boldsymbol{\omega}_i, \beta)$ close to zero. These empirical moments are the sample analogues of the population moments of the form

$$E\left[sg(\boldsymbol{\omega}, \beta)\right] = 0 \tag{1.22}$$

which are referred to as the unweighted selected population moments (Prokhorov and Schmidt, 2009; Wooldridge, 2002b). The name emphasizes that they are evaluated at the

selected rather than the full population of interest and differentiates them from the weighted selected population moments defined below. The selectivity problem occurs exactly because 1.22 may not hold; in other words, the value β_o that solves 1.21 may not also solve 1.22 (Prokhorov and Schmidt, 2009). If that happens, the estimate for β_o obtained through this procedure is not generally consistent. In fact, its consistency and potential solutions for the data selection problem will depend on the relationship between the selection process and both the dependent and independent variables.

1.3.1 Data Selection under Ignorability

A straightforward solution is to solve the nonrandom sampling problem using inverse probability weighting (IPW) as shown by Wooldridge (2002b, 2007). To be able to use IPW we need some variables that are reasonable predictors of selection as described in Wooldridge (2007). This is formally stated as an "ignorability" of selection assumption.

Assumption 1 (Wooldridge, 2007, Assumption 3.1) *(i)* $\boldsymbol{\omega}_i$ *is observed whenever* $s_i = 1$*;*

(ii) For a random vector z_i *such that* $P(s_i = 1 \mid \boldsymbol{\omega}_i, z_i) = P(s_i = 1 \mid z_i) \equiv p(z_i)$*;*

(iii) For all $z \in Z \subset \mathbb{R}^J$*,* $p(z) > 0$*;*

(iv) z_i *is observed whenever* $s_i = 1$*.*

Item (ii) in this assumption requires that $s \perp \boldsymbol{\omega} \mid z$. In other words, the selection has to be independent of the y and x conditional on z. As discussed at length by Wooldridge (2007), assumption 1 encompasses a variety of selection schemes common in the missing data literature, including "missing at random", "variable probability sampling", "selection on observables' etc. This allows, for example, that the probability of observing $\boldsymbol{\omega}_i$ to depend on the stratum in which $\boldsymbol{\omega}_i$ falls into; or that z_i is observed only along with $\boldsymbol{\omega}_i$; or that partial information is known about the incompletely observed data. Assumption 1 does not

apply to the "selection on unobservables"[2] case as generally used in econometrics. I will not explore these possibilities directly here, referring the reader to Wooldridge (2007).

Assume that a conditional density determining selection is correctly specified and that a maximum likelihood estimator of the selection model is available.

Assumption 2 (Wooldridge, 2007, Assumption 3.2) *(i) $G(z,\gamma)$ is a parametric model for $p(z)$, where $\gamma \in \Gamma \subset \mathbb{R}^{p2}$ and $G(z,\gamma) > 0$ for all $z \in Z$ and $\gamma \in \Gamma$;*

(ii) There exists γ_o in the interior of Γ such that $p(z) = G(z,\gamma_o)$;

(iii) For a random vector v_i such that $D(v_i \mid \boldsymbol{\omega}_i, z_i) = D(v_i \mid z_i)$, the estimator $\widehat{\gamma}$ solves a conditional maximum likelihood problem of the form

$$\max_{\boldsymbol{\gamma}\in\Gamma} \sum_{i=1}^{n} \ln\left[f(v_i \mid z_i, \gamma)\right] \tag{1.23}$$

where $f(v \mid z, \gamma) > 0$ is a conditional density function known up to the parameters γ_o, and $s_i = h(v_i, z_i)$ for some nonstochastic function $h(\cdot,\cdot)$;

(iv) The solution to 1.23 has the first-order representation

$$\sqrt{n}(\widehat{\gamma} - \gamma_o) = \left\{E\left[d_i\left(\gamma_o\right) d_i\left(\gamma_o\right)'\right]\right\}^{-1} \left(n^{-\frac{1}{2}} \sum_{i=1}^{n} d_i\left(\gamma_o\right)\right) + o_p(1)$$

with $d_i\left(\gamma\right) \equiv \frac{\nabla\gamma f(v_i|z_i,\gamma)'}{f(v_i|z_i,\gamma)}$, which is the $p_2 \times 1$ score vector for the MLE.

The assumption above requires standard regularity conditions about $G(z,\gamma)$, including smoothness of the parametric model. Even though this restricts the possibilities to model the selection process, it includes the most used probability models used in the literature. By doing so, we concentrate on the impacts of nonsmoothness in the model of interest and provide results about the use of IPW in correcting sample selection for those cases. Assumption 2 covers the cases presented by Wooldridge (2002b) in which the conditional log-likelihood was for a binary response model. The advantage of using this slightly more

[2] A quantile regression estimator for the case wheen selection is on unobservables is provided by Buchinsky (1998)

complicated framework is to allow z_i to be only partially observed and to permit s_i to be a function of another random variable v_i which includes a broader class of selection problems. For a deeper discussion on the extensions allowed by assumption 2, see Wooldridge (2007).

Note that the MLE estimator for γ_o described above can be obtained in a GMM setting as follows.

Let $\widehat{\gamma}$ the Maximum Likelihood Estimator (MLE) of $\boldsymbol{\gamma}_o$, that is $\widehat{\boldsymbol{\gamma}}$ solves

$$\max_{\boldsymbol{\gamma}\in\Gamma} \sum_{i=1}^{n} \ln\left[f(v_i \mid z_i,\gamma)\right]$$

Define $g_2(z,\boldsymbol{\gamma},s) \equiv d(\gamma) = \frac{\nabla_\gamma f(v_i|z_i,\gamma)'}{f(v_i|z_i,\gamma)}$ and $g_{n2}(\boldsymbol{\gamma}) \equiv n^{-1}\sum_{i=1}^{n} g_2(z_i,\gamma,s_i)$. Hence, $g_{n2}(\boldsymbol{\gamma}) \xrightarrow{p} g_{2o}(\boldsymbol{\gamma}) \equiv E\left[g_2(z,\gamma,s)\right]$. Then, the problem above is characterized by the following first order conditions

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} g_2(\mathbf{z}_i,\widehat{\boldsymbol{\gamma}},s_i) &= n^{-1}\sum_{i=1}^{n}\left[\frac{\nabla_\gamma f(v_i \mid z_i,\widehat{\gamma})'}{f(v_i \mid z_i,\widehat{\gamma})}\right] \\ &= n^{-1}\sum_{i=1}^{n} d_i(\widehat{\gamma}) = o_p(n^{-\frac{1}{2}}) \end{aligned}$$

and,

$$E\left[g_2(\mathbf{z},\boldsymbol{\gamma}_o,s)\right] = E\left[d(\gamma_o)\right] = 0$$

Under assumption 1, the following lemma, presented in Wooldridge (2002b) is valid.

Lemma 1 (Wooldridge, 2002b, Lemma 3.1) *Under the conditions presented in Assumptions 1 and 2, for any real-valued function $g(\boldsymbol{\omega})$ such that $E\left[\frac{|g(\boldsymbol{\omega},\boldsymbol{\beta}_o)|}{G(z,\gamma_o)}\right] < \infty$,*

$$E\left\{\left[\frac{s}{G(z,\gamma_o)}\right] g(\boldsymbol{\omega},\beta_o)\right\} = E\left\{\left[\frac{s}{p(z)}\right] g(\boldsymbol{\omega},\beta_o)\right\} = E\left[g(\boldsymbol{\omega},\beta_o)\right] \tag{1.24}$$

Lemma 1 suggests that we use the sampling probabilities to consistently estimate $\boldsymbol{\beta}_o$ Consider the weighted selected population moments that weight 1.22 by the inverse of the

selection probability:

$$E\left\{\left[\frac{s}{G(z,\gamma_o)}\right]g(\boldsymbol{\omega},\beta_o)\right\}=0 \tag{1.25}$$

Given an estimator for γ_o, $\widehat{\gamma}$, we can form $G(z_i,\widehat{\gamma})$ for all i with $s_i=1$ and we are able to obtain consistent estimates for β_o by using the weighted selected population moments 1.25 as described in Wooldridge (2007). Note that, by the Law of Large Numbers and Law of Iterated Expectations, assumptions 1, 2 and consistency of $\widehat{\gamma}$ for γ_o (see Wooldridge, 2002b, theorem 3.1).

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\frac{s_i}{G(z_i,\widehat{\gamma})}g(\boldsymbol{\omega}_i,\boldsymbol{\beta}) \xrightarrow{p} E\left[\frac{s_i}{p(z_i)}g(\boldsymbol{\omega}_i,\boldsymbol{\beta})\right] \\
= \;& E\left\{E\left[\frac{s_i}{p(z_i)}g(\boldsymbol{\omega}_i,\boldsymbol{\beta})\mid \boldsymbol{\omega}_i,z_i\right]\right\} \\
= \;& E\left\{\frac{1}{p(z_i)}E\left[s\mid \boldsymbol{\omega}_i,z_i\right]E\left[g(\boldsymbol{\omega}_i,\boldsymbol{\beta})\mid \boldsymbol{\omega}_i,z_i\right]\right\} \\
= \;& E\left\{\frac{p(z_i)}{p(z_i)}E\left[g(\boldsymbol{\omega}_i,\boldsymbol{\beta})\mid \boldsymbol{\omega}_i,z_i\right]\right\} \\
= \;& E\left[E\left[g(\boldsymbol{\omega}_i,\boldsymbol{\beta})\mid \boldsymbol{\omega}_i,z_i\right]\right]=E\left[g(\boldsymbol{\omega}_i,\boldsymbol{\beta})\right]=g_o(\boldsymbol{\beta})
\end{aligned}$$

Therefore,

$$n^{-1}\sum_{i=1}^{n}\frac{s_i}{G(z_i,\widehat{\gamma})}g(\boldsymbol{\omega}_i,\boldsymbol{\beta}_o)\xrightarrow{p} g_o(\boldsymbol{\beta}_o)=0$$

Hence, this provides a set of valid moment conditions that could be used to estimate β_o.

Efficiency Comparisons

The relative efficiency of the estimators for β_o that use IPW to correct a missing data problem under assumption 1 and 2 can be analyzed under the framework developed in section 1.2. Consider the two sets of moment functions

$$\begin{aligned}
g_1(\boldsymbol{\omega},z,\beta,\gamma,s) &= \frac{s}{G(z,\gamma)}g(\boldsymbol{\omega},\boldsymbol{\beta}) \\
g_2(z,\gamma,s) &\equiv d(\gamma)=\frac{\nabla_\gamma f(v\mid z,\gamma)'}{f(v\mid z,\gamma)}
\end{aligned}$$

and the following moment conditions are valid,

$$E\left[g_1(\boldsymbol{\omega}, z, \beta, \gamma, s)\right] = 0 \tag{1.26}$$

$$E\left[g_2(z, \gamma_o, s)\right] = 0 \tag{1.27}$$

Any of the estimators discussed in section 1.2 can be used, differing on the set of moment conditions used and the knowledge about the weights.

Under the assumptions on the moment conditions and the selection process discussed in this section, the following lemma holds.

Lemma 2 *If the conditions of Theorem 1 and 2; Assumptions 1 and 2 hold, and the moment conditions are defined by 1.26 and 1.27, then* $G_{12} = C_{12}C_{22}^{-1}G_{22}$.

By using this result, we can see that under these assumptions, the results of Theorem 4 can be directly applied to this specific case and the ONE-STEP, TWO-STEP and KNOW-γ-JOINT estimators will be no less efficient than the KNOW-γ.

Theorem 5 *Under the conditions of Lemma 2 , ONE-STEP, KNOW-γ-JOINT and TWO-STEP are no less efficient for* β_o *than KNOW-γ. Furthermore, ONE-STEP and KNOW-γ-JOINT are equally efficient for* β_o.

Hence, unless $C_{12} = 0$ (in which case the four estimators would be equally efficient), using ONE-STEP or TWO-STEP that estimate γ_o through MLE produce more efficient estimates for β_o than using known weights (if we knew them) in the KNOW-γ estimator. The KNOW-γ-JOINT estimator is as efficient as ONE-STEP as well, indicating that the knowledge of γ_o is not useful in terms of the efficiency of the estimates for β_o. The efficiency gains relatively to KNOW-γ are due to the use the information in the second set of moment conditions.

Therefore, the puzzle described in Wooldridge (2002b, 2007) that KNOW-γ is inefficient relative to TWO-STEP, extends to a larger set of estimators in which the original set of unweighted moment conditions is nonsmooth as it was discussed by Chen et al. (2008) and

Hitomi et al. (2008). In these cases we are better off estimating the weights by a conditional MLE than knowing them. Nonetheless, the TWO-STEP estimator is dominated by both ONE-STEP and KNOW-γ-JOINT and those should be used to obtain relatively efficient estimates of β_o.

It is important to note that the framework developed in this chapter does not extend directly to semiparametric cases in which the probability of selection is estimated nonparametrically. That can be a serious inconvenience when we have limited information about the selection process and would benefit from a more flexible estimator to those probabilities. However, as it is shown in the section 1.3.2 we can obtain consistent estimates for β_o even if using misspecified selection probabilities, as long as the data selection is exogenous.

1.3.2 Data Selection under Exogeneity of Selection

The literature in sample selection has long established that sample selection does not necessarily cause bias in unweighted estimators. As shown in Wooldridge (2007) if selection is exogenous conditional on the vector of covariates x the estimators of interest using the unweighted moment conditions will be consistent and, in fact, more efficient (Prokhorov and Schmidt, 2009) than the weighted estimators. Following Wooldridge (2007), I analyze the properties of the estimators obtained under exogenous selection but with potential misspecification of the selection model. The main results about consistency of the estimators for the parameters of interest shown in Wooldridge (2007) and Prokhorov and Schmidt (2009) remain unaltered by the fact that the moment conditions are allowed to be nonsmooth as summarized below.

Consider that we have a potentially misspecified model for the probability of selection given by $G(z, \gamma^*)$, which is not necessarily equal to the true $p(z_i)$. Assume that the estimate $\widehat{\gamma}$ obtained based on that model is consistent to some parameter vector γ^* and $\sqrt{n}(\widehat{\gamma} - \gamma^*) = O_p(1)$.

In this case, the weighted moment condition

$$n^{-1}\sum_{i=1}^{n}\frac{s_i}{G(z,\widehat{\gamma})}g(\boldsymbol{\omega}_i,\boldsymbol{\beta}_o)\longrightarrow^{p} E\left[\frac{s}{G(z,\gamma^*)}g(\boldsymbol{\omega},\boldsymbol{\beta})\right] \tag{1.28}$$

instead of $E\left[g(\boldsymbol{\omega},\boldsymbol{\beta})\right]=0$, as seen in section 1.3.1.

Assume that the selection process is exogenous conditional on z.

Assumption 3 (Wooldridge, 2007, Assumption 4.1) *(i) $\boldsymbol{\omega}_i$ is observed whenever $s_i = 1$;*

(ii) For a random vector z_i such that $P(s_i = 1 \mid \boldsymbol{\omega}_i, z_i)$ =$P(s_i = 1 \mid z_i)$ ≡$p(z_i)$;

(iii) z_i is observed whenever $s_i = 1$.

(iv) $\beta_o \in B$ solves the problem

$$E\left[g(\boldsymbol{\omega},\beta) \mid z\right]=0$$

for all $z \in Z$.

This assumption is the same as in Prokhorov and Schmidt (2009) and as shown by them in Lemma 4.1 and Theorem 4.1 (p.53), which are not altered due to the use of nonsmooth objective functions, it implies

$$E\left[g(\boldsymbol{\omega},\beta) \mid z,s\right]=0$$

Hence, any function of z and s is uncorrelated with $g(\boldsymbol{\omega},\beta)$ and both weighted and unweighted moment conditions hold in the selected sample for *any* weighting (that is a function of z and s) that we could use. Therefore, the weighted moment condition in equation 1.28 will hold in the selected sample for any misspecified model $G(z,\gamma^*)$. Obviously, this holds for the unweighted moment conditions as well since it is equal to the special case in which $G(z,\gamma^*)=1$.

Then, we conclude that under exogeneity of selection, the IPW estimator for β_o proposed is consistent, regardless of the misspecification of the model for probability of selection[3].

[3] This conclusion is equivalent to Theorem 4.1 in Wooldridge (2007), extending it for nonsmooth objective functions.

This robustness is an important feature of the IPW procedure and adds to its usefulness in applications.

1.4 Examples

1.4.1 Quantile Regression under Ignorability of Selection

Quantile regression is one of the main motivations for this research. As an example of the use of the results presented here, consider I am interested in estimating the conditional quantile function (CQF) of a random variable y conditional on a vector of explanatory variables x. This is defined by,

$$Q_\tau(Y \mid X) = \inf \{y : F_Y(y \mid X) \geq \tau\}$$

where $\tau \in (0, 1)$ indexes the τ^{th} quantile of the conditional distribution of Y. Suppose that the CQF is a linear model

$$Y = X'\beta_{\tau_o} + \varepsilon$$

and that $Q_\tau(\varepsilon \mid X) = 0$. Then,

$$Q_\tau(Y \mid X) = X'\beta_{\tau_o}$$

In the population, β_o solves the following problem

$$\min_{\beta \in \mathbf{B}} E\left(\rho_\tau(Y - X'\beta_\tau)\right)$$

where, $\rho_\tau(u) = (\tau - 1\,[u \leq 0])u$

Given a random sample from the population of size n, it is possible to obtain consistent estimates of β_o by a standard quantile regression (QR) estimator.

$$\min_{\beta \in \mathbf{B}} n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta_\tau)$$

Note that the population minimization problem has the following first order conditions

$$E\left\{\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right\} = 0$$

and their sample analogue is (Buchinsky, 1998)

$$n^{-1}\sum_{i=1}^{n}\left(\tau - 1\left[y_i - x_i'\widehat{\beta_\tau} \leq 0\right]\right)x_i = o_p(n^{-\frac{1}{2}})$$

Hence, we frame this problem as a GMM estimator that uses as moment conditions the first order conditions of the QR problem, since these identify β_{τ_o}. However, suppose a random sample of (y, x) is not observed. We have a selection problem such that the full vector (y_i, x_i) is observed only if a certain binary variable that defines selection equals the unity, $s_i = 1$, if $s_i = 0$ at least some part of (y_i, x_i) is not observed. Then, in the selected sample, we can only estimate

$$n^{-1}\sum_{i=1}^{n} s_i\left\{\left(\tau - 1\left[y_i - x_i'\widehat{\beta_\tau} \leq 0\right]\right)x_i\right\} = o_p(n^{-\frac{1}{2}})$$

which is the sample analogue of

$$E\left\{s\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right]\right\} = 0$$

but the value β_{τ_o} that solves the population moment condition does not necessarily solve the selected population moment condition. Additionally, assume that the probability of selection can be written as a parametric function of some vector of variables (x_i, z_i) and parameters γ_o and that conditional on z_i, the terms of x_i that are not included in z_i and y_i are irrelevant for the probability of selection (Assumption 1).

$$P(s_i = 1 \mid y_i, x_i, z_i) = P(s_i = 1 \mid z_i) \equiv p(z_i, \gamma_o)$$

In this situation, we can estimate consistent and asymptotically normal estimates for β_{τ_o} using the selected sample by weighting the original observations by the inverse of the

probability of selection. Note that,

$$\begin{aligned} & E\left\{\frac{s}{p(z,\gamma_o)}\left[\left(\tau-1\left[y-x'\beta_{\tau o}\leq 0\right]\right)x\right]\right\} \\ = \; & E\left\{E\left[\frac{s}{p(z,\gamma_o)}\left[\left(\tau-1\left[y-x'\beta_{\tau o}\leq 0\right]\right)x\right]\mid x,y,z\right]\right\} \\ = \; & E\left\{\frac{E\left(s\mid z\right)}{p(z,\gamma_o)}\left[\left(\tau-1\left[y-x'\beta_{\tau o}\leq 0\right]\right)x\right]\right\} \\ = \; & E\left\{\left[\left(\tau-1\left[y-x'\beta_{\tau o}\leq 0\right]\right)x\right]\right\}=0 \end{aligned}$$

then $E\left\{\frac{s}{p(z,\gamma_o)}\left[\left(\tau-1\left[y-x'\beta_{\tau o}\leq 0\right]\right)x\right]\right\}=0$ holds. Therefore, we can estimate $\beta_{\tau o}$ by using those weighted moment conditions. Naturally, we would need to estimate the weights if they are unknown.

Let the true selection model be a standard binary response model for simplicity. Then, estimate the selection of probability by MLE, or more conveniently, a GMM procedure that uses the first order conditions of the MLE for the selection model as moment conditions. The MLE maximization problem and its first order condition are given by, respectively,

$$\max_{\boldsymbol{\gamma}\in\Gamma}\sum_{i=1}^{N}\left\{s_i\ln\left[p(z_i,\boldsymbol{\gamma})\right]+(1-s_i)\ln\left[1-p(z_i,\boldsymbol{\gamma})\right]\right\} \tag{1.29}$$

$$n^{-1}\sum_{i=1}^{n}\left[\nabla'_{\gamma}p(z_i,\widehat{\gamma})\frac{s_i-p(z_i,\widehat{\gamma})}{p(z_i,\widehat{\gamma})\left(1-p(z_i,\widehat{\gamma})\right)}\right]=0 \tag{1.30}$$

where the estimator for γ_o is defined as the vector $\widehat{\gamma}$. Again, 1.30 is the sample analogue of the following moment condition,

$$E\left[\nabla'_{\gamma}p(z,\gamma_o)\frac{s-p(z,\gamma_o)}{p(z,\gamma_o)\left(1-p(z,\gamma_o)\right)}\right]=0$$

Hence, we have two sets of moment conditions that can be used to estimate both the selection model and the conditional median model. The GMM estimator in this case would be given by any of the four estimators proposed in section 1.2, with

$$\begin{aligned} g_{n1}(\theta) &= n^{-1}\sum_{i=1}^{n}\frac{s_i}{p(z_i,\gamma)}\left\{\left(\tau-1\left[y_i-x_i'\beta_{\tau}\leq 0\right]\right)x_i\right\} \\ g_{n2}(\gamma) &= n^{-1}\sum_{i=1}^{n}\left[\nabla'_{\gamma}p(z_i,\gamma)\frac{s_i-p(z_i,\gamma)}{p(z_i,\gamma)\left(1-p(z_i,\gamma)\right)}\right] \end{aligned}$$

the variance of the estimates will depend on the choice of estimator as stated by Theorem 4.

To estimate the variance of the estimated parameters we need to obtain valid estimates for the components of G in the variance of $\hat{\theta}$.Note that, for example,

$$\begin{aligned}
G_{11} &\equiv \nabla_\beta E[g_1(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o)] = \nabla_\beta E\left\{\frac{s}{p(z,\gamma_o)}\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right]\right\} \\
&= \nabla_\beta E\left\{E\left[\frac{s}{p(z,\gamma_o)}\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right] \mid z, x, s\right]\right\} \\
&= \nabla_\beta E\left\{\frac{s}{p(z,\gamma_o)}E\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right) \mid z, x, s\right]x\right\} \\
&= \nabla_\beta E\left\{\frac{s}{p(z,\gamma_o)}\left(\tau - F_{y|z,x,s}(x'\beta_{\tau_o})\right)x\right\} \\
&= E\left\{\frac{s}{p(z,\gamma_o)}f_{y|z,x,s}(x'\beta_{\tau_o})x'x\right\}
\end{aligned}$$

hence, consistent estimates can be obtained by the sample analogue,

$$\begin{aligned}
\widehat{G}_{11} &= n^{-1}\sum_{i=1}^{n}\frac{s_i}{p(z_i,\hat{\gamma})}\widehat{f}_{y|z,x,s}(x_i'\widehat{\beta}_\tau)x_i'x_i \\
\widehat{G}_{12} &= n^{-1}\sum_{i=1}^{n} -\frac{\nabla_\gamma' p(z_i,\hat{\gamma})}{[p(z_i,\hat{\gamma})]^2}s_i\left[\left(\tau - 1\left[y_i - x_i'\widehat{\beta}_\tau \leq 0\right]\right)x_i\right] \\
\widehat{G}_{22} &= n^{-1}\sum_{i=1}^{n}\left[\nabla_\gamma' p(z_i,\hat{\gamma})\left(\frac{s_i - p(z_i,\hat{\gamma})}{p(z_i,\hat{\gamma})\left(1 - p(z_i,\hat{\gamma})\right)}\right)^2\nabla_\gamma p(z_i,\hat{\gamma})\right]
\end{aligned}$$

where the last equality is a direct application of GIME and $\widehat{f}_{y|z,x,s}(\cdot)$ is a suitable estimator of the conditional density of y, commonly by a kernel estimator.

Note that the same asymptotic variance formula for the KNOW-γ estimator for $\widehat{\beta}_\tau$ is obtained by a simple extension of the results for weighted quantile regression presented in Koenker (2005) as shown in claim 1 in the appendix.

Since the conditions in Theorem 5 hold, we will obtain more efficient estimates by estimating the inverse probability weights than using the "true" weights, characterizing the puzzle described in the literature (Wooldridge, 2002b, 2007). The relatively more efficient estimate for β_{τ_o} is given by the one-step estimator that jointly estimates both the probability weights and the parameters of interest, β_{τ_o}.

One interesting point to note is that, even this relatively restrictive model for the CQF, which assumes linearity, can be very insightful about the potentially nonlinear true CQF. As discussed in detail by Angrist, Chernozhukov, and Fernández-Val (2006), a linear quantile regression provides the best linear approximation of the true CQF in the sense that it minimizes a weighted mean square error loss function. So even if we have reasons to believe that the true CQF in which we are interested is nonlinear, the use of a linear quantile regression in the example above would provide us with the "best linear approximation" to it in a similar way that a linear OLS model offers the best linear approximation to the conditional mean function. Hence, by using IPW to correct the selection bias caused by missing data we can recover this linear approximation to the CQF of interest, even if we don't know its true specification.

Nevertheless, this framework can be applied to nonlinear conditional quantiles of the form $Q_\tau(Y \mid X) = m\left(X, \beta_{\tau_o}\right)$, with

$$\begin{aligned}
g_{n1}(\theta) &= n^{-1}\sum_{i=1}^{n}\frac{s_i}{p(z_i,\gamma)}\left\{\left(\tau - 1\left[y_i - m\left(x_i,\beta_\tau\right)\leq 0\right]\right)\nabla_\beta m\left(x_i,\beta_\tau\right)\right\} \\
\widehat{G}_{11} &= n^{-1}\sum_{i=1}^{n} -\frac{s_i}{p(z_i,\hat{\gamma})}\widehat{f}_{y|z,x,s}\left(m\left(x_i,\widehat{\beta}_\tau\right)\right)\nabla'_\beta m\left(x_i,\widehat{\beta}_\tau\right)\nabla_\beta m\left(x_i,\widehat{\beta}_\tau\right) \\
\widehat{G}_{12} &= n^{-1}\sum_{i=1}^{n} -\frac{\nabla'_\gamma p(z_i,\hat{\gamma})}{\left[p(z_i,\hat{\gamma})\right]^2}s_i\left[\left(\tau - 1\left[y_i - m\left(x_i,\widehat{\beta}_\tau\right)\leq 0\right]\right)\nabla_\beta m\left(x_i,\widehat{\beta}_\tau\right)\right]
\end{aligned}$$

and the remaining equations unchanged.

1.4.2 Instrumental Variable Quantile Regression

Consider a simplified version of the IVQR estimator described in Chernozhukov and Hansen (2006). Focus on the basic linear model that allow for heterogeneous effects given by,

$$Y_d = q(d,x,\tau) = d'\alpha_\tau + x'\beta_\tau$$

where d is a vector of (potentially endogenous) multi-valued treatment variables and x is a vector of covariates. Under the conditions described in Assumption 1 of Chernozhukov and

Hansen (2006), the IVQR estimator of the vector of parameters $(\alpha(\tau)', \beta(\tau)')'$ proposed in that paper approximately solves the estimating equation[4]:

$$n^{-1}\sum_{i=1}^{n}(1\left[y_i - d_i'\alpha_\tau - x_i'\beta_\tau \leq 0\right] - \tau)(x_i', \widehat{\Phi}_{i\tau}')' = o_p(n^{-\frac{1}{2}})$$

where $\widehat{\Phi}_{i\tau} \equiv \widehat{\Phi}_\tau(\tau, x_i, z_i)$ is a vector of transformations of the instruments. In a simple model $\widehat{\Phi}_{i\tau}$ can be formed by the least squares projection of d on z and x (and its powers) (Chernozhukov and Hansen, 2006, 2008). In that simple case, we could write the sample analogue of the moment conditions that will identify the parameters of the model as

$$\begin{aligned} g_{n1}(\theta) &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{(1\left[y_i - d_i'\alpha_\tau - x_i'\beta_\tau \leq 0\right] - \tau)(x_i', (x_i', z_i')\gamma)'\right\} \\ g_{n2}(\gamma) &= n^{-1}\sum_{i=1}^{n}(x_i', z_i')'[d_i - (x_i', z_i')\gamma] \end{aligned}$$

Hence, the analysis developed in section 1.2 can be applied to the IVQR estimator proposed by Chernozhukov and Hansen (2005, 2006, 2008) and the results shown above are valid in its scope. Nevertheless, it is important to note that the framework developed here does not extend directly to semiparametric cases in which the "first stage" is estimated nonparametrically. That can be a serious inconvenience when we have limited information about the form of the transformation on the vector of instruments that would be preferable in estimating IVQR.

1.5 Conclusion

This chapter (i) extends the GMM efficiency and redundancy results of Prokhorov and Schmidt (2009) to nonsmooth objective functions; (ii) analyzes the extent to which these results could be useful in the context of inverse probability weighting (IPW) as a mechanism to correct missing data issues, thus allowing its use in the LAD and quantile regression

[4] For simplicity I'm assuming that the weights $\widehat{V}_{i\tau}$ in Chernozhukov and Hansen (2006) are equal to the unit.

framework; (iii) verifies the conditions under which the puzzle of selectivity literature, i.e., that weighting using known probabilities of selection leads to a less efficient estimate than using estimated probabilities of selection (Wooldridge, 2002b, 2007; Prokhorov and Schmidt, 2009; Hitomi et al., 2008), is valid under nonsmoothness of the objective functions that characterize the models of interest; and (iv) shows that even in that case the widely used two-step estimator is relatively less efficient than a one-step joint estimator.

Section 1.2 extends results on redundancy and efficiency due to Prokhorov and Schmidt (2009) that can now be applied to a wide range of contexts in which nonsmooth objective functions can be useful, including LAD, quantile regression, censored LAD and quantile treatment effects. Joint estimation of nuisance parameters and parameters of interest is more efficient than a two-step procedure or knowing the true nuisance parameters in the nonsmooth case. This springs from the information contained in the correlation between both sets of moment conditions which is useful, even when γ_o is known. Using only the first set of moment conditions and known values of γ_o in the estimation procedure does not use the additional information embedded in the second set of moment conditions, being inefficient. Some possible consistent estimators for the variance of both sets of parameters are presented.

Section 1.3 analyzes the missing data problem described in Wooldridge (2007). The selection model is estimated by a conditional MLE procedure, but the assumptions about the selection model are weak enough to cover most of the common parametric selection processes in the literature, like attrition, variable probability, "missing at random", etc. One important case not covered is "selection on unobservables". The results from Wooldridge (2007) and Prokhorov and Schmidt (2009) extend to nonsmooth objective functions. If we use both sets of moment conditions, knowledge about the nuisance parameters is not useful for the efficiency of the estimates of the parameters of interest. Additionally, the moment conditions that are associated with the selection model are not redundant, except in special cases. Estimating the parameters of interest using only the first set of moment conditions

with known probabilities of selection as weights is inefficient because it ignores information in the second set of moment conditions. This is the type of puzzle referred to in the selectivity literature, specially in the IPW approach to missing data.

In summary, IPW can be used to correct missing data problems when the model of interest is based on nonsmooth objective functions. Furthermore, two-step estimation of β_o is more efficient than using known probabilities of selection. Nonetheless, the two-step estimator is dominated by a one-step joint estimation procedure that uses both the weighted moment conditions and the selection model's conditions. Hence, the analysis by Prokhorov and Schmidt (2009) extends to the relative efficiency of an IPW approach to deal with missing data problems in which the moment conditions of interest are nonsmooth, encompassing, for example, LAD, quantile regression, Censored LAD and IVQR.

Finally, two illustrative examples of interesting models are provided that are encompassed by the general framework developed in this work. The first is a quantile regression model with missing data and, the second one is a simplified version of the Instrumental Variable Quantile Regression estimator (IVQR) presented by Chernozhukov and Hansen (2006).

CHAPTER 2

Fixed Bandwidth Asymptotics for Regression Discontinuity Designs

2.1 Introduction

Regression discontinuity (RD) designs have been propelled to the spotlight of economic analysis in recent years[1], especially in the policy and treatment evaluation literatures, as a form of estimating treatment effects in a non-experimental setting. The appeal of RD comes from the relative weak assumptions necessary for the identification of treatment effects and inference, which rely on RD's "quasi-experimental" characteristics.

The standard approach to derive the asymptotic properties of estimates obtained in RD settings relies on the traditional assumption that the bandwidth, h, used in the estimation procedure shrinks towards zero as the sample size grows. This guarantees identification of the parameter of interest under mild conditions. Hence, the asymptotic distribution of the estimator used as the basis for inference depends crucially on this small-h condition.

[1] Lee and Lemieux (2009), in a broad review of the RD literature, compile a list of more than 60 papers that apply RD design to many different contexts. The overwhelming majority of the papers have been published in the last decade.

In practice, to obtain an estimate of the parameter of interest and perform inference about it, the empiricist is required to choose a fixed bandwidth greater than zero. Hence, even though the asymptotic theory requires that $h \to 0$, in practice $h > 0$ and fixed. Asymptotic distributions that treat h as fixed can provide a more refined approximation of the asymptotic behavior of the estimator than those derived under the assumption that $h \to 0$.

This chapter derives the asymptotic distribution for the local polynomial estimator when the bandwidth is allowed to be any positive real number. The results shown in section 2.5 provide a new, fixed-h, approximation to the estimator's bias and variance that incorporate the bandwidth size chosen by the researcher.

Corollary 2 shows that the standard small-h asymptotic distribution of the parameter of interest is a special case of fixed-h in which $h \to 0$. Also, corollary 3 shows that when a fixed $h > 0$ is used the standard small-h result for the variance of the estimators is equivalent to assuming that the density of the running variable and the conditional variance of the outcome variable are constant around the discontinuity.

The increased theoretical interest in RD started with Hahn, Todd, and Van der Klaauw (1999, 2001), who presented the conditions for identification of the average treatment effect of interest and its estimation exploiting discontinuities in the probability of treatment provision, which are determined by the so-called running variable. They also derived the asymptotic distribution of the estimators by looking at a shrinking bandwidth around the discontinuity. Porter (2003) provided widely used results on the asymptotic properties of the estimators for the treatment effect of interest, obtaining limiting distributions for estimators based on local polynomial regression and partially linear estimation.

Imbens and Lemieux (2008) and Lee and Lemieux (2009) offer a broad review of the theoretical and applied literature with emphasis on the identification of the parameter of interest and its potential interpretation as a weighted average treatment effect.

The analysis of asymptotic properties of estimators for fixed bandwidths has received some attention in other literatures. Notably, Neave (1970), in the framework of spectral

density estimation, obtains more accurate approximations to the variance of nonparametric spectral estimates by acknowledging that, with a finite sample, the bandwidth used is fixed. He asserts that, in the context of his paper, the assumption equivalent to the bandwidth converging to zero: "(...) is a convenient assumption mathematically in that, in particular, it ensures consistency of the estimates, but it is unrealistic when such results are used as approximations to the finite case(...)"(Neave, 1970, p. 70). Also, Fan (1998) provides an alternative approximation for goodness-of-fit tests for density function estimates in which the bandwidth used in the test is fixed, obtaining improved approximations to the asymptotic behavior of the test and more appropriate critical values for inference.

The same can be said in the regression discontinuity design. Even though $h \to 0$ is a convenient assumption that guarantees consistency of the estimates of the average treatment effect, it will be unrealistic. It is of theoretical and practical interest to obtain more accurate asymptotic distributions by treating h as fixed so that the theory used for inference is more accurate and aligned with the practice of applied economists.

Monte Carlo simulations in section 2.7.1 indicate that, compared with small-h, asymptotic distributions derived based on fixed-h better characterize the behavior of the estimators and provide improved inference about the treatment effect, reducing size distortions in tests and better approximating the bias in the estimates. These improvements are more important when the bandwidth is farther from zero, as one would expect.

Section 2.6 proposes estimators for the asymptotic variance based on the fixed-h results and provides evidence, through Monte Carlo simulations (section 2.7.2), that the feasible inference incorporates the inference improvements predicted by the theory, suggesting that the theoretical gains in robustness can be translated to practical benefits in applied work. Section 2.7.2 also compares the performance of the small-h standard error estimators proposed in the literature in performing inference. The fixed-h variance estimators can improve markedly over small-h estimators in the presence of some forms of heteroskedasticity. Simulations using heteroskedastic errors have provided evidence that feasible tests based on the

fixed-h approach obtain better coverage, outperforming small-h starting at relatively small bandwidths.

Interestingly, in the case of the widely used local linear estimator with homoskedastic errors the variance estimators based on small-h asymptotics suggested in the literature produce well behaved tests with similar size performance to the fixed-h variance estimators, performing better than the standard theory would expect.

2.2 Model

The interest lies in estimating the average treatment effect, τ, of a certain treatment or policy that affects part of a population of interest. As discussed in Porter (2003); Imbens and Lemieux (2008) and Lee and Lemieux (2009), RD designs are closely associated with the treatment effect literature.[2] There are two types of RD designs, sharp and fuzzy, and they differ as to how treatment is assigned to a certain observation and the impact of the discontinuity in its assignment. I will focus on the sharp design in this section and emphasize the differences of the fuzzy design when needed.

2.2.1 Sharp Regression Discontinuity Design

In the sharp design, the treatment status, D, is a deterministic function of a so called "running" or "forcing" variable, x, such that,

$$d_i = \left\{ \begin{array}{l} 1 \text{ if } x_i \geq \overline{x} \\ 0 \text{ if } x_i < \overline{x} \end{array} \right\}$$

where $\overline{x}$ is the *known* cut-off point. Then, let Y_1 and Y_0 be the potential outcomes corresponding to the two possible treatment assignments. As usual, we cannot observe both potential outcomes, having access only to $Y = dY_1 - (1-d)Y_0$. As described by Hahn et al.

[2] Angrist and Pischke (2009) provide a simple introduction to the intuition of regression discontinuity.

(2001) and Porter (2003), under a smoothness assumption that $E\left[Y_j \mid X = x\right]$ is continuous at $\overline{x}$ for $j = 0, 1$, the average treatment effect can be estimated by comparing points just above and just below the discontinuity. The discontinuity in treatment assignment at $\overline{x}$ provides the opportunity for identifying the average treatment effect at the cutoff without any additional parametric functional form restrictions on the conditional expectations of the outcome variable. The average causal effect of the treatment at the discontinuity is Imbens and Lemieux (2008)

$$\begin{aligned} \tau_S &\equiv E\left[Y_1 - Y_0 \mid X = \overline{x}\right] \\ &= \lim_{x \downarrow \overline{x}} E\left[Y \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[Y \mid X = x\right] \end{aligned}$$

where the second equality holds under some smoothness assumptions regarding the conditional expectations (discussed below). The sharp regression discontinuity design uses the discontinuity in the conditional expectation of Y given X to uncover the average treatment effect. If the treatment effect is deemed constant across individuals, τ_S is the effect of treatment for each individual in the population. If we allow the treatment effect to differ among individuals, τ_S is the average treatment effect for individuals at the cutoff. Interestingly, Lee and Lemieux (2009) show that the so-called RD gap obtained by the comparison of observations just above and just below the cutoff can be interpreted as a weighted average treatment effect across all individuals, not only the individuals around the cutoff. In this case each individual would have weights directly proportional to the ex ante likelihood that an individual's realization of X will be close to the threshold. For a comprehensive review of RD designs and their applications and interpretation, see Lee and Lemieux (2009).

2.2.2 Fuzzy Regression Discontinuity Design

In the fuzzy design the probability of receiving treatment still changes discontinuously at the threshold, but is not required to go from 0 to 1, allowing for a smaller jump in the probability

of receiving treatment at the cutoff,

$$\lim_{x \downarrow \overline{x}} \Pr(d \mid X = x) \neq \lim_{x \uparrow \overline{x}} \Pr(d \mid X = x)$$

This framework allows for a greater range of applications since it includes cases in which the incentives to receive (or assign) treatment change discontinuously at the threshold, but are not strong enough to induce all individuals above it to be treated (and those below not to be treated). The average treatment effect at the cutoff can be identified by the ratio of the change in the conditional expectation for the outcome variable to the change in the conditional probability of receiving treatment (Imbens and Lemieux, 2008):

$$\tau_F \equiv \frac{\lim_{x \downarrow \overline{x}} E\left[Y \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[Y \mid X = x\right]}{\lim_{x \downarrow \overline{x}} E\left[d \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[d \mid X = x\right]}$$

This parameter's interpretation is closely linked to the instrumental variables approach. As emphasized by Hahn, Todd, and Van der Klaauw (2001); Imbens and Lemieux (2008) and Lee and Lemieux (2009), a causal interpretation of this ratio requires the same assumptions for local average treatment effects (LATE) presented in Imbens and Angrist (1994). For that we assume monotonicity, i.e., that the treatment status is non-increasing in the cutoff value, or, as stated by Lee and Lemieux (2009, p. 23): "(...)X crossing the cutoff cannot simultaneously cause some units to take up and others to reject the treatment." Also, crossing the cutoff cannot affect the outcome other than by the receipt of treatment, otherwise we would erroneously attribute changes in the conditional expectation of Y due to changes in X to the treatment.

Under these additional assumptions, τ_F has an interpretation similar to the IV estimator, the average treatment effect for the individuals at the threshold (due to the RD design) and only for those whose participation on treatment was affected by the cutoff. Those individuals are described as compliers in the Average Treatment Effect literature[3]. Hence (Imbens and Lemieux, 2008),

[3] See Imbens and Angrist (1994); Hahn, Todd, and Van der Klaauw (2001); Imbens and Lemieux (2008) and Lee and Lemieux (2009).

$$\tau_F \equiv E\left[Y_1 - Y_0 \mid \text{individual is a complier and } X = \overline{x}\right]$$

Similarly as in the sharp RD design, Lee and Lemieux (2009) show that the fuzzy RD design estimator can be interpreted as a weighted LATE with an individual's weight directly proportional to the ex ante likelihood that an individual's realization of X will be close to the threshold.

2.3 Estimators

I analyze estimates for the parameter of interest, τ_S or τ_F, obtained by local polynomial estimators. In applied work local polynomial estimators are a staple for estimation of treatment effects in RD settings. This is partially due to their easy implementation, nice properties and by the fact that the local linear estimator has been the focus on several papers that helped to disseminate the technique (Hahn, Todd, and Van der Klaauw, 1999, 2001; Imbens and Lemieux, 2008; Lee and Lemieux, 2009). Theoretically as well, local polynomial estimators are attractive for estimation in the regression discontinuity setting given its nice boundary behavior as described by Fan and Gijbels (1996).

The order p local polynomial estimator is defined as follows. In the sharp design case, given data $(y_i, x_i)_{i=1,2,...,n}$, let $d_i = 1[x_i \geq \overline{x}]$, $k(\cdot)$ be a kernel function, h denote a bandwidth that controls size of the local neighborhood to be averaged over. Also, define the $p+1 \times 1$ vector $Z(x) = \left(1, (x-\overline{x}), (x-\overline{x})^2, ..., (x-\overline{x})^p\right)'$ and let $\left(\widehat{\alpha}_{p+}, \widehat{\beta}_{p+}\right)'$ be the solution to the minimization problem:[4]

$$\min_{a,b_1,...,b_p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left[y_i - a - b_1(x_i - \overline{x}) - ... - b_p(x_i - \overline{x})^p\right]^2$$

[4] Note that in the sharp RD design, d_i will be identical to the treatment assignment variable D_i since the probability of being treated is zero below the threshold and one above it.

and similarly $\left(\widehat{\alpha}_{p-}, \widehat{\beta}_{p-}\right)$ minimizes the same objective function but with $1 - d_i$ replacing d_i. The estimator of the parameter of interest is given by

$$\widehat{\tau}_S \equiv \widehat{\alpha}_p = \widehat{\alpha}_{p+} - \widehat{\alpha}_{p-}$$

This estimator fits a polynomial on X on a neighborhood just above and below the cutoff for treatment, $\overline{x}$, and encompasses some familiar estimators as special cases.

Case 1 *If $p = 0$ is used, the* Nadaraya-Watson *estimator is obtained. The Nadaraya-Watson estimator takes a kernel weighted average of observations at each side of the discontinuity and its difference.*

$$\begin{aligned}
\widehat{\alpha} &= \frac{n^{-1}\sum_i h^{-1} k\left(\frac{\overline{x}-x_i}{h}\right) y_i d_i}{n^{-1}\sum_j h^{-1} k\left(\frac{\overline{x}-x_j}{h}\right) d_j} - \frac{n^{-1}\sum_i h^{-1} k\left(\frac{\overline{x}-x_i}{h}\right) y_i (1-d_i)}{n^{-1}\sum_j h^{-1} k\left(\frac{\overline{x}-x_j}{h}\right) (1-d_j)} \\
&= \frac{\sum_i k\left(\frac{\overline{x}-x_i}{h}\right) y_i d_i}{\sum_j k\left(\frac{\overline{x}-x_j}{h}\right) d_j} - \frac{\sum_i k\left(\frac{\overline{x}-x_i}{h}\right) y_i (1-d_i)}{\sum_j k\left(\frac{\overline{x}-x_j}{h}\right) (1-d_j)}
\end{aligned}$$

If in addition we use the rectangular kernel, $\widehat{\alpha}$ simplifies to be the difference of the means of y_i in the bandwidths above and below the cutoff.

Case 2 *If the rectangular kernel is used, the* local least squares *estimator of y_i on a polynomial of $(x_i - \overline{x})$ with order p is obtained on the neighborhood on each side of the cutoff. Moreover, if the condition $\widehat{\beta}_{p+} = \widehat{\beta}_{p-}$ is imposed, the estimator $\widehat{\alpha}_p$ is the coefficient on d_i on the* OLS *regression of y_i on d_i and the polynomial of $(x_i - \overline{x})$ with order p using the data inside the bandwidth on both sides of the cutoff.*

In both the theoretical and applied literatures, emphasis has been given to the case in which a linear model ($p = 1$) on X is fitted on each side of the cutoff (Hahn, Todd, and Van der Klaauw, 1999, 2001; Imbens and Lemieux, 2008; Lee and Lemieux, 2009).

Case 3 *If $p = 1$, the* local linear *estimator is obtained. For the rectangular kernel $\widehat{\alpha}$ simplifies to the difference of the intercepts from the linear regression of y_i on 1 and $(x_i - \overline{x})$ in the*

ranges above and below the cutoff. If, additionally, $\widehat{\beta}_{1+} = \widehat{\beta}_{1-}$ *is imposed, the estimator of the ATE of interest is the coefficient for* d_i *on the* OLS *regression of* y_i *on* 1, d_i *and* $(x_i - \overline{x})$ *using the data inside the ranges on both sides of the cutoff.*

2.4 Assumptions

To derive the asymptotic distribution of the estimator for τ, the following assumptions are sufficient.

Assumption 4 *(a)* $k(\cdot)$ *is a symmetric, bounded, Lipschitz function, zero outside a bounded set;* $\int k(u)du = 1$.

(b) For a positive integer s, $\int k(u)u^j du = 0$, $1 \leq j \leq s-1$.

Assumption 4 allows for higher order kernels[5] and a bounded support set for the kernel avoids the use of a trimming function.

Let f_o denote the marginal density of x and $m(x)$ denote the conditional expectation of y given x minus the discontinuity, i.e., $m(x) = E[y \mid x] - \alpha 1[x \geq \overline{x}]$, where $\overline{x}$ is the value of the running variable in which the discontinuity occurs.

Assumption 5 *Suppose the data* $(y_i, x_i)_{i=1,2,...,n}$ *is i.i.d. and* α *is defined by*

$$\alpha = \lim_{x \downarrow \overline{x}} E[y \mid X = x] - \lim_{x \uparrow \overline{x}} E[y \mid X = x]$$

For some compact interval $\aleph$ *of* x *with* $\overline{x} \in int(\aleph)$, f_o *is* l_f *times continuously differentiable and bounded away from zero;* $m(x)$ *is* l_m *times continuously differentiable for* $x \in \aleph \setminus \{\overline{x}\}$, *and* m *is continuous at* $\overline{x}$ *with finite right and left-hand derivatives to order* l_m.

In the sharp RD design $\tau_S = \alpha$ and the average treatment effect is obtained directly from the discontinuity in the conditional expectation of Y. In the following, I discuss the

[5] If $s \geq 3$, the kernel has to be negative for some region of its domain to satisfy part (b) of the assumption.

estimation of α and interpret it as the estimate for the average treatment effect of interest. For the Fuzzy RD design the average treatment effect will be given by the ratio of two such discontinuities, the conditional expectations of the outcome and probability of receiving "treatment".

Assumption 5 guarantees smoothness of the density of x and the conditional expectation of y on both sides of the discontinuity while allowing for different right and left-side derivatives of m at $\overline{x}$. Also, bounding the density of x on the neighborhood around $\overline{x}$ guarantees there is density ("data") around the discontinuity to estimate the jump size.

Assumption 6 describes the behavior of the moments of the outcome variable around the discontinuity. Define, $\varepsilon = y - E[y \mid X = x] = y - m(x) - \alpha 1[x \geq \overline{x}]$.

Assumption 6 *(a)* $\sigma^2(x) = E\left[\varepsilon^2 \mid X = x\right]$ *is continuous for* $x \neq \overline{x}$, $x \in \aleph$, *and right and left-hand limits at* $\overline{x}$ *exist.*

(b) For some $\zeta > 0$, $E\left[|\varepsilon|^{2+\zeta} \mid X = x\right]$ *is uniformly bounded on* $\aleph$.

Assumption 6(a) allows the conditional variance of the outcome variable to be a function of the running variable and assures it is well behaved around the cutoff. Part (b) bounds the moments so that a central limit theorem can be applied.

The fixed-h asymptotic distributions described in section 2.5 do not require additional assumptions over what is used in the standard, small-h literature, *e.g.*, Hahn, Todd, and Van der Klaauw (2001); Porter (2003) etc.

2.5 Asymptotic Distributions

This section develops the asymptotic distribution for the local polynomial estimator of the average treatment effect for a fixed bandwidth, h.

Theorem 6 *Suppose Assumptions 4 (a) and 6 hold. If Assumption 5 (a) holds with* $l_m \geq p+1$ *and* l_f *any nonnegative integer. If* h *is fixed and positive, as* $n \to \infty$*, then*

$$\sqrt{nh}(\widehat{\alpha}_p - \alpha_p^*) \xrightarrow{d} N\left(0, V_{fixed-h}\right) \tag{2.1}$$

where

$$\begin{aligned} V_{fixed-h} &= e_1' \left[\left(\Gamma_+^*\right)^{-1} \Delta_+^* \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^* \left(\Gamma_-^*\right)^{-1} \right] e_1 \qquad (2.2) \\ \alpha_p^* &= \alpha + B_{fixed-h} \\ B_{fixed-h} &= e_1' \left\{ \begin{array}{l} \left(\Gamma_+^*\right)^{-1} \left[\int_0^\infty k(u) Z(\overline{x}+uh) m(\overline{x}+uh) f_o(\overline{x}+uh) du\right] - \\ -\left(\Gamma_-^*\right)^{-1} \left[\int_0^\infty k(u) Z(\overline{x}-uh) m(\overline{x}-uh) f_o(\overline{x}-uh) du\right] \end{array} \right\} \qquad (2.3) \end{aligned}$$

and

$$\Gamma_{+(-)}^* = \begin{bmatrix} \gamma_0^{+(-)} & \cdots & \gamma_p^{+(-)} \\ \vdots & \ddots & \vdots \\ \gamma_p^{+(-)} & \cdots & \gamma_{2p}^{+(-)} \end{bmatrix}, \; \Delta_{+(-)}^* = \begin{bmatrix} \delta_0^{+(-)} & \cdots & \delta_p^{+(-)} \\ \vdots & \ddots & \vdots \\ \delta_p^{+(-)} & \cdots & \delta_{2p}^{+(-)} \end{bmatrix},$$

$e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}'$, $\gamma_j^+ = \int_0^\infty k(u) u^j f_o(\overline{x}+uh) du$,

$\gamma_j^- = (-1)^j \int_0^\infty k(u) u^j f_o(\overline{x}-uh) du$,

$\delta_j^+ = \int_0^\infty k^2(u) u^j \sigma^2(\overline{x}+uh) f_o(\overline{x}+uh) du$,

$\delta_j^- = (-1)^j \int_0^\infty k^2(u) u^j \sigma^2(\overline{x}-uh) f_o(\overline{x}-uh) du$

The proof is given in the appendix.

Theorem 6 provides the asymptotic distribution for the local polynomial estimator of the parameter of interest for any bandwidth value.

The formula for asymptotic variance explicitly takes into consideration the choice of bandwidth, without assuming $h \to 0$. The fixed-h approach used in theorem 6 captures the impact of h on the asymptotic variance, $V_{fixed-h}$. Even though the asymptotic variance formulas are somewhat cumbersome, these are still functions of known data and can be calculated for given functions $f_o(x)$ and $\sigma^2(x)$ or estimated in a dataset (see section 2.6).

The bias term that arises under the fixed-h assumption does not vanish as the sample size increases as suggested by the standard approximations but, for a given bandwidth, it

converges to B_{ar}. The bias is the difference of the (scaled) linear projection for $m(x)$ on Z evaluated at $x = \overline{x}$ (i.e., the difference in intercepts) inside the bandwidth above and below the cutoff. Intuitively, the bias in $\widehat{\alpha}$ is a difference between the conditional expectation of the outcome above and below the cutoff that would have arisen in the absence of treatment, i.e., the difference that would have happened nevertheless and are erroneously attributed to the treatment or policy being analyzed. The fixed-h approach tackles the bias problem "head on", making explicit the impact of the bandwidth choice on the bias of the estimate obtained.

The local polynomial approach mitigates the bias problem if it is able to approximate $m(x)$ appropriately, since it partially captures changes in $m(x)$ above and below the cutoff that would exist even in the absence of treatment by using the higher order polynomials.

Note that, as $h \to 0$ the results for the asymptotic distribution of $\widehat{\alpha}$ in theorem 6 approach the asymptotic variance and bias of small-h asymptotics (Porter, 2003).

Corollary 2 *If the conditions in theorem 6 hold, $h \to 0$, then the asymptotic variance and bias for $\widehat{\alpha}_p$ are equal to the small-h approximation (Porter, 2003)*

$$V_{small-h} = \frac{\sigma^{2+}(\overline{x}) + \sigma^{2-}(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \tag{2.4}$$

$$\alpha^*_{small-h} = \alpha + B_{small-h}$$

$$B_{small-h} = \frac{\lim_{h\to 0} h^{p+1}}{(p+1)!} [m^{(p+1)+}(\overline{x}) - (-1)^{p+1} m^{(p+1)-}(\overline{x})] e_1' \Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix} \tag{2.5}$$

and

$$\Gamma = \begin{bmatrix} \gamma_0 & \cdots & \gamma_p \\ \vdots & \ddots & \vdots \\ \gamma_p & \cdots & \gamma_{2p} \end{bmatrix}, \Delta = \begin{bmatrix} \delta_0 & \cdots & \delta_p \\ \vdots & \ddots & \vdots \\ \delta_p & \cdots & \delta_{2p} \end{bmatrix},$$

$e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}'$, $\gamma_j = \int_0^\infty k(u) u^j du$, $\delta_j = \int_0^\infty k^2(u) u^j du$ *and* $m^{(l)+(-)}(x)$ *is the* l^{th} *right (left)-hand derivative of* $m(x)$ *at point* x.

Additionally, the small-h asymptotic distribution variance and bias in corollary 2 are equal to that obtained by assuming that $f_o(x)$ and $\sigma^2(x)$ are constant around the cutoff and that

$m(x)$ can be exactly approximated by a polynomial of order $p+1$.

Corollary 3 *Under the assumptions in theorem 6, if $h > 0$ and, in the bandwidth around the cutoff, $f_o(x)$ and $\sigma^2(x)$ are constant and $m(x)$ can be exactly approximated by an expansion of order $p+1$. Then, the asymptotic variance and bias of $\sqrt{nh}(\widehat{\alpha}_p - \alpha)$ obtained by fixed-h (theorem 6) and small-h (Porter, 2003) are the same.*

Focusing on the formula for asymptotic variance in both fixed-h and small-h approaches, it is clear that the refinements obtained by fixed-h are due to incorporating the behavior of $f_o(x)$ and $\sigma^2(x)$ in the ranges around the cutoff, while small-h ignores it by considering only the values at the cutoff, $f_o(\overline{x})$ and $\sigma^2(\overline{x})$. Hence, the benefits in using fixed-h asymptotics are expected to be larger when the density of X and the conditional variance change markedly inside the bandwidths around the cutoff, i.e., heteroskedasticity inside the bandwidth could lead to poor performance by the small-h variance approximation relative to fixed-h.

It is relevant to note that both fixed-h and small-h asymptotic approximations are based on the same estimator for $\widehat{\alpha}$. For a given bandwidth the bias present in the estimate is set. Small-h asymptotics may lead one to ignore the bias by arguing to have chosen the bandwidth to "undersmooth". However, once a bandwidth is chosen the bias is given and should not be ignored.

To clarify the intuition on the results in theorem 6, it is interesting to analyze the special case of the Nadaraya-Watson estimator.

Case 4 *For the* Nadaraya-Watson *estimator case, we have*

$$\sqrt{nh}(\widehat{\alpha} - \alpha^*_{NW}) \xrightarrow{d} N\left(0, V_{NW}\right)$$

where

$$\begin{aligned}
V_{NW} &= \frac{\int_0^\infty k^2(u)\,\sigma^{2+}(\overline{x}+uh) f_o(\overline{x}+uh)du}{\left(\int_0^\infty k(u)\, f_o(\overline{x}+uh)du\right)^2} + \frac{\int_0^\infty k^2(u)\,\sigma^{2-}(\overline{x}-uh) f_o(\overline{x}-uh)du}{\left(\int_0^\infty k(u)\, f_o(\overline{x}-uh)du\right)^2} \\
\alpha^*_{NW} &= \alpha + B_{NW} \\
B_{NW} &= \frac{\int_0^\infty k(u)\, m(\overline{x}+uh) f_o(\overline{x}+uh)du}{\int_0^\infty k(u)\, f_o(\overline{x}+uh)du} - \frac{\int_0^\infty k(u)\, m(\overline{x}-uh) f_o(\overline{x}-uh)du}{\int_0^\infty k(u)\, f_o(\overline{x}-uh)du}
\end{aligned}$$

If the rectangular kernel is used, the asymptotic variance and bias simplify to

$$
\begin{aligned}
V_{NW} &= \frac{\int_0^\infty \sigma^{2+}(\overline{x}+uh) f_o(\overline{x}+uh)du}{\left(\int_0^\infty f_o(\overline{x}+uh)du\right)^2} + \frac{\int_0^\infty \sigma^{2-}(\overline{x}-uh) f_o(\overline{x}-uh)du}{\left(\int_0^\infty f_o(\overline{x}-uh)du\right)^2} \\
B_{NW} &= \frac{\int_0^\infty m(\overline{x}+uh) f_o(\overline{x}+uh)du}{\int_0^\infty f_o(\overline{x}+uh)du} - \frac{\int_0^\infty m(\overline{x}-uh) f_o(\overline{x}-uh)du}{\int_0^\infty f_o(\overline{x}-uh)du}
\end{aligned}
$$

The asymptotic variance is given by a weighted average of the conditional variance of Y above and below the cutoff, and that the asymptotic bias is simply the difference in the (local) averages of $m(x)$ above and below the cutoff, i.e., the difference in outcome that would have arisen even in the absence of treatment.

Intuitively, it is interesting to draw a parallel of the results in theorem 6 with the issue of model misspecification. The problem of estimating the ATE at the cutoff discussed here can be seen as one of correctly estimating $E[Y \mid X]$ on both sides of the cutoff. In this sense, the local polynomial estimator is a polynomial approximation to the unknown conditional expectation inside the bandwidth on each side, not different from standard parametric methods. By choosing a relatively small bandwidth we are fitting the conditional expectation on a restricted support and, hence, expect a polynomial of order p to produce a better fit than if we were trying to fit $E[Y \mid X]$ globally, this is the benefit associated with a nonparametric approach, since it allows the conditional expectation to be unrestricted *outside* the bandwidth.

Clearly, one does not expect the conditional expectation in the bandwidth to be completely described by a polynomial of the chosen order p, so we can draw some intuition by looking at the asymptotic results in Theorem 6 as those arising from potentially misspecified models (White, 1982, 1996). The estimator converges to α^* which is not equal to the parameter of interest but still provides relevant information about the population.

2.5.1 Fuzzy Regression Discontinuity Design

In the Fuzzy RD design the estimator of the parameter of interest is given by the ratio

$$\widehat{\tau}_F = \frac{\widehat{\alpha}}{\widehat{\theta}}$$

where $\widehat{\alpha}$ is any of the estimators described in the previous section and $\widehat{\theta}$ is the estimator for the change in the probability of being in the treated group at the cutoff. Note that $\widehat{\theta}$ is obtained by using the estimators described above with the treatment assignment variable, D_i, as the dependent variable.

To obtain the asymptotic distribution of the fuzzy RD estimator, the delta method can be used, similarly to the result in Porter (2003).

Theorem 7 *If*

$$\begin{pmatrix} \sqrt{nh}(\widehat{\alpha} - \alpha^*) \\ \sqrt{nh}(\widehat{\theta} - \theta^*) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} V_\alpha & C_{\alpha\theta} \\ C_{\alpha\theta} & V_\theta \end{bmatrix} \right)$$

then

$$\sqrt{nh}\left(\frac{\widehat{\alpha}}{\widehat{\theta}} - \frac{\alpha^*}{\theta^*} \right) \xrightarrow{d} N\left(0, \frac{1}{\theta^{*2}} V_\alpha - 2\frac{\alpha^*}{\theta^{*3}} C_{\alpha\theta} + \frac{\alpha^{*2}}{\theta^{*4}} V_\theta \right)$$

where $\alpha^* = \alpha + B_\alpha$, $\theta^* = \theta + B_\theta$ and B_α and B_θ are the bias terms for the estimators as defined in theorem 6 for local polynomial estimators.

The proof of the proposition follows directly from the Delta Method and is omitted. The condition of multivariate normality required in this proposition follows from usual multivariate central limit theorem using a Cramer-Wold device (James, 2004; Pagan and Ullah, 1999). Note that,

$$\begin{aligned} \frac{\alpha^*}{\theta^*} &= \frac{\alpha + B_\alpha}{\theta + B_\theta} \\ &= \frac{\alpha + B_\alpha}{\theta} \frac{\theta}{\theta + B_\theta} \\ &= \frac{\alpha}{\theta} \frac{\theta}{\theta + B_\theta} + \frac{B_\alpha}{\theta} \frac{\theta}{\theta + B_\theta} \end{aligned} \tag{2.6}$$

for given values of α and θ, if $|\theta| < |\theta + B_\theta|$ then $0 \leq \left| \frac{\theta}{\theta + B_\theta} \right| \leq 1$. Clearly, if there is no bias in the estimate for α or θ, i.e., $B_\alpha = 0$ and $B_\theta = 0$, the fuzzy design RD estimator will

be consistent for the true treatment effect. If $B_\alpha = 0$ and $B_\theta \neq 0$, the estimator will suffer an attenuation bias and tests for the null hypotheses that the treatment is unimportant will be conservative. If $B_\alpha \neq 0$ and $B_\theta = 0$, the estimator's bias is similar to the one seem for the sharp RD design, only being scaled by $\frac{1}{\theta}$. Finally, if $B_\alpha \neq 0$ and $B_\theta \neq 0$, any increase in B_α increases the bias in the ATE estimator but there will be a trade-off regarding the size of B_θ since its impact in the first and second terms will be in opposite directions.

All the terms that appear in the asymptotic distribution above, except for $C_{\alpha\theta}$, can be obtained from theorem 6 by using local polynomial estimators discussed in section 2.3. It is necessary to specify $C_{\alpha\theta}$ in order to obtain the asymptotic distribution of the estimator in the fuzzy RD design.

Theorem 8 *Suppose $\sigma_{\varepsilon\eta} = E[\varepsilon\eta \mid X = x]$ is continuous for $x \neq \overline{x}$, $x \in \aleph$ and the left and right-hand limits at $\overline{x}$ exist. If $\widehat{\alpha}$ and $\widehat{\theta}$ are the local polynomial estimators and the conditions of theorem 6 hold for both estimators, then*

$$C_{\alpha\theta} = e_1' \left[\left(\Gamma_+^*\right)^{-1} \Delta_+^\rho \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^\rho \left(\Gamma_-^*\right)^{-1} \right] e_1$$

where $\Delta_{+(-)}^\rho = \begin{bmatrix} \rho_0^{+(-)} & \cdots & \rho_p^{+(-)} \\ \vdots & \ddots & \vdots \\ \rho_p^{+(-)} & \cdots & \rho_{2p}^{+(-)} \end{bmatrix}$,

$\rho_j^+ = \int_0^\infty k^2(u)\, u^j \sigma_{\varepsilon\eta}(\overline{x} + uh) f_o(\overline{x} + uh) du$,

$\rho_j^- = (-1)^j \int_0^\infty k^2(u)\, u^j \sigma_{\varepsilon\eta}(\overline{x} - uh) f_o(\overline{x} - uh) du$, Γ_+^* *and* Γ_-^* *are defined as in previous Corollaries.*

As $h \to 0$, the standard small-h asymptotic covariance is the same as the one in theorem 8, as one would expect.

Corollary 4 *Letting $h \longrightarrow 0$ in the expressions of theorem 8, then the asymptotic covariance, $C_{\alpha\theta}$, obtained by fixed-h (theorem 8) and small-h (Porter, 2003) are the same:*

$$C_{\alpha\theta} = \frac{\sigma_{\varepsilon\eta}^+(\overline{x}) + \sigma_{\varepsilon\eta}^-(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1$$

Also, a result similar to the corollary 3 is readily available.

Corollary 5 *Under the assumptions in theorem 8, if $h > 0$, and in the bandwidth around the cutoff, $f_o(x)$ and $\sigma_{\varepsilon\eta}(x)$ are constant, then the asymptotic covariance, $C_{\alpha\theta}$, obtained by fixed-h (theorem 8) and small-h (Porter, 2003) are the same.*

In the case of the Nadaraya-Watson estimator, the asymptotic covariance simplifies in similar fashion to the asymptotic variance in formula (2.6) and provides intuition about the refinements obtained by the fixed-h asymptotic distribution relative to small-h. Those improvements arise from incorporating the behavior of $\sigma_{\varepsilon\eta}(x)$ and $f_o(x)$ in the range around the cutoff while small-h does not.

Case 5 *For the* Nadaraya-Watson *estimator we have*

$$C_{\alpha\theta} = \frac{\int_0^\infty k^2(u)\,\sigma_{\varepsilon\eta}(\overline{x}+uh) f_o(\overline{x}+uh)du}{\left(\int_0^\infty k(u)\, f_o(\overline{x}+uh)du\right)^2} + \frac{\int_0^\infty k^2(u)\,\sigma_{\varepsilon\eta}(\overline{x}-uh) f_o(\overline{x}-uh)du}{\left(\int_0^\infty k(u)\, f_o(\overline{x}-uh)du\right)^2}$$

If the rectangular kernel is used $C_{\alpha\theta}$ simplifies to

$$C_{\alpha\theta} = \frac{\int_0^\infty \sigma_{\varepsilon\eta}(\overline{x}+uh) f_o(\overline{x}+uh)du}{\left(\int_0^\infty f_o(\overline{x}+uh)du\right)^2} + \frac{\int_0^\infty \sigma_{\varepsilon\eta}(\overline{x}-uh) f_o(\overline{x}-uh)du}{\left(\int_0^\infty f_o(\overline{x}-uh)du\right)^2}$$

2.6 Variance Estimators

To be able to perform inference about α using the information in a given sample, appropriate estimates for the unknown terms in the asymptotic variance formulas from theorem 6 are necessary. Note that the components of the asymptotic variance of $\sqrt{nh}(\widehat{\alpha}_p - \alpha_p^*)$ can be written as expectations of population quantities and estimated using sample analogues. We have

$$
\begin{aligned}
\gamma_j^+ &= E\left[h^{-1}k\left(\frac{\overline{x}-x}{h}\right)\left(\frac{\overline{x}-x}{h}\right)^j d\right], \\
\gamma_j^- &= E\left[h^{-1}k\left(\frac{\overline{x}-x}{h}\right)\left(\frac{\overline{x}-x}{h}\right)^j (1-d)\right], \\
\delta_j^+ &= E\left[h^{-1}k\left(\frac{\overline{x}-x}{h}\right)^2\left(\frac{\overline{x}-x}{h}\right)^j d\varepsilon^2\right], \\
\delta_j^- &= E\left[h^{-1}k\left(\frac{\overline{x}-x}{h}\right)^2\left(\frac{\overline{x}-x}{h}\right)^j (1-d)\varepsilon^2\right]
\end{aligned}
$$

Then, define the sample analog estimators of those quantities as

$$
\begin{aligned}
\widehat{\gamma}_j^+ &= (nh)^{-1}\sum_{i=1}^n k\left(\frac{\overline{x}-x_i}{h}\right)\left(\frac{\overline{x}-x_i}{h}\right)^j d_i, \\
\widehat{\gamma}_j^- &= (nh)^{-1}\sum_{i=1}^n k\left(\frac{\overline{x}-x_i}{h}\right)\left(\frac{\overline{x}-x_i}{h}\right)^j (1-d_i), \\
\widehat{\delta}_j^+ &= (nh)^{-1}\sum_{i=1}^n k\left(\frac{\overline{x}-x_i}{h}\right)^2\left(\frac{\overline{x}-x_i}{h}\right)^j d_i\widehat{\varepsilon}_i^2, \\
\widehat{\delta}_j^- &= (nh)^{-1}\sum_{i=1}^n k\left(\frac{\overline{x}-x_i}{h}\right)^2\left(\frac{\overline{x}-x_i}{h}\right)^j (1-d_i)\widehat{\varepsilon}_i^2;
\end{aligned}
$$

which are consistent by standard arguments based on the Law of Large Numbers. The residuals used in these estimators will depend on the order of the local polynomial used to estimate the Average Treatment Effect of interest and are given by

$$
\begin{aligned}
\widehat{\varepsilon}_i = y_i &- d_i\left(\widehat{\alpha}_{p+} + \widehat{\beta}_{1,p+}(x_i-\overline{x}) + ... + \widehat{\beta}_{p,p+}(x_i-\overline{x})^p\right) \\
&- (1-d_i)\left(\widehat{\alpha}_{p-} + \widehat{\beta}_{1,p-}(x_i-\overline{x}) + ... + \widehat{\beta}_{p,p-}(x_i-\overline{x})^p\right)
\end{aligned}
$$

Even though these estimators requires the calculation of $4(2p+1)$ terms[6] to obtain the plug-in estimator of the fixed-h variance-covariance matrix,

$$
\left[\left(\widehat{\Gamma}_+^*\right)^{-1}\widehat{\Delta}_+^*\left(\widehat{\Gamma}_+^*\right)^{-1} + \left(\widehat{\Gamma}_-^*\right)^{-1}\widehat{\Delta}_-^*\left(\widehat{\Gamma}_-^*\right)^{-1}\right], \tag{2.7}
$$

[6] In fact, since we are interested only on the estimate for the ATE, α, one can potentially only estimate the terms of both matrices that show up at the element at the $[1,1]$ position of the variance-covariance matrix for the estimators.

these are very simple averages of the data and kernel weights.

Porter (2003) suggests an estimator for the variance of $\widehat{\alpha}$ using the small-h approximation in corollary 2 which requires only the estimation of the conditional variance of the errors at the cutoff approaching both from right and left and the density of x at the cutoff.[7]

$$\widehat{\sigma}^{2+}(\overline{x}) = \frac{(nh)^{-1}\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right) d_i \widehat{\varepsilon}_i^2}{\frac{1}{2}\widehat{f}_O(\overline{x})}, \tag{2.8}$$

$$\widehat{\sigma}^{2-}(\overline{x}) = \frac{(nh)^{-1}\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right) (1-d_i) \widehat{\varepsilon}_i^2}{\frac{1}{2}\widehat{f}_O(\overline{x})}, \tag{2.9}$$

$$\widehat{f}_O(\overline{x}) = (nh)^{-1}\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right), \tag{2.10}$$

then,

$$\frac{\widehat{\sigma}^{2+}(\overline{x}) + \widehat{\sigma}^{2-}(\overline{x})}{\widehat{f}_O(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \tag{2.11}$$

is the estimator for the asymptotic variance matrix.

This small-h variance estimator avoids estimating each component of the matrices by assuming $h \to 0$, which is similar to assuming that $f_O(x)$ and $\sigma^2(x)$ are constant in the bandwidth around the cutoff as shown in corollary 3. The matrix $\Gamma^{-1}\Delta\Gamma^{-1}$ can be calculated directly because it is a deterministic function of the kernel. A drawback of the variance estimator in formula (2.11) is the need to estimate $f_O(\overline{x})$, which is not necessary if one uses the fixed-h variance estimator in formula (2.7). To obtain $\widehat{f}_O(\overline{x})$ we need to choose a kernel and a bandwidth for the density estimator, increasing the number of tuning parameters to be chosen. A natural choice would be both the kernel and bandwidth used in the estimation of the parameter of interest. In section 2.7, I present evidence that using the same bandwidth not only saves one the trouble of choosing another bandwidth, but also provides more reliable inference than choosing a bandwidth that differs from the one used to estimate τ.

[7] The estimator presented in formula (2.11) is not exactly the one presented in Porter (2003). He never suggested a specific estimator $\widehat{f}_O(\overline{x})$, so I chose the standard Rosenblatt-Parzen kernel estimator for $f_O(\overline{x})$ presented in Pagan and Ullah (1999).

For the Nadaraya-Watson estimator, the variance estimator simplifies greatly.

Case 6 *The fixed-h estimator of the asymptotic variance for the* Nadaraya-Watson *estimator is given by*

$$\begin{aligned}
&\frac{(nh)^{-1}\sum_{i=1}^{n} k^2\left(\frac{\overline{x}-x_i}{h}\right) d_i\widehat{\varepsilon}_i^2}{\left((nh)^{-1}\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right) d_i\right)^2} + \frac{(nh)^{-1}\sum_{i=1}^{n} k^2\left(\frac{\overline{x}-x_i}{h}\right)(1-d_i)\widehat{\varepsilon}_i^2}{\left((nh)^{-1}\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right)(1-d_i)\right)^2} \\
= \quad &\frac{(nh)\sum_{i=1}^{n} k^2\left(\frac{\overline{x}-x_i}{h}\right) d_i\widehat{\varepsilon}_i^2}{\left(\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right) d_i\right)^2} + \frac{(nh)\sum_{i=1}^{n} k^2\left(\frac{\overline{x}-x_i}{h}\right)(1-d_i)\widehat{\varepsilon}_i^2}{\left(\sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right)(1-d_i)\right)^2}
\end{aligned}$$

where n_l and n_u are the number of observations used below and above the cutoff, respectively.

Let $\bar{\widehat{\varepsilon}}_u^2 = n_u^{-1}\sum_{\overline{x}\le x_i\le \overline{x}+h}\widehat{\varepsilon}_i^2$ and $\bar{\widehat{\varepsilon}}_l^2 = n_l^{-1}\sum_{\overline{x}-h\le x_i\le \overline{x}}\widehat{\varepsilon}_i^2$. For the case of the rectangular kernel, equation (2.12) simplifies to

$$\frac{nh}{n_u}\bar{\widehat{\varepsilon}}_u^2 + \frac{nh}{n_l}\bar{\widehat{\varepsilon}}_l^2$$

and if $n_u = n_l$, simplifies further to

$$\frac{2nh}{n_u+n_l}\left(\bar{\widehat{\varepsilon}}_u^2 + \bar{\widehat{\varepsilon}}_l^2\right) \tag{2.12}$$

which is the estimator proposed for the asymptotic variance by Imbens and Lemieux (2008) in the local linear case, adapted for the Nadaraya-Watson Estimator.

Imbens and Lemieux (2008) propose the plug-in estimator of formula (2.12) for $\frac{\sigma^{2+}(\overline{x})+\sigma^{2-}(\overline{x})}{f_O(\overline{x})}$ and obtain their estimate for the asymptotic variance of the local linear estimator by scaling it by $e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1$. Note that $e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1$ equals 4 for the local linear estimator and 1 for the Nadaraya-Watson estimator. If higher polynomial orders are used in the estimator, the only change in the formula for the variance estimator is the scaling term.

In fact, both small-h (Porter, 2003; Imbens and Lemieux, 2008) variance estimators are based on an estimate for $\frac{\sigma^{2+}(\overline{x})+\sigma^{2-}(\overline{x})}{f_O(\overline{x})}$ and a scaling parameter that depends on the order of the polynomial and kernel used on the estimation of the parameter of interest.

In section 2.7 I present simulation evidence on test coverage using the variance estimators in formulas (2.7) and (2.11).

2.7 Simulations

This section presents simulation evidence displaying the empirical coverage of a standard t-statistic used to perform inference about the treatment effect of interest. All simulations are based on a Sharp RD design. The objective of the simulations is to evaluate the relative performance of the asymptotic distributions obtained by fixed-h and small-h regarding inference about the parameter of interest. As shown by corollary 3, assuming that $h \to 0$ provides asymptotic variance and bias approximations that are equal to the ones obtained by assuming that the probability density function of X and the conditional variance of the outcome are constant in the bandwidth around the cutoff. In fact, one would reasonably expect that the approximations should be similar for bandwidth values close to zero. Evidence from simulations presented below indicates that inference about the treatment effect of interest using the fixed-h theoretical approximation has better size behavior than the small-h approach, especially for larger bandwidths. Simulations using feasible estimators for the asymptotic variance indicate that tests based on fixed-h approach can improve over tests based on small-h, especially for larger bandwidths and when some forms of heteroskedasticity are present, however fixed-h can show slightly worse size behavior on tests that use small bandwidths.

Let X be the running variable, Y is the outcome variable for which we would like to estimate the average treatment effect at the cutoff and u be the error term. The details of the simulations are listed below.

- Sample size (n): 750
- Number of replications of the experiment: 2,000
- X is drawn from a $Normal(50, 100)$
- u is drawn from a $Normal(0, 1)$
- The cutoff for receiving treatment is $\overline{x} = 55$.

- The treatment variable is defined as $d_i = 1\left[x_i \geq \overline{x}\right]$
- The bandwidths range from 0.2 to 20, or from $\frac{1}{50}$ to 2 standard deviations of the running variable.

The empirical coverages presented are the fraction of rejections in the 2,000 repetitions for a test of size 5% (two-sided). I analyze 5 different data generating processes (DGPs) for the outcome variable, Y.

- DGP 1: $y_i = \mu + \alpha d_i + u_i$
- DGP 2: $y_i = \mu + \beta_1 x_i + \alpha d_i + u_i$
- DGP 3: $y_i = \mu + \beta_1 x_i + \beta_2 x_i^2 + \alpha d_i + u_i$
- DGP 4: $y_i = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \alpha d_i + u_i$
- DGP 5: $y_i = \exp\left(\frac{x}{20}\right) + \alpha d_i + u_i$

The true value of the parameters is $\mu = 3$, $\alpha = 10$, $\beta_1 = 0.5$, $\beta_2 = -0.005$, $\beta_3 = 00002$.

Two estimators for the parameter of interest α are used in the simulations, the first is the Nadaraya-Watson estimator presented in case 1 and, second, the widely used local linear estimator presented on case 3. For both estimators I use the Bartlett kernel[8].

The next subsection compares the test coverages obtained by the theoretical fixed-h and small-h asymptotic distributions derived in theorem 6 and corollary 2. The results obtained are infeasible since they depend on knowledge about $f_o(x)$, $\sigma^2(x)$ and $m(x)$ around the cutoff. Nevertheless, they demonstrate the theoretical improvements that fixed-h provides over small-h asymptotics. Subsection 2.7.2 compares the empirical coverages obtained with (feasible) estimated standard errors.

[8] Similar results were obtained when using a rectangular kernel and a truncated gaussian kernel (weighted so that the kernel integrates to one). They are available from the author upon request.

2.7.1 Simulations for Infeasible Inference

Nadaraya-Watson Estimator

The first set of figures[9] show the empirical coverage of the test for the (true) null hypotheses that $\alpha = 10$ when the Nadaraya-Watson estimator is used to obtain $\widehat{\alpha}$ and the (infeasible) variances for $\sqrt{nh}(\widehat{\alpha} - \alpha^*)$ presented in theorem 6 and corollary 2 are used.

For DGP 1, shown in figure B.1, the dependent variable does not depend on X directly and no bias is expected in the estimates for α using the Nadaraya-Watson estimator since the relationship between Y and X would be correctly captured even for larger bandwidths. As expected, the empirical size for the tests using fixed-h and small-h standard errors approximations behave very similarly for small bandwidths, but the differences increase with the bandwidth, suggesting that the fixed-h asymptotic distribution presented in theorem 6 provide a better approximation for the behavior of the estimator $\widehat{\alpha}$.

Figures B.2, B.3 and B.4 refer to DGP 2 in which Y is linearly related to X. In general, a large bias on the Nadaraya-Watson estimate is expected to arise for any bandwidth away from zero, since the estimator does not capture the relationship between Y and X. Hence, the estimates erroneously attribute differences in $m(x)$ above and below the cutoff to the treatment or policy. The steep decline on the empirical coverage in figure B.2 reflects the deleterious effects of the bias on the estimate and inference. This effect overwhelms the gains obtained by the better approximation for the variance of the estimates.

Nevertheless, since the DGP, $f_o(x)$ and $\sigma^2(x)$ are known I can obtain a bias approximation for this estimator using B_{NW} for fixed-h and $B_{small-h}$[10] with $p = 0$ for small-h. Figure B.3 shows the empirical coverage for the (infeasible) bias corrected test.

To better understand the role of the improved bias and standard error approximations separately, figure B.4 adds the empirical coverage that would be obtained if small-h bias and fixed-h's variance approximation were used to obtain the test-statistic and vice-versa. In this

[9] See appendix.

[10] Formulas (2.6) and (2.5), respectively.

case, the majority of the improvement is due to better (infeasible) approximation of the bias but the more precise calculations for the standard error provide non-trivial improvement.

The results for the remainder DGPs are qualitatively similar to the ones observed for DGP 2 and the graphs are omitted for brevity.[11]The "speed" with which the bias becomes a problem for inference varies depending on the DGP, but in general it becomes relevant for relatively small bandwidths.

To be fair, the comparisons between the Nadaraya-Watson bias approximations by fixed-h and small-h are not adequate for any true model in which the relationship of y and x could be described by a polynomial of order higher than linear (or order higher than $p+1$ in the local polynomial case) since the small-h approximation (Porter, 2003) describes the bias as being a function of the derivative of $m(x)$ limiting its accuracy to more complex functional forms. Nevertheless, from the simulations it is clear that the fixed-h approximation for the bias developed in theorem 6 better describes the asymptotic behavior of the estimator than the small-h bias given by corollary 2.

In summary, for all the simulations we have evidence that the asymptotic distribution of $\sqrt{nh}(\widehat{\alpha} - \alpha^*)$ is best described by the fixed-h approach developed in theorem 6, which explicitly considers the effects of the choice of bandwidth, than by the standard small-h asymptotic approximation, which assumes that $h \to 0$. The gains in the approximation are, as one would expect, larger for bandwidths further away from zero.

The importance of the asymptotic bias is substantial in the Nadaraya-Watson estimator's case and serves as cautionary evidence of the risks of dismissing the presence of bias in the estimation by arguing some "undersmoothing" in the choice of bandwidth. The bias can be greatly reduced by the use of local polynomial estimators (see next subsection).

[11]The graphs are available from the author upon request.

Local Polynomial Estimator

This section presents simulations in which the "empiricist's favorite" *local linear* estimator is used. This estimator has been a staple in the applied literature that uses RD designs and has been shown to have nice theoretical bias reduction properties.

For DGPs 1 and 2, no bias is expected in the estimates, since the local linear estimator correctly captures the relationship between Y and X inside any of the bandwidths used. In these cases, the local linear estimator correctly captures the DGP on the bandwidth and no bias arises.

The empirical coverage for DGP 2 is presented on figure B.5.[12] For smaller bandwidths, the use of small-h asymptotic variance generates similar empirical coverages to the ones obtained using the refined fixed-h variance approximation, but there is a significant decrease in the small-h coverage as the bandwidth increases, with the fixed-h approach outperforming the standard approximation on both DGPs 1 and 2. The improvement increases with the bandwidth size as one would expect.

For the remaining DGPs (X has a quadratic, cubic or exponential relationship to Y) both the asymptotic bias and variance approximations are relevant.[13] Figures B.6, B.7 and B.8 show the empirical coverage under DGP 3. Figure B.6 compares the test coverages using fixed-h versus small-h standard error approximations while ignoring the bias. It is clear that the general pattern observed till this moment remains, with fixed-h outperforming small-h, specially for larger bandwidths. Figure B.7 graphs the empirical coverage obtained by (infeasible) bias corrected tests and; figure B.8 separate the gains due to improvement in bias and variance refinements by adding the graphs of the "counterfactual" coverages

[12] The empirical coverages for DGP 1 and 2 are very similar. Only DGP 2's graph is reported here.

[13] As discussed in section 2.5 the local polynomial estimator could be analyzed under a parametric framework as the problem of estimating a (potentially) misspecified model. (White, 1982, 1996).

that calculate the test statistics using small-h bias and fixed-h variance approximation and vice-versa.[14]

For DGP 3 and 4, the bias do not seem to be greatly important, being successfully reduced by the use of the local linear estimator. Naturally then, the difference in bias approximations is not the main source of improvement in the empirical coverage as can be seem in figure B.8. This contrasts with the results for the same DGPs using the Nadaraya-Watson estimator. In that case, the bias had a large effect on the test's coverage and the majority of the gains associated with the use of fixed-h asymptotics were due to the bias' refinement.

For DGP 5, even though the bias is substantially mitigated by the use of local linear estimator, the empirical coverage is still significantly reduced for bandwidths larger than one standard error of the running variable ($h = 10$). As in the case of the Nadaraya-Watson estimator, the bias is an important component of the improvement in coverage and infeasible fixed-h bias correction correctly captures the bias and provides coverages that outperform the small-h approach.

As described in section 2.5, the refinements obtained by fixed-h are due to considering the behavior of $f_o(x)$ and $\sigma^2(x)$ inside the bandwidth, while small-h in effect ignores it. Note that the previous simulations were based on DGPs with homoskedastic errors. In the presence of heteroskedasticity, one would expect the improvements of the fixed-h approximation to be even more important.

To exemplify the distortions heteroskedasticity can create and how well the fixed-h asymptotic approximation can capture it, I have simulated empirical coverages for two heteroskedastic cases. For the first and second cases the standard error of the error term is defined as $\sigma(x) = 1 + 0.25x^2$ and $\sigma(x) = 1 + 0.25\,(x - \overline{x})^2$, respectively.[15]The (infeasible) tests based

[14]The graphs for GDPs 1, 4 and 5 are available from the author upon request.

[15]These examples are not intended to be representative of any empirical problem commonly faced in the applied literature and are intended to highlight the behavior of the fixed-h and small-h approximations in different heteroskedastic contexts.

on the fixed-h asymptotic approximation behave very well on both cases, highlighting its robustness. In the first case, for GDPs 2 and 3 (figures B.9 and B.10) , the small-h asymptotic approximation holds up relatively well, with a pattern similar to the one obtained in the homoskedastic case. In contrast, the second case in figures B.11 and B.12, the small-h based test has a steep decline[16]in coverage as the bandwidth increases, since it is not able to properly capture the effect of the heteroskedasticity in its asymptotic variance.

The difference of the small-h performance in the two cases can provide useful intuition to when its weaknesses can prove most relevant. The second case was designed to be a "worst case scenario" heteroskedasticity for small-h asymptotics since the conditional variance of the error at the cutoff, $\sigma^2(\overline{x})$ is at the extreme of the range of values assumed by $\sigma^2(x)$ in any given bandwidth. As can be seen from formula (2.4), the small-h and fixed-h asymptotic variances will be more similar the closer $\sigma^2(\overline{x})$ is from the "weighted average" of $\sigma^2(x)$ inside the bandwidths. In the first case, since $\sigma^2(\overline{x})$ is at the "middle" of the range for the conditional variance, the distortion produced by the heteroskedasticity is less marked than in the second case.

Some points are worth emphasizing. First, the general pattern is that, as expected, the empirical coverages obtained using the fixed-h results from theorem 6 outperform the small-h approximations, especially for larger bandwidths.

Second, both the asymptotic variance and asymptotic bias refined calculations improve the precision of inference relative to the standard approach. For smaller bandwidths small-h asymptotics provide similar coverages to the fixed-h approach, making clear that the core difference is due to the suitability of the restrictions imposed on $f_o(x)$, $\sigma^2(x)$ and $m(x)$ as the bandwidth increases (corollary 3). Naturally, those restrictions tend to be less realistic for larger bandwidths.

[16]Note the change in the scale of the y-axis, which now emcompasses the interval from 0 to 1.

Third, the use of the local linear estimator reduces significantly the coverage distortion[17] introduced by the bias present in the estimates, even when the linear approximation does not fully capture the local relationship between Y and X. This is in line with the results in Porter (2003) and justifies the reliance on the local linear estimator in applications.

Fourth, in the presence of heteroskedasticity, the small-h asymptotic approximation can have very poor performance, while the fixed-h approach still provides a reliable asymptotic approximation for the estimator's behavior.

2.7.2 Simulations for Feasible Inference

The simulations in the previous subsection have established that fixed-h asymptotic distribution approximations based on theorem 6 improve over the usual approximations in the literature, with better test size behavior by incorporating the choice of bandwidth by the researcher on the formulas for asymptotic variance and bias. In obtaining those results I used knowledge about the true DGP that is unavailable to the practitioner when implementing such estimators.

As described in section 2.6 natural estimators for the asymptotic variance of the parameters of interest are readily available and can be easily calculated for a given sample. This section presents simulations for the empirical coverage of the tests using two different estimated standard errors. The first one is based on the fixed-h asymptotic distribution and is given by formula (2.7). The second is proposed by Porter (2003) and described by formula (2.11).

Figures B.13 and B.14 present the empirical coverages for the tests based on the Nadaraya-Watson estimator for DGPs 1 and 2 described above.[18] It also includes the coverages obtained

[17] Note the change in the y-axis' range relatively to most cases presented in the Nadaraya-Watson simulations.

[18] The graphs for GDPs 3, 4 and 5 in the Nadaraya-Watson estimator case and 1, 4 and 5 in the local linear estimator case are available from the author upon request.

by the (infeasible) theoretical formulas[19]so one can compare the feasible coverage relative to the infeasible coverage in section 2.7.1. Figures B.15 and B.16 perform the same exercise using the local linear estimator for DGPs 2 and 3.

In figure B.13,[20]it is clear that even though both tests tend to overreject for small bandwidths, due to the small amount of data available in those cases, the coverages obtained by the fixed-h variance estimators provide meaningful improvements for larger bandwidths over the tests based on small-h variance estimators.

For DGPs 2 through 5, the presence of a strong bias overwhelms the tests even for relatively small bandwidths, as expected given the results from section 2.7.1. Nevertheless, the general pattern that the variance estimators based on formula (2.7) reflect the theoretical gains is maintained.

Similarly, when the local linear estimator is used, the empirical coverage obtained using the fixed-h standard errors' estimator incorporates the gains of improved inference described in the theory and shown in the infeasible simulations even for large bandwidths. As in the Nadaraya-Watson case, the tests overreject for very small bandwidths, probably due to the relative small amount of data available on these cases but hold very good size behavior for larger bandwidths.

Surprisingly, when the local linear estimator is used, the tests obtained using small-h standard error estimates behave very similarly to fixed-h ones especially for larger bandwidths, for which the results in section 2.7.1 would lead one to expect a significantly smaller coverage based on the small-h approach.

In fact, the small-h estimators provide tests with better size behavior for small (close

[19]Those lines, named fixed-h and small-h, are exactly the same presented in Figures B.1 and B.2, respectively.

[20]In figure B.13, the Nadaraya-Watson estimator is used for DGP 1, correctly capturing the relationship between Y and X around the cutoff. As discussed in section 2.7.1 there is no asymptotic bias in this case.

to zero) bandwidths, due to the fact that fixed-h standard errors require the estimation of several more terms for the components of the asymptotic variance, suffering more acutely with the restricted amount of data on the smaller bandwidths.

It seems that the small-h variance estimators are benefiting from the fact that, in practice, one cannot actually restrict the bandwidth too close to zero. Since the estimator for the standard errors sums across $\widehat{\varepsilon}_i^2$ it (partially) captures the behavior of $f_O(x)$ in the range around the cutoff that the small-h asymptotic approximation ignores by forcing $h \to 0$.

As discussed in section 2.7.1 the presence of heteroskedasticity can generate substantial problems for the size of tests using the theoretical small-h approximation. Figures B.17 through B.20 show simulations for the coverage of feasible tests using the fixed-h and small-h asymptotic variance estimators in the two heteroskedastic cases described in section 2.7.1.[21]

The results clearly show that, differently from the homoskedastic case, the fixed-h variance estimator produces better test sizes than the one based on the small-h approach. In the first case (figures B.17 and B.18), the use of the fixed-h estimator in formula (2.7) provides better empirical size for bandwidths larger than 5. It is important to emphasize that, even though small-h tests have better sizes for small bandwidths, both tend to overreject due to constrained data availability, and a researcher would be ill advised to use too small of a bandwidth.

In figures B.19 and B.20, the second case, the fixed-h variance estimator produces tests with coverage very close to the test's nominal size, while for the small-h the coverage rapidly increases to 1 as the bandwidth increases. Hence, there is evidence that heteroskedasticity can be accurately captured by tests based on fixed-h asymptotic approximations but small-h estimators can produce tests which perform substantially worse.

It is possible that these results are somewhat dependent on the DGPs chosen and, even though the empirical coverages obtained are similar using any of the asymptotic variance

[21] For the first case the standard error of the error term is defined as $\sigma(x) = 1 + 0.25x^2$, in the second case ("worst case scenario") it is given by $\sigma(x) = 1 + 0.25\,(x - \overline{x})^2$.

estimators in some cases, it seems the fixed-h standard error estimator is a "safer choice" for practitioners since it is based on a more robust asymptotic approximation and its computation is very easy once a kernel and bandwidth are chosen. Using standard error estimates based on small-h asymptotics can lead to serious size distortions for larger bandwidths, especially in the presence of heteroskedasticity, even in the absence of bias.

Furthermore, the fixed-h variance estimator has the advantage of not requiring the estimation of $f_o(\overline{x})$. This entails the choice of (potentially different) kernel and bandwidth for $\widehat{f_o}(\overline{x})$. The additional choice of these two tuning parameters might significantly alter the empirical size of the tests performed about $\widehat{\tau}$ and depends on the discretion of the researcher.

To exemplify this issue, figures B.21 and B.22 show the simulated empirical coverages obtained by using the small-h variance estimator for DGPs 2 and 3[22] using the Bartlett kernel for five different scenarios. Each scenario differs by the choice of the bandwidth, h_f, used in formula (2.10) to obtain $\widehat{f_o}(\overline{x})$. The first reproduces the small-h result described above by choosing the same bandwidth used to estimate $\widehat{\tau}$, i.e., $h_f = h$, the other lines are the empirical coverages obtained by using bandwidth of 1, 5, 10 and 20[23] for $\widehat{f_o}(\overline{x})$ independent of the bandwidth used for $\widehat{\tau}$.

The choice of bandwidth on the estimation of $\widehat{f_o}(\overline{x})$ can have a relevant impact on the test coverages. Interestingly, choosing the same bandwidth as used in estimating the parameter of interest provides more stable empirical coverages for a wide range of h relative to the cases in which the bandwidths are different. The cautious practitioner using the small-h variance estimator would be well advised to choose the same bandwidth for both estimators.

One key problem (especially the ones with the Nadaraya-Watson estimator), is how to deal with the bias in practice. The bias is a main contributor for the divergence between the empirical and nominal sizes of the tests being performed. Adequate estimation of the bias based on the results on theorem 6 would require observing or estimating the counterfactual

[22] The graphs for GDPs 1, 4 and 5 are available from the author upon request.

[23] $\frac{1}{20}$, $\frac{1}{4}$, $\frac{1}{2}$ and 1 standard deviations of the running variable, respectively.

conditional expectation of y around the cutoff in the absence of treatment, which is not available for most cases where the RD design is relevant.[24]

2.8 Conclusion

The use of regression discontinuity designs to obtain estimates of treatment effect, τ, has been widely used in recent years by researchers in economics. Special attention has been given to the use of local polynomial estimators to obtain the ATE of interest.

The standard literature on RD designs (Hahn, Todd, and Van der Klaauw, 2001; Porter, 2003; Imbens and Lemieux, 2008) assumes that the bandwidth around the discontinuity, h, shrinks fast enough towards zero, $h \to 0$, to guarantee identification of the parameter of interest (small-h asymptotics).

This chapter derives, in the RD design context, a refined asymptotic distribution for the local polynomial estimators by treating h as fixed. This fixed-h asymptotics explicitly acknowledges the fact that a researcher has to choose bandwidth to implement the estimator and are usually bounded in their ability to reduced the bandwidth size by data availability constraints.

The fixed-h asymptotic distributions obtained are more precise and provide refined inference relative to asymptotic distributions based on small-h approach. The fixed-h asymptotic approximation provides more precise formulas for both bias and variance of the estimators of interest (theorem 6). The standard small-h asymptotic bias and variance can be obtained

[24]The small-h asymptotic bias approximation (Porter, 2003) lends itself for estimation, since estimates for $m^{(p+1)}$ above and below the cutoff can be obtained. However, the results in section 2.7.1 indicate that this would be a relatively poor approximation. Furthermore, to a large degree, bias reduction could be obtained by increasing the order of the local polynomial fitted above and below the cutoff, reducing the "misspecification" in the model (see section 2.5).

by allowing $h \to 0$ in the fixed-h distribution (corollary 2). Also, when $h > 0$ the small-h result for the variance of the estimators is equivalent to assume that the density of the running variable and the conditional variance of the dependent variable are constant around the cutoff (corollary 3). Similar results are shown for both sharp and fuzzy RD designs.

Simulations provide evidence that fixed-h asymptotic distributions more accurately describe the behavior for both bias and variances than the usual small-h results used in the literature. This is reflected on improved test size, specially when larger bandwidths are used.

Simple feasible estimators for the refined, fixed-h, standard errors are provided and shown to incorporate the theoretical gains of the improved approximations in simulations. These estimators are simple to implement and have the advantage of not requiring the estimation of the density of the running variable at the discontinuity. The fixed-h variance estimators can improve markedly over small-h estimators in the presence of heteroskedasticity and should be generally preferred. Simulations using heteroskedastic errors have provided evidence that feasible tests based on the fixed-h approach obtain better coverage, outperforming small-h starting at relatively small bandwidths.

Interestingly, in the case of the widely used local linear estimator with homoskedastic errors the variance estimators based on small-h asymptotics suggested in the literature produce well behaved tests with similar size performance to the fixed-h variance estimators, performing better than the standard theory would expect. In other words, when errors are homoskedastic (inside the bandwidth) the inability of the empiricist to mimic what theory suggests ends up improving the properties of the tests and its robustness to the choice of bandwidth, relative to what the theory that spawned those estimators would have provided.

The results indicate that the fixed-h standard error estimator is a "safer choice" for practitioners since it is based on a more robust asymptotic approximation and its computation is very easy once a kernel and bandwidth are chosen.

CHAPTER 3

Asymptotic Properties of Quantile Regression for Standard Stratified Samples

3.1 Introduction

Quantile Regression (QR) has been widely used in the social sciences in recent decades, and provides a useful characterization of the distributional features of variables in which the researcher is interested. In economics, for example, a very natural use of quantile regression has been to analyze the wage structure and potential differences in the determinants of the observed wages at different levels of the wage distribution, e.g., Albrecht et al. (2003); Buchinsky (1998, 2001); Machado and Mata (2005); Martins and Pereira (2004) and Melly (2005).

In those analyses it is very common to use datasets generated by stratified sampling. In standard stratified sampling (SS sampling) the population is divided into J mutually exclusive, exhaustive strata, and a random sample of size N_j is taken from stratum j. As described by Wooldridge (2001) when stratification is based on exogenous variables it

usually does not cause serious problems. The usual estimators that ignore stratification are consistent and asymptotically normal and the usual variance estimators are still valid.

When stratification is based on endogenous variables, the standard unweighted estimators are generally inconsistent. Wooldridge (2001) studies the asymptotic properties of general weighted M-estimators under SS sampling which will be consistent, asymptotically normal and provides estimators for standard errors of the parameters of interest that can be used to perform inference for general stratification. However, those results are not directly applicable to the quantile regression case due to the nonsmoothness in the objective function that provides the QR estimates.

This chapter fills that gap, extends the analysis to the quantile regression case, analyzes the asymptotic properties of the weighted QR estimates under general SS sampling, and provides consistent estimators for the standard errors that take the stratification of the data into account. Under exogenous stratification the usual unweighted QR estimators are still valid as well as its standard error estimates.

3.2 The Quantile Regression Population Problem

We are interested in estimating the conditional quantile function (CQF) of a random variable Y conditional on a vector of q explanatory variables X. This is defined by,

$$Q_\tau(Y \mid X) = \inf \{y : F_Y(y \mid X) \geq \tau\}$$

where $\tau \in (0, 1)$ indexes the τ^{th} quantile of the conditional distribution of Y. Let the CQF be described by a *known* function $g(\cdot)$ of the parameters and the explanatory variables,

$$Q_\tau(Y \mid X) = g\left(X, \beta_{\tau o}\right)$$

A special case of interest is given by the linear model[1]

[1] This formulation assumes the error term is additive and, hence, separable. For a

$$Y = X'\beta_{\tau_o} + \varepsilon$$

with $Q_\tau(\varepsilon \mid X) = 0$. Throughout this chapter I concentrate on the linear CQF, since it is the most widely used by practitioners and for ease of exposition. Nevertheless, the results presented are valid for a nonlinear, correctly specified CQF, $g(\cdot)$. In the population, β_{τ_o} solves the following problem

$$\min_{\beta_\tau \in \mathbf{B}} E\left[\rho_\tau\left(Y - X'\beta_\tau\right)\right] \tag{3.1}$$

where, $\rho_\tau(u) = (\tau - 1[u \leq 0])u$ and $\mathbf{B} \in \mathbb{R}^K$ is the parameter space.

Given a random sample from the population of size n, it is possible to obtain consistent estimates of β_o by a standard quantile regression (QR) estimator.

$$\min_{\beta_\tau \in \mathbf{B}} n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta_\tau) \tag{3.2}$$

Note that the minimization problem has the following first order conditions and sample analogue (Buchinsky, 1998)

$$\begin{aligned} E\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x\right] &= 0 \\ n^{-1}\sum_{i=1}^{n}\left(\tau - 1\left[y_i - x_i'\widehat{\beta_\tau} \leq 0\right]\right)x_i &= 0 \end{aligned} \tag{3.3}$$

Hence, we can frame this problem as a GMM estimator that uses as moment conditions the first order conditions of the QR problem that identify β_{τ_o}. Under random sampling, the standard QR procedures can be used to estimate β_{τ_o} and perform inference.

3.2.1 Quantile Regression under Stratified Sampling

Suppose our sample is obtained by a standard stratification scheme as formally described by Wooldridge (2001). Assume that the population is divided into J nonempty, mutually

treatment of the more general formulation with (potentially) non-separable ε see Powell (1991).

exclusive, and exhaustive strata, $W_1, W_2, ..., W_J$, where J is a finite integer. Let w denote a random variable having the population distribution of interest.

Definition 5 *Standard stratified sampling: For $j = 1, ..., J$, draw a random sample of size N_j from stratum j. For each j, denote this random sample by $\{w_{ij} : i = 1, 2, ..., N_j\}$.*

The strata sample sizes N_j are nonrandom. Therefore, the total sample size, $N = N_1 + ... + N_J$, is nonrandom. Notice that for a given j, $\{w_{ij} : i = 1, 2, ..., N_j\}$ is an independent, identically distributed (i.i.d.) sequence having the same distribution conditional on being part of a strata, $D(w|w \in W_j)$.

Then, one can rewrite the minimization problem and its moment conditions as

$$\begin{aligned} \min_{\beta_\tau \in \mathbf{B}} E\left[\rho_\tau(Y - X'\beta_\tau)\right] &= \min_{\beta_\tau \in \mathbf{B}} \sum_{j=1}^{J} Q_j E\left[\rho_\tau(Y - X'\beta_\tau)|w \in W_j\right] \\ E\left[\left(\tau - 1\left[y - x'\beta_{\tau o} \leq 0\right]\right)x\right] &= \sum_{j=1}^{J} Q_j E\left[\left(\tau - 1\left[y - x'\beta_{\tau o} \leq 0\right]\right)x|w \in W_j\right] = 0 \end{aligned} \tag{3.4}$$

where $Q_j = P(w \in W_j)$, $j = 1, ..., J$ and its sample analogue

$$\begin{aligned} \sum_{j=1}^{J} Q_j \left[\frac{1}{N_j}\sum_{i=1}^{N_j}\left(\tau - 1\left[y_{ij} - x'_{ij}\widehat{\beta}_\tau \leq 0\right]\right)x_{ij}\right] &= 0 \\ \frac{1}{N}\sum_{j=1}^{J}\frac{Q_j}{H_j}\left[\sum_{i=1}^{N_j}\left(\tau - 1\left[y_{ij} - x'_{ij}\widehat{\beta}_\tau \leq 0\right]\right)x_{ij}\right] &= 0 \end{aligned}$$

with $H_j = \frac{N_j}{N}$. This can be rewritten as,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{Q_j}{H_j}\left(\tau - 1\left[y_i - x'_i\widehat{\beta}_\tau \leq 0\right]\right)x_i = 0 \tag{3.5}$$

This is the empirical moment condition that is used to estimate the parameters of interest, defining the *weighted quantile regression* estimator. This estimator is consistent for the parameters of interest under standard stratified sampling (Wooldridge, 2001, theorem 3.1).[2]

[2] As a minor point note that if one wants to implement the weighted estimator by applying a standard quantile regression to weighted data, the weights for each observation

The asymptotic distribution of the weighted quantile regression estimator can be obtained by a direct application of Newey and McFadden (1994) Theorem 7.1,with careful consideration to the formulation of $Var\left[g(\beta_{\tau_o})\right]$ due to the stratified nature of the data.

Corollary 6 *If the conditions in Newey and McFadden (1994) theorem 7.1 hold,* $\left\{w_{ij} : i = 1, 2, ..., N_j; j = 1, 2, ..., J\right\}$ *follows the standard stratified sample scheme,* $\frac{N_j}{N} \to \overline{H}_j > 0$ *as* $N \to \infty$ *for each* j*. Then* $\sqrt{N}\left(\widehat{\beta}_\tau - \beta_{\tau_o}\right) \overset{a}{\sim} N\left(0, A_w^{-1} B_w A_w^{-1}\right)$*, where*

$$A_w = E\left[f_{y|x}(g\left(x, \beta_{\tau_o}\right))\overset{\bullet}{g}\overset{\bullet}{g}'\right]$$

and

$$B_w = \sum_{j=1}^{J} \frac{Q_j^2}{\overline{H}_j} Var\left[\left(\tau - 1\left[y - g\left(x, \beta_{\tau_o}\right) \leq 0\right]\right)\overset{\bullet}{g}|w \in W_j\right]$$

where $\overset{\bullet}{g} \equiv \frac{\partial g(x,\beta)}{\partial \beta}\mid_{\beta=\beta_{\tau_o}}$

In the special case of the linear CQF, $g\left(X, \beta\right) = x'\beta$*,*

$$A_w = E\left[f_{y|x}\left(x'\beta_{\tau_o}\right)xx'\right]$$

and

$$B_w = \sum_{j=1}^{J} \frac{Q_j^2}{\overline{H}_j} Var\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x|w \in W_j\right]$$

Corollary 6 provides a general form for the asymptotic variance of the quantile regression estimators under standard stratification. Two main points are relevant when analyzing B_w. The first, which is general to the standard stratification literature, is that we cannot replace $Var\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau_o}\right) \leq 0\right]\right)\overset{\bullet}{g}_i|w \in W_j\right]$ by the outer product of the score as in the random sampling case because in general $E\left[\left(\tau - 1\left[y - x'\beta_{\tau_o} \leq 0\right]\right)x|w \in W_j\right] \neq 0$,

---

will be given by $\frac{Q_{j_i}}{H_{j_i}}$ instead of the $\left(\frac{Q_{j_i}}{H_{j_i}}\right)^{\frac{1}{2}}$ usually required when implementing least squares estimators and its variants.

as pointed out by Wooldridge (2001). It is also interesting to note that, differently from the standard results in quantile regression for random sampling, B_w does not simplify to $\tau(1-\tau)E[xx']$ in this case. That is due to the fact that the variance of the binary variable $1\left[y_{ij} - x'_{ij}\beta_{\tau o} \leq 0\right]$ is not necessarily the same for each stratum, in other words, $x'_{ij}\beta_{\tau o}$ will not represent the τ quantile in every stratum.

If we assume that

$$E\left[\left(\tau - 1\left[y - x'\beta_{\tau o} \leq 0\right]\right)x | w \in W_j\right] = 0, \text{ for } j = 0, 1, 2, \ldots, J. \tag{3.6}$$

then an alternative formula for B_w is available.

$$B_w = \sum_{j=1}^{J} \frac{Q_j^2}{\overline{H}_j} E\left[\left(\tau - 1\left[y - x'\beta_{\tau o} \leq 0\right]\right)^2 xx' | w \in W_j\right]$$

A sufficient condition for equation 3.6 is that the conditional distribution of Y is independent from strata, in which case $E\left[1\left[y - g\left(x, \beta_{\tau o}\right) \leq 0\right] | x, w \in W_j\right] = \tau$ for all j. Then

$$B_w = \tau(1-\tau)\sum_{j=1}^{J} \frac{Q_j^2}{\overline{H}_j} E\left[xx' | w \in W_j\right]$$

3.2.2 Quantile Regression Estimation under Exogenous Stratification

If we are modeling the quantiles of Y given X, and stratification is based solely on the conditioning variables X, stratification is said to be exogenous. Let the sample space for X be partitioned into J nonempty, mutually exclusive, and exhaustive strata, $\chi_1, \chi_2, ..., \chi_J$, where J is a finite integer, then we can analyze the effects of stratification under the framework of the previous section.

The weighted quantile regression estimator in equation 3.5 can be used both when the stratification is exogenous or not. Under exogenous stratification the *unweighted quantile*

regression estimator is also consistent (Wooldridge, 2001).[3] This estimator drops the weights $\frac{Q_j}{H_j}$, solving

$$\frac{1}{N}\sum_{i=1}^{N}\left(\tau - 1\left[y_i - x_i'\widehat{\beta}_\tau \leq 0\right]\right)x_i = \sum_{j=1}^{J} H_j \left[\frac{1}{N_j}\sum_{i=1}^{N_j}\left(\tau - 1\left[y_{ij} - x_{ij}'\widehat{\beta}_\tau \leq 0\right]\right)x_{ij}\right] = 0 \tag{3.7}$$

so that each stratum average is just weighted by its sample frequency, i.e., this is the usual QR estimator that would be used under random sampling.

Under exogenous stratification the unweighted quantile regression estimator has asymptotic distribution as described in the corollary below.

Corollary 7 *If the conditions in Newey and McFadden (1994) theorem 7.1 hold,* $\left\{w_{ij} : i = 1, 2, ..., N_j; j = 1, 2, ..., J\right\}$ *follows the standard stratified sample scheme, and, stratification is a deterministic function of* X, $\frac{N_j}{N} \to \overline{H}_j > 0$ *as* $N \to \infty$ *for each* j. *Then* $\sqrt{N}\left(\widehat{\beta}_\tau - \beta_{\tau_o}\right) \overset{a}{\sim} N\left(0, A_u^{-1}B_uA_u^{-1}\right)$, *where*

$$A_u = \sum_{j=1}^{J}\overline{H}_j E[f_{y|x}\left(g\left(x, \beta_{\tau_o}\right)\right)\overset{\bullet}{g}\overset{\bullet}{g}'|x \in \chi_j]$$

and

$$B_u = \tau(1-\tau)\sum_{j=1}^{J}\overline{H}_j E\left[\overset{\bullet}{g}\overset{\bullet}{g}'|x \in \chi_j\right]$$

where $\overset{\bullet}{g} \equiv \frac{\partial g(x,\beta)}{\partial \beta}\,|_{\beta=\beta_{\tau_o}}$

[3] As emphasized in Wooldridge (2001) the consistency of the unweighted estimator will hold under exogenous stratification if the conditional quantiles of Y are correctly specified. The unweighted estimator solves

$$\min_{\beta_\tau \in \mathbf{B}}\left\{\sum_{j=1}^{J}\overline{H}_j E\left[\rho_\tau\left(Y - X'\beta_\tau\right)|x \in \chi_j\right]\right\}$$

which solution is not necessarily β_{τ_o} under misspecification.

In the special case of the linear CQF, $g(X,\beta) = x'\beta$,

$$A_u = \sum_{j=1}^{J} \overline{H}_j E[f_{y|x}\left(x'\widehat{\beta}_{\tau o}\right) xx'|x \in \chi_j]$$

and

$$B_u = \tau(1-\tau) \sum_{j=1}^{J} \overline{H}_j E\left[xx'|x \in \chi_j\right]$$

3.2.3 Sequence of Quantile Regressions

The discussion above considers only the estimation of a single quantile regression for a given value τ but one might be interested in estimating several quantile regressions for diverse points of the conditional distribution of Y. As emphasized by Buchinsky (1998), because the coefficients are estimated utilizing the same data with different weighting schemes, the estimators will be correlated.

Consider that we are still interested in estimating the linear conditional quantile function for p separate quantiles, τ,

$$Y = X'\beta_{\tau_r} + \varepsilon_{\tau_r}$$

and that $Q_{\tau r}(\varepsilon_{\tau r} \mid X) = 0$ for $r = 1, \ldots, p$. Also, let $0 < \tau_1 < \tau_2 < \ldots < \tau_p < 1$, and $\beta'_\tau = (\beta'_{\tau_1}, \beta'_{\tau_2}, \ldots, \beta'_{\tau_p})$ For each τ_r define

$$\psi_r(y, x, \beta_{\tau_r}) \equiv \left(\tau_r - 1\left[y - x'\beta_{\tau_r} \leq 0\right]\right) x$$

and $\Psi\left(y, x, \beta_{\tau_1}, \beta_{\tau_2}, \ldots, \beta_{\tau_p}\right)' = \left[\psi_1(y, x, \beta_{\tau_1})', \psi_2(y, x, \beta_{\tau_2})', \ldots, \psi_p(y, x, \beta_{\tau_p})'\right]$.
Hence

$$E\left[\Psi\left(y, x, \beta_{\tau_1}, \beta_{\tau_2}, \ldots, \beta_{\tau_p}\right)\right] = \sum_{j=1}^{J} Q_j E\left[\Psi\left(y, x, \beta_{\tau_1}, \beta_{\tau_2}, \ldots, \beta_{\tau_p}\right) | w \in W_j\right] = 0$$

By the analogy principle, the estimator $\widehat{\beta}_\tau$ for β_τ solves

$$\frac{1}{N}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j}\Psi\left(y_{ij},x_{ij},\widehat{\beta}_{\tau_1},\widehat{\beta}_{\tau_2},\ldots,\widehat{\beta}_{\tau p}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{Q_{j_i}}{H_{j_i}}\Psi\left(y_i,x_i,\widehat{\beta}_{\tau_1},\widehat{\beta}_{\tau_2},\ldots,\widehat{\beta}_{\tau p}\right)=0$$

which can be solved separately if no cross-quantile restrictions are imposed on $\widehat{\beta}_{\tau_1},\widehat{\beta}_{\tau_2},\ldots,\widehat{\beta}_{\tau p}$ or simultaneously otherwise. Then, we can show that $\widehat{\beta}'_\tau$ follows an asymptotic multivariate normal distribution as well.

Corollary 8 *If the conditions in Newey and McFadden (1994) theorem 7.1 hold,* $\left\{w_{ij}:i=1,2,...,N_j;j=1,2,...,J\right\}$ *follows the standard stratified sample scheme,* $\frac{N_j}{N}\to\overline{H}_j>0$ *as* $N\to\infty$ *for each* j. *Then* $\sqrt{N}\left(\widehat{\beta}_\tau-\beta_\tau\right)\overset{a}{\sim}N\left(0,\Lambda_\tau\right)$, *where* $\Lambda_\tau=\left\{\Lambda_{\tau_{l,k}}\right\}_{l,k=1,...p}$ *with typical element defined as*

$$\begin{aligned}\Lambda_{\tau_{l,k}} &= E\left[f_{y|x}(g\left(x,\beta_{\tau_l}\right))\overset{\bullet}{g}_l\overset{\bullet}{g}_l'\right]^{-1}\times\\ &\quad\times\left[\sum_{j=1}^{J}\frac{Q_j^2}{\overline{H}_j}Cov\left[\psi_l(y,x,\beta_{\tau_l}),\psi_k(y,x,\beta_{\tau_k})|w\in W_j\right]\right]\times\\ &\quad\times E\left[f_{y|x}(g\left(x,\beta_{\tau_k}\right))\overset{\bullet}{g}_k\overset{\bullet}{g}_k'\right]^{-1}\end{aligned}$$

where, in this case, $\psi_r(y,x,\beta_{\tau_r})=\left(\tau_r-1\left[y-g\left(x,\beta_{\tau_r}\right)\leq 0\right]\right)\overset{\bullet}{g}_r$ *for* $r=l,k$ *and* $\overset{\bullet}{g}_l\equiv\frac{\partial g(x,\beta)}{\partial\beta}\mid_{\beta=\beta_{\tau_l}}$.

In the special case of the linear CQF, $g\left(X,\beta\right)=x'\beta$

$$\begin{aligned}\Lambda_{\tau_{l,k}} &= E\left[f_{y|x}\left(x'\beta_{\tau_l}\right)xx'\right]^{-1}\times\\ &\quad\times\left[\sum_{j=1}^{J}\frac{Q_j^2}{\overline{H}_j}Cov\left[\psi_l(y,x,\beta_{\tau_l}),\psi_k(y,x,\beta_{\tau_k})|w\in W_j\right]\right]\times\\ &\quad\times E\left[f_{y|x}\left(x'\beta_{\tau_k}\right)xx'\right]^{-1}\end{aligned}$$

with $\psi_r(y,x,\beta_{\tau_r})=\left(\tau_r-1\left[y-x\beta_{\tau_r}\leq 0\right]\right)x$ *for* $r=l,k$.

Once again, the notable difference of the result relative to the usual QR under random sampling is that the center term on $\Lambda_{\tau_{l,k}}$ does not simplify as neatly as in standard the

random sampling case, since the covariance of the binary variables $1\left[y_{ij}-x'_{ij}\beta_{\tau_l}\leq 0\right]$ and $1\left[y_{ij}-x'_{ij}\beta_{\tau_k}\leq 0\right]$ are not necessarily the same for each stratum.

3.3 Asymptotic Variance Estimation

To perform inference about the parameter's estimates, $\widehat{\beta}_\tau$, we need to obtain valid estimators for its asymptotic variance. From corollary 6 we have that for linear CQF

$$\begin{aligned} Var\left[\sqrt{N}\left(\widehat{\beta}_\tau-\beta_{\tau_o}\right)\right] &= A_w^{-1}B_wA_w^{-1} \\ &= E\left[f_{y|x}(x'\beta_{\tau_o})x'x\right]^{-1}\times \\ &\quad\times\left[\sum_{j=1}^{J}\frac{Q_j^2}{\overline{H}_j}Var\left[\left(\tau-1\left[y-x'\beta_{\tau_o}\leq 0\right]\right)x|w\in W_j\right]\right]\times \\ &\quad\times E\left[f_{y|x}(x'\beta_{\tau_o})x'x\right]^{-1} \end{aligned}$$

Natural estimators for A_w and B_w, as suggested by Wooldridge (2001), are given by

$$\begin{aligned} \widehat{A}_w &\equiv \sum_{j=1}^{J}Q_j\left(N_j^{-1}\sum_{i=1}^{N_j}\widehat{f}_{y|x,w\in W_j}(x'_{ij}\widehat{\beta}_\tau)x_{ij}x'_{ij}\right) \\ &= N^{-1}\sum_{i=1}^{N}\frac{Q_{j_i}}{H_{j_i}}\widehat{f}_{y|x,w\in W_j}(x'_i\widehat{\beta}_\tau)x_ix'_i \\ \widehat{B}_w &\equiv \sum_{j=1}^{J}\frac{Q_j^2}{H_j}\left(N_j^{-1}\sum_{i=1}^{N_j}\left[s_{ij}\left(\widehat{\beta}_\tau\right)-\overline{s_j\left(\widehat{\beta}_\tau\right)}\right]'\left(\left[s_{ij}\left(\widehat{\beta}_\tau\right)-\overline{s_j\left(\widehat{\beta}_\tau\right)}\right]\right)\right) \\ &= N^{-1}\sum_{j=1}^{J}\left(\frac{Q_j}{H_j}\right)^2\left(\sum_{i=1}^{N_j}\left[s_{ij}\left(\widehat{\beta}_\tau\right)-\overline{s_j\left(\widehat{\beta}_\tau\right)}\right]'\left[s_{ij}\left(\widehat{\beta}_\tau\right)-\overline{s_j\left(\widehat{\beta}_\tau\right)}\right]\right) \end{aligned}$$

where $s_{ij}\left(\widehat{\beta}_\tau\right)\equiv\left(\tau-1\left[y_{ij}-x'_{ij}\widehat{\beta}_\tau\leq 0\right]\right)x_{ij}$, $\overline{s_j\left(\widehat{\beta}_\tau\right)}=N_j^{-1}\sum\limits_{i=1}^{N_j}s_{ij}\left(\widehat{\beta}_\tau\right)$ and $\frac{Q_{j_i}}{H_{j_i}}=\frac{Q_j}{H_j}$ for the stratum j of x_i. In the more general, nonlinear framework, a similar estimator for the covariance matrix is given by replacing $s_{ij}\left(\widehat{\beta}_\tau\right)\equiv\left(\tau-1\left[y_{ij}-g\left(x_{ij},\widehat{\beta}_\tau\right)\leq 0\right]\right)\widehat{\bullet}{g}_i$

and the outer terms $\left[N^{-1}\sum_{i=1}^{N}\frac{Q_{j_i}}{H_{j_i}}\widehat{f}_{y|x,w\in W_j}\left(g\left(x_{ij},\widehat{\beta}_\tau\right)\right)\overset{\widehat{\bullet}}{g}_i\overset{\widehat{\bullet}}{g}_i'\right]^{-1}$ with $\overset{\widehat{\bullet}}{g}_i \equiv \frac{\partial g\left(x_{ij},\beta\right)}{\partial\beta}|_{\beta=\widehat{\beta}_\tau}$.

Then,

$$\widehat{Var}\left[\sqrt{N}\left(\widehat{\beta}_\tau-\beta_{\tau_o}\right)\right]=\widehat{A}_w^{-1}\widehat{B}_w\widehat{A}_w^{-1}$$

To estimate the out of diagonal terms $\Lambda_{\tau_{l,k}}$ when we are interested in performing inference about the parameters on a sequence of quantile regressions, we can use a similar approach for estimating the middle term by

$$\sum_{j=1}^{J}\frac{Q_j^2}{H_j}\left(N_j^{-1}\sum_{i=1}^{N_j}\left[s_{ij}\left(\widehat{\beta}_{\tau_l}\right)-\overline{s_j\left(\widehat{\beta}_{\tau_l}\right)}\right]'\left(\left[s_{ij}\left(\widehat{\beta}_{\tau_k}\right)-\overline{s_j\left(\widehat{\beta}_{\tau_k}\right)}\right]\right)\right)$$

and we can still use the same estimators, $\widehat{A}_w$, for the outer terms of $\Lambda_{\tau_{l,k}}$, noticing that they are based on the estimates for τ_l and τ_k, respectively.

Finally, under exogenous stratification and correct specification of the underlying CQF, the score will generally have a zero conditional mean for all X when evaluated at β_{τ_o}. Then the asymptotic variance estimator for the *unweighted quantile regression estimator* (with linear CQF) is given by:

$$\widehat{Var}\left[\sqrt{N}\left(\widehat{\beta}_\tau-\beta_{\tau_o}\right)\right]=\widehat{A}_u^{-1}\widehat{B}_u\widehat{A}_u^{-1}$$

with

$$\begin{aligned}\widehat{A}_u &= \frac{1}{N}\sum_{i=1}^{N}\widehat{f}_{y|x}(x_i'\widehat{\beta}_\tau)x_ix_i'\\ \widehat{B}_u &= \tau(1-\tau)\frac{1}{N}\sum_{i=1}^{N}x_ix_i'\end{aligned}$$

which are the usual variance matrix estimators from the quantile regression literature under random sampling as described by Buchinsky (1998) and Koenker (2005). This confirms that the usual quantile regression estimators and standard errors are valid when the sample

is exogenously stratified, this is the same result as obtained by Wooldridge (2001) for the general M-estimators with smooth objective functions.

A main issue, which is specific to quantile regression, is that we need to estimate $\widehat{f}_{y|x,w\in W_j}(x_i'\widehat{\beta}_\tau)$ taking in consideration the stratification. An intuitive approach is to take advantage of the fact that we have assumed random sampling for each stratum and apply a standard nonparametric density estimator for each stratum and simply plug in the formula above. For example, using the Rosenblatt-Parzen kernel estimator described in Pagan and Ullah (1999)

$$\widehat{f}_{y|x,w\in W_j}(x_{ij}'\widehat{\beta}_\tau) = \left(N_j h_{nj}\right)^{-1} \sum_{i=1}^{N_j} K\left(\frac{y_{ij} - x_{ij}'\widehat{\beta}_\tau}{h_{nj}}\right)$$

where $K(\cdot)$ is a kernel function and h_{nj} is a bandwidth parameter such that $h_{nj} \to 0$ and $\sqrt{N_j}h_{nj} \to \infty$. Another option is to bypass the estimation of $\widehat{f}_{y|x,w\in W_j}(x_i'\widehat{\beta}_\tau)$ itself and revert to the estimator for

$$A_{w_j} = E[f_{y|x,w\in W_j}\left(x'\beta_{\tau o}\right)x'x|w \in W_j]$$

that is referred to as the Powell Sandwich by Koenker (2005). This takes account of the fact that estimating A_{w_j} is just as estimating a matrix weighted density estimator (Koenker, 2005).

$$\widehat{A}_{w_j} = \left(N_j h_{nj}\right)^{-1} \sum_{i=1}^{N_j} K\left(\frac{y_{ij} - x_{ij}'\widehat{\beta}_\tau}{h_{nj}}\right) x_{ij}x_{ij}'$$

Then, we can obtain $\widehat{A}_w = \sum_{j=1}^{J} Q_j \widehat{A}_{w_j}$. Powell (1991) has shown that under some additional conditions regarding $f(\cdot)$, $\widehat{A}_{w_j}$ converges is probability to A_{w_j}. Two drawbacks of this estimator are the necessity of choosing kernels and bandwidths (which could be different for each stratum) and the fact that, in practice, the researcher might have some stratum for which only a small amount of data is available, reducing the confidence in the obtained

estimates for $\widehat{A}_{w_j}$.[4]

When the stratification is exogenous, we can take advantage of the fact that $f_{y|x,x\in\chi_j}(x'_{ij}\beta_\tau) = f_{y|x}(x'_i\beta_\tau)$ for all strata and obtain $\widehat{f}_{y|x}(x'_{ij}\widehat{\beta}_\tau)$ or $\widehat{A}_w$ directly as

$$\begin{aligned}\widehat{f}_{y|x}(x'_i\widehat{\beta}_\tau) &= (Nh_n)^{-1}\sum_{i=1}^{N} K\left(\frac{y_i - x'_i\widehat{\beta}_\tau}{h_n}\right)\\ \widehat{A}_w &= (Nh_n)^{-1}\sum_{i=1}^{N} K\left(\frac{y_i - x'_i\widehat{\beta}_\tau}{h_n}\right)x_i x'_i\end{aligned}$$

The choice of estimator for $\widehat{f}_{y|x,w\in W_j}(x'_{ij}\widehat{\beta}_\tau)$ as well as any nuisance parameters associated with its estimation is an important issue that remains open. It seems advisable for researchers to be cautious regarding the impacts of the choice of estimator and nuisance parameters on the standard errors used in QR. Pagan and Ullah (1999) present several possible estimators for such densities and discuss their advantages and drawbacks.

3.4 Conclusion

This chapter addressed the issue of inference on quantile regressions when the data is obtained through standard stratified sampling. Extending results from Wooldridge (2001) I derive the asymptotic distribution of the weighted quantile regression estimator for the case with general stratification and of the unweighted quantile regression estimator in the case that the stratification is a deterministic function of the conditioning variables. Valid estimators for the asymptotic variance matrix of those estimators are provided.

The results shown here provide confirmation to the intuition that the results for general M-estimators with smooth objective functions transfer neatly to the quantile regression case. This adds to literature a more careful treatment of the quantile regression case for stratified sampling that had not been available.[5]

[4] For the nonlinear case, one can use $\widehat{A}_{w_j} = \left(N_j h_{nj}\right)^{-1}\sum_{i=1}^{N_j} K\left(\frac{y_{ij} - g\left(x_{ij},\widehat{\beta}_\tau\right)}{h_{nj}}\right)\widehat{\bullet}_i\widehat{\bullet}'_i g_i g_i'$.

[5] And this fact has probably played a significant role in the absence of probability

Weighted quantile regression provides a simple and reliable way to deal with data obtained through stratified sampling, albeit requiring adjustments to the usual standard errors in the literature. Valid estimators for the standard errors are provided. Under exogenous stratification one could use the usual unweighted estimators, which retain its properties of consistency (Wooldridge, 2001) and asymptotic normality. Even more relevant, in that case, some usual standard error estimators in the literature (Koenker, 2005; Buchinsky, 1998, etc) are still valid.

weighted and "survey" methods for quantile regression in popular statistical packages like STATA.

APPENDICES

APPENDIX A

Proofs to "GMM Efficiency and IPW for Nonsmooth Functions"

Proof. [Proof of Theorem 1] For $V_{ONE-STEP}$, $V_{KNOW-\gamma}$ and $V_{KNOW-\gamma-JOINT}$ this result is a direct application of known results in the literature (see, e.g., p. 2186 in Newey and McFadden 1994 or more generally p. 1594 in Chen, Linton and Van Keilegom 2003) and the simplifications that take effect by the use of the appropriate weighting matrix. For $V_{TWO-STEP}$ I rely on the approximations used by Newey and McFadden (1994) in theorem 7.2 and Pakes and Pollard (1989) theorem 3.3 and lemma 3.5. Following Pakes and Pollard (1989), I claim that $g_n(\theta)$ is very well approximated by the linear function

$$\begin{aligned} L_n(\theta) &= \begin{bmatrix} L_{n1}(\theta) \\ L_{n2}(\theta) \end{bmatrix} = g_n(\theta_o) + G(\theta - \theta_o) \\ &= \begin{bmatrix} g_{n1}(\boldsymbol{\beta}_o, \boldsymbol{\gamma}_o) + G_{11}(\boldsymbol{\beta} - \boldsymbol{\beta}_o) + G_{12}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_o) \\ g_{n2}(\boldsymbol{\gamma}_o) + G_{22}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_o) \end{bmatrix} \end{aligned}$$

within a $O_p(n^{-\frac{1}{2}})$ neighborhood of θ_o. More precisely, I need the approximation error to be of order $o_p(n^{-\frac{1}{2}})$ at $\widehat{\theta}$ and at θ^* which minimizes $\|L_n(\theta)\|$ globally. In the case analyzed

here,

$$\begin{aligned}
\left\|g_n(\widehat{\theta}) - L_n(\widehat{\theta})\right\| &= \left\|g_n(\widehat{\theta}) - g_n(\theta_o) - G(\widehat{\theta} - \theta_o)\right\| \\
&= \left\|g_n(\widehat{\theta}) - g_n(\theta_o) - G(\widehat{\theta} - \theta_o) - g_o(\widehat{\theta}) + g_o(\widehat{\theta})\right\| \\
&\leq \left\|g_n(\widehat{\theta}) - g_o(\widehat{\theta}) - g_n(\theta_o)\right\| + \left\|g_o(\widehat{\theta}) - G(\widehat{\theta} - \theta_o)\right\| \\
&\leq o_p(1)n^{-\frac{1}{2}}\left[1 + \sqrt{n}\left\|(\widehat{\theta} - \boldsymbol{\theta}_o)\right\|\right] + o_p(\left\|(\widehat{\theta} - \boldsymbol{\theta}_o)\right\|) \\
&= o_p(n^{-\frac{1}{2}})
\end{aligned}$$

where in the last equality I used the fact that $\left\|(\widehat{\theta} - \boldsymbol{\theta}_o)\right\| \leq O_p(n^{-\frac{1}{2}})$ (see Newey and McFadden, 1994, p. 2191). To correspond to a minimum of $\|L_n(\theta)\|$, the vector $G(\theta^* - \theta_o)$ must be equal to the linear projection of $-g_n(\theta_o)$ onto the space G. Hence,

$$G(\theta^* - \theta_o) = -G(G'G)^{-1}G'g_n(\theta_o)$$

from this equation, we can obtain

$$\sqrt{n}(\theta^* - \theta_o) = -\sqrt{n}(G'G)^{-1}G'g_n(\theta_o)$$

from Pakes and Pollard (1989, lemma 3.5) the result above holds for the case in which we use the appropriate positive semidefinite weighting matrix $\widehat{W}$ that converges in probability to W , in which case

$$\sqrt{n}(\theta^* - \theta_o) = -\sqrt{n}(G'\widehat{W}G)^{-1}G'\widehat{W}g_n(\theta_o)$$

as shown by Pakes and Pollard (1989, p. 2042) under the conditions listed above θ^* and $\widehat{\theta}$ are close enough in this shrinking neighborhood around θ_o such that we can write

$$\sqrt{n}(\widehat{\theta} - \theta_o) = \sqrt{n}(\theta^* - \theta_o) + o_p(1)$$

Hence, for the first step estimator, the following approximation is valid

$$\sqrt{n}(\widehat{\gamma} - \gamma_o) = -\sqrt{n}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_{n2}(\boldsymbol{\gamma}_o) + o_p(1) \tag{A.1}$$

Then, for the second step, using the same results, we can approximate

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}-\beta_o) &= -\sqrt{n}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}g_{n1}(\boldsymbol{\beta}_o,\widehat{\boldsymbol{\gamma}})+o_p(1)\\
&= -\sqrt{n}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}\left[g_{n1}(\boldsymbol{\beta}_o,\boldsymbol{\gamma}_o)+G_{12}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_o)\right]+o_p(1)\\
&= -\sqrt{n}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}g_{n1}(\boldsymbol{\beta}_o,\boldsymbol{\gamma}_o)+\\
&+\sqrt{n}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}\times \qquad \text{(A.2)}\\
&\times G_{22}'C_{22}^{-1}g_{n2}(\boldsymbol{\gamma}_o)+o_p(1)
\end{aligned}$$

then, by combining A.1 and A.2 we can write

$$\sqrt{n}(\widehat{\theta}-\theta_o)=B\sqrt{n}g_n(\boldsymbol{\theta}_o)+o_p(1)$$

where,

$$B=\begin{bmatrix}B_{11} & B_{12}\\ 0 & B_{22}\end{bmatrix}$$

with

$$\begin{aligned}
B_{11} &= -\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}\\
B_{12} &= \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}\\
B_{22} &= -\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}
\end{aligned}$$

hence,

$$V_{TWO-STEP}=B\Sigma B'$$

■

Proof. [Proof of Corollary 1] The proof follows directly from Prokhorov and Schmidt (2009) since theorem 3 has shown that the variance structure of the four estimators considered is the same as in Prokhorov and Schmidt (2009). The proof that the result hold directly for the case in which the objective functions considered are nonsmooth is available under request.

■

Proof. [Proof of Corollary 2] Note that the asymptotic variance of $\sqrt{n}(\widehat{\beta}_{TWO-STEP}-\beta_o)$ can be rewritten as (note that $B_{12}=B_{11}G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}$)

$$\begin{aligned}
V(\widehat{\beta}_{TWO-STEP}) &= B_{11}C_{11}B_{11}' + B_{12}C_{21}B_{11}' + B_{11}C_{12}B_{12}' + B_{12}C_{22}B_{12}' \\
&= B_{11}E\left[g_1(\boldsymbol{\omega}_i^*,\theta)g_1(\boldsymbol{\omega}_i^*,\theta)'\right]B_{11}' + B_{12}E\left[g_2(\boldsymbol{\omega}_i^*,\gamma)g_1(\boldsymbol{\omega}_i^*,\theta)'\right]B_{11}' + \\
&\quad + B_{11}E\left[g_1(\boldsymbol{\omega}_i^*,\theta)g_2(\boldsymbol{\omega}_i^*,\gamma)'\right]B_{12}' + B_{12}E\left[g_2(\boldsymbol{\omega}_i^*,\gamma)g_2(\boldsymbol{\omega}_i^*,\gamma)'\right]B_{12}' \\
&= B_{11}E\left[\begin{array}{c}\left(g_1(\boldsymbol{\omega}_i^*,\theta) - G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)\right)\times \\ \times\left(g_1(\boldsymbol{\omega}_i^*,\theta) - G_{12}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)\right)'\end{array}\right]B_{11}' \\
&\qquad \text{if } G_{12}=C_{12}C_{22}^{-1}G_{22} \\
&= B_{11}E\left[\begin{array}{c}\left(g_1(\boldsymbol{\omega}_i^*,\theta) - C_{12}C_{22}^{-1}G_{22}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)\right)\times \\ \times\left(g_1(\boldsymbol{\omega}_i^*,\theta) - C_{12}C_{22}^{-1}G_{22}\left(G_{22}'C_{22}^{-1}G_{22}\right)^{-1}G_{22}'C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)\right)'\end{array}\right]B_{11}'
\end{aligned}$$

Since it is assumed that G_{22} is invertible,

$$= \; B_{11}E\left[\left(g_1(\boldsymbol{\omega}_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)\right)\left(g_1(\boldsymbol{\omega}_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)\right)'\right]B_{11}'$$

If we define $e_i = g_1(\boldsymbol{\omega}_i^*,\theta) - C_{12}C_{22}^{-1}g_2(\boldsymbol{\omega}_i^*,\gamma)$, and $D_o = E\left[e_i e_i'\right]$, we can write,

$$V(\widehat{\beta}_{TWO-STEP}) = \left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}G_{11}'C_{11}^{-1}D_oC_{11}^{-1\prime}G_{11}\left(G_{11}'C_{11}^{-1}G_{11}\right)^{-1}$$

In this case, we can write the variance of the two-step estimator for β_o in a quadratic form in which the term in the middle of the matrix is the residual of the linear projection of the first set of moment conditions on the second set of moment conditions.

If, in addition to the conditions above, we assume G_{11} is invertible, the result follows.

$$V(\widehat{\beta}_{TWO-STEP}) = G_{11}^{-1}D_oG_{11}^{-1\prime}$$

■

Proof. [Proof of Lemma 2] First, note that,

$$\begin{aligned}
E\left[sg_2(z,\gamma_o,s)' \mid z\right] &= E\left[s\frac{\nabla_\gamma f(v_i \mid z_i,\gamma)'}{f(v_i \mid z_i,\gamma)} \mid z\right] \\
&= \int_{-\infty}^{\infty} s\frac{\nabla_\gamma f(v \mid z,\gamma)'}{f(v \mid z,\gamma)} f(v \mid z,\gamma)dv \\
&= \int_{-\infty}^{\infty} h(v,z)\nabla_\gamma f(v \mid z,\gamma)'dv \\
&= \nabla_\gamma \left[\int_{-\infty}^{\infty} h(v,z)f(v \mid z,\gamma)'dv\right] \\
&= \nabla_\gamma E\left[s \mid z\right] \\
&= \nabla_\gamma p(z,\gamma_o)
\end{aligned}$$

this is nonzero in general. Hence,

$$\begin{aligned}
C_{12} &= E\left[g_1(\boldsymbol{\omega}^*,\beta_o,\gamma_o,s)g_2(z,\boldsymbol{\gamma}_o,s)'\right] \\
&= E\left[\frac{s}{p(z,\boldsymbol{\gamma}_o)}g(\boldsymbol{\omega},\boldsymbol{\beta}_o)g_2(z,\gamma_o,s)'\right] \\
&= E\left[E\left[\frac{s}{p(z,\boldsymbol{\gamma}_o)}g(\boldsymbol{\omega},\boldsymbol{\beta}_o)g_2(z,\gamma_o,s)' \mid z\right]\right] \\
&= E\left[E\left[\frac{1}{p(z,\boldsymbol{\gamma}_o)}g(\boldsymbol{\omega},\boldsymbol{\beta}_o)sg_2(z,\gamma_o,s)' \mid z\right]\right] \\
&= E\left[\frac{1}{p(z,\boldsymbol{\gamma}_o)}E\left[g(\boldsymbol{\omega},\boldsymbol{\beta}_o) \mid z\right]E\left[sg_2(z,\gamma_o,s)' \mid z\right]\right], \text{ by ignorability} \\
&= E\left[\frac{g(\boldsymbol{\omega},\boldsymbol{\beta}_o)}{p(z,\boldsymbol{\gamma}_o)}E\left[sg_2(z,\gamma_o,s) \mid z\right]'\right] \\
&= E\left[\frac{g(\boldsymbol{\omega},\boldsymbol{\beta}_o)}{p(z,\boldsymbol{\gamma}_o)}\nabla_\gamma p(z,\gamma_o)\right]
\end{aligned}$$

which is generally nonzero.

Analyzing G_{12},

$$\begin{aligned}
G_{12} &= \nabla_\gamma E[g_1(\boldsymbol{\omega}^*,\beta_o,\gamma_o,s)] \\
&= \nabla_\gamma E\left[\frac{s}{p(z,\boldsymbol{\gamma}_o)}g(\boldsymbol{\omega},\boldsymbol{\beta}_o)\right]
\end{aligned}$$

since, $g_1(\boldsymbol{\omega}^*, \beta_o, \gamma_o, s) = \frac{s}{p(\mathbf{z},\boldsymbol{\gamma})} g(\boldsymbol{\omega}, \boldsymbol{\beta})$, is smooth in γ,

$$\begin{aligned}
G_{12} &= E\left[\nabla_\gamma \left(\frac{s}{p(\mathbf{z}, \boldsymbol{\gamma}_o)}\right) g(\boldsymbol{\omega}, \boldsymbol{\beta}_o)\right] \\
&= E\left[-\frac{s}{(p(\mathbf{z}, \boldsymbol{\gamma}_o))^2} \nabla_\gamma p(\mathbf{z}, \boldsymbol{\gamma}_o) g(\boldsymbol{\omega}, \boldsymbol{\beta}_o)\right] \\
&= E\left[-\frac{s}{p(\mathbf{z}, \boldsymbol{\gamma}_o)} g(\boldsymbol{\omega}, \boldsymbol{\beta}_o) \frac{\nabla_\gamma p(\mathbf{z}, \boldsymbol{\gamma}_o)}{p(\mathbf{z}, \boldsymbol{\gamma}_o)}\right] \\
&= -E\left[E\left[\frac{s}{p(\mathbf{z}, \boldsymbol{\gamma}_o)} g(\boldsymbol{\omega}, \boldsymbol{\beta}_o) \frac{\nabla_\gamma p(\mathbf{z}, \boldsymbol{\gamma}_o)}{p(\mathbf{z}, \boldsymbol{\gamma}_o)}\right] \shortmid z\right], \text{ by LIE} \\
&= -E\left[\frac{E(s \shortmid z)}{p(\mathbf{z}, \boldsymbol{\gamma}_o)} E\left[g(\boldsymbol{\omega}, \boldsymbol{\beta}_o) \shortmid z\right] \frac{\nabla_\gamma p(\mathbf{z}, \boldsymbol{\gamma}_o)}{p(\mathbf{z}, \boldsymbol{\gamma}_o)}\right] \\
&= -E\left[g(\boldsymbol{\omega}, \boldsymbol{\beta}_o) \frac{\nabla_\gamma p(\mathbf{z}, \boldsymbol{\gamma}_o)}{p(\mathbf{z}, \boldsymbol{\gamma}_o)}\right], \text{ since } E\left[s \shortmid z\right] = p(\mathbf{z}, \boldsymbol{\gamma}_o) \\
&= -C_{12}
\end{aligned}$$

Then, to prove the lemma 2 I need that $G_{22} = -C_{22}$, which follows from the Generalized Information Equality (remembering $g_2(\mathbf{z},\gamma_o,s)$ is a smooth function).

$$\begin{aligned}
G_{22} &= \nabla_\gamma E[g_2(\mathbf{z}, \boldsymbol{\gamma}_o, s)] \\
&= E[\nabla_\gamma g_2(\mathbf{z}, \boldsymbol{\gamma}_o, s)] \\
&= -E\left[g_2(\mathbf{z},\gamma_o,s) g_2(\mathbf{z},\gamma_o,s)'\right] = -C_{22}
\end{aligned}$$

hence, $G_{12} = -C_{12} = -C_{12}(-C_{22}^{-1} G_{22}) = C_{12} C_{22}^{-1} G_{22}$. ■

Proof. [Proof of Theorem 2] This follows directly from Lemma 2 and statements 9 and 10 in Corollary 1. ■

Claim 1 *Consider the conditional quantile function*

$$Q_\tau(Y \mid X) = X' \beta_{\tau_o}$$

and the weighted linear quantile estimator obtained as

$$\widehat{\beta}_\tau = \arg \min_{b \in \mathbb{R}^p} \sum w_i \rho_\tau (y_i - x_i' b)$$

for some known weight w_i that could be a function of exogenous variables. Under conditions 7 and 8, we have

$$\sqrt{n}\left(\widehat{\beta_\tau}-\beta_\tau\right)\sim N\left(0,\tau(1-\tau)D_1^{-1}D_oD_1^{-1}\right)$$

with, $D_2=\lim_{n\to\infty}\sum_{i=1}^n w_i f_i(x_i'\beta_{\tau_o})x_ix_i'$ *and* $D_o=\lim_{n\to\infty}\sum_{i=1}^n w_i^2x_i'x_i$

Assumption 7 *For $Y_1,Y_2,\ldots,Y_n$ independent random variables with distribution functions $F_1,F_2,\ldots,F_n$, $\{F_i\}$ are absolutely continuous with continuous densities $f_i(\cdot)$ and weights, w_i, uniformly bounded away from 0 and ∞ at the points $f_i\left(x_i'\beta_{\tau_o}\right)$ for every i.*

Assumption 8 *There exist positive definite matrices D_o and D_1 such that*

i) $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n w_i^2x_ix_i'=D_o$

ii) $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n w_if_i(x_i'\beta_{\tau_o})x_ix_i'=D_1$

iii) $\max\frac{\|x_i\|}{\sqrt{n}}\to 0$

Proof. [Proof of Claim 1] This proof follows the steps presented on Koenker (2005, p. 120).

Consider $u_i=y_i-x_i'\beta_{\tau_o}$, then

$$\begin{aligned}\widehat{\beta_\tau} &= \arg\min_{b\in\mathbb{R}p}\sum_{i=1}^n w_i\left[\left(y_i-x_i'b\right)\left(\tau-1\left[y_i-x_i'b\leq 0\right]\right)\right]\\ &= \arg\min_{b\in\mathbb{R}p}\sum_{i=1}^n w_i\rho_\tau(u_i)\end{aligned}$$

Consider the following convex objective function, with unique minimizer at $\sqrt{n}\left(\widehat{\beta_\tau}-\beta_{\tau_o}\right)$,

$$Z_n(\delta)=\sum_{i=1}^n w_i\left[\rho_\tau\left(u_i-x_i'\frac{\delta}{\sqrt{n}}\right)-\rho_\tau(u_i)\right]$$

using Knight's identity $\rho_\tau(u-v)-\rho_\tau(u)=-v\Psi_\tau(u)+\int_0^v\left(1\left[u\leq S\right]-1\left[u\leq 0\right]\right)dS$, with $\Psi_\tau(u)=\tau-1\left[u\leq 0\right]$

$$\begin{aligned}Z_n(\delta) &= \sum_{i=1}^n w_i\left[-x_i'\frac{\delta}{\sqrt{n}}\Psi_\tau(u_i)+\int_0^{x_i'\frac{\delta}{\sqrt{n}}}\left(1\left[u_i\leq S\right]-1\left[u_i\leq 0\right]\right)dS\right]\\ &= Z_{1n}(\delta)+\sum_{i=1}^n Z_{2ni}(\delta)=Z_{1n}(\delta)+Z_{2n}(\delta)\end{aligned}$$

Note that, by the Lindeberg-Feller central limit theorem,

$$\begin{aligned} Z_{1n}(\delta) &= -\delta' \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i x_i' \Psi_\tau(u_i) \\ &= -\delta' \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i x_i' \left(\tau - 1\left[u_i \leq 0\right]\right) \\ &\sim -\delta' W \\ W &\sim N\left(0, \tau(1-\tau) \lim_{n\to\infty} \sum_{i=1}^{n} w_i^2 x_i x_i'\right) \end{aligned}$$

Also,

$$\begin{aligned} Z_{2n}(\delta) &= \sum_{i=1}^{n} Z_{2ni}(\delta) \\ &= \sum_{i=1}^{n} E\left[Z_{2ni}(\delta)\right] + \sum_{i=1}^{n} Z_{2ni}(\delta) - E\left[Z_{2ni}(\delta)\right] \end{aligned}$$

but,

$$\begin{aligned} \sum_{i=1}^{n} E\left[Z_{2ni}(\delta)\right] &= \sum_{i=1}^{n} w_i \int_0^{x_i'\frac{\delta}{\sqrt{n}}} E\left[1\left[u_i \leq S\right] - 1\left[u_i \leq 0\right]\right] dS \\ &= \sum_{i=1}^{n} w_i \int_0^{x_i'\frac{\delta}{\sqrt{n}}} F_i(x_i'\beta_{\tau o} + S) - F_i(x_i'\beta_{\tau o}) dS \end{aligned}$$

let $S = \frac{t}{\sqrt{n}}$, then

$$\begin{aligned} \sum_{i=1}^{n} E\left[Z_{2ni}(\delta)\right] &= \frac{1}{n} \sum_{i=1}^{n} w_i \int_0^{x_i'\delta} \sqrt{n}\left[F_i\left(x_i'\beta_{\tau o} + \frac{t}{\sqrt{n}}\right) - F_i(x_i'\beta_{\tau o})\right] dt \\ &= \frac{1}{n} \sum_{i=1}^{n} w_i \int_0^{x_i'\delta} f_i(x_i'\beta_{\tau o}) t dt + o(1) \\ &= \frac{1}{2n} \sum_{i=1}^{n} w_i f_i(x_i'\beta_{\tau o}) \delta' x_i x_i' \delta + o(1) \\ &\to \frac{1}{2}\delta' \left[\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_i f_i(x_i'\beta_{\tau o}) x_i x_i'\right] \delta = \frac{1}{2}\delta' D_1 \delta \end{aligned}$$

Under A2(iii):

$$Z_n(\delta) \sim Z_o(\delta) = -\delta' W + \frac{1}{2}\delta' D_1 \delta$$

then

$$\begin{aligned} \sqrt{n}\left(\widehat{\beta_\tau} - \beta_\tau\right) &= \widehat{\delta_n} = \arg\min Z_n(\delta) \sim \widehat{\delta_o} = \arg\min Z_o(\delta) \\ \widehat{\delta_o} &= D_1^{-1} W \end{aligned}$$

hence,

$$\sqrt{n}\left(\widehat{\beta_\tau} - \beta_\tau\right) \sim N\left(0, \tau(1-\tau) D_1^{-1} D_o D_1^{-1}\right)$$

■

APPENDIX B

Figures to "Fixed Bandwidth Asymptotics for Regression Discontinuity Designs"

B.1 Simulations for Infeasible Inference

B.1.1 Nadaraya-Watson Estimator

Figure B.1. Nadaraya-Watson Estimator - DGP: No X - Homosk. Errors

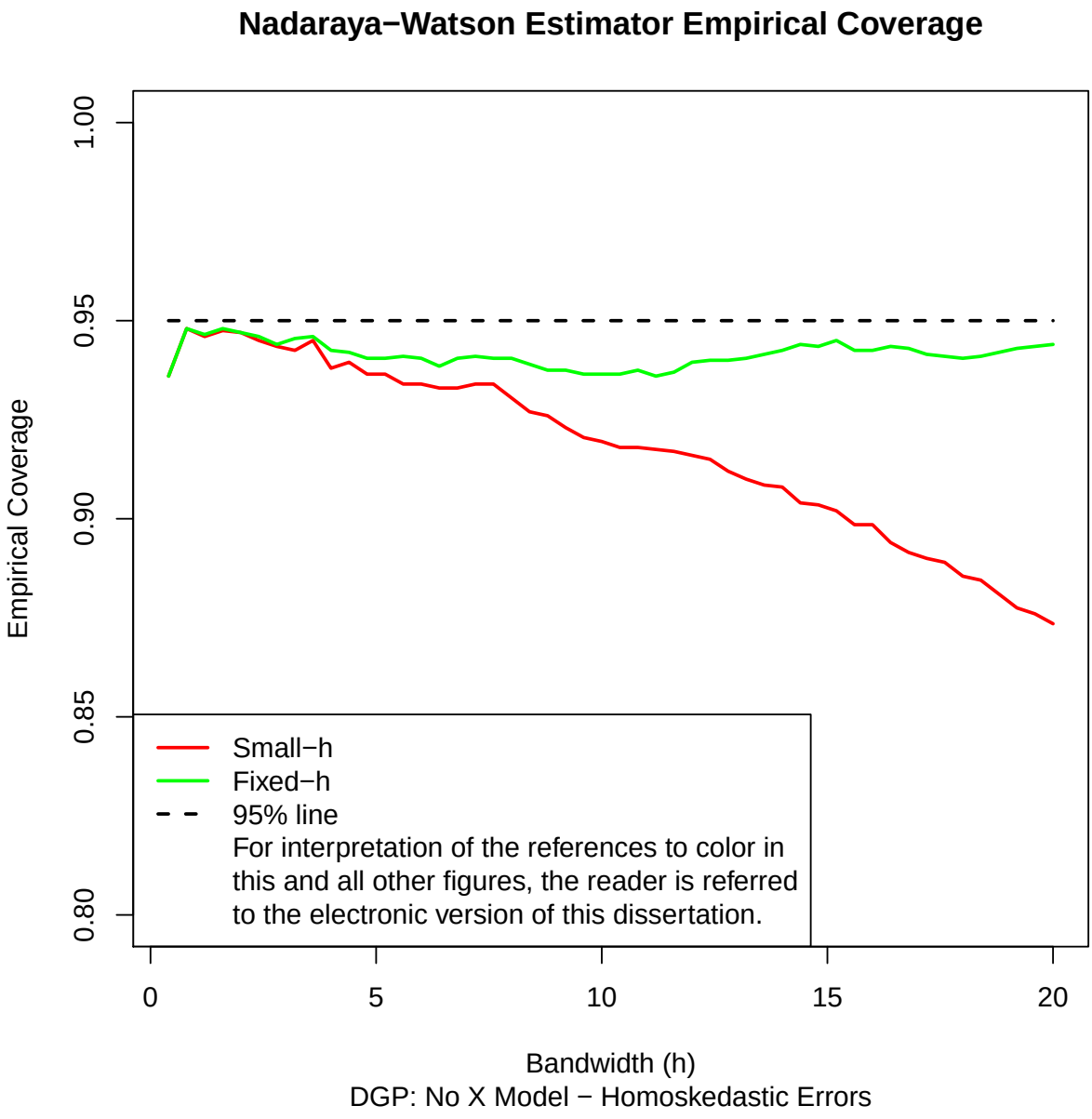


Figure B.2. Nadaraya-Watson Estimator - DGP: Linear - Homosk. Errors

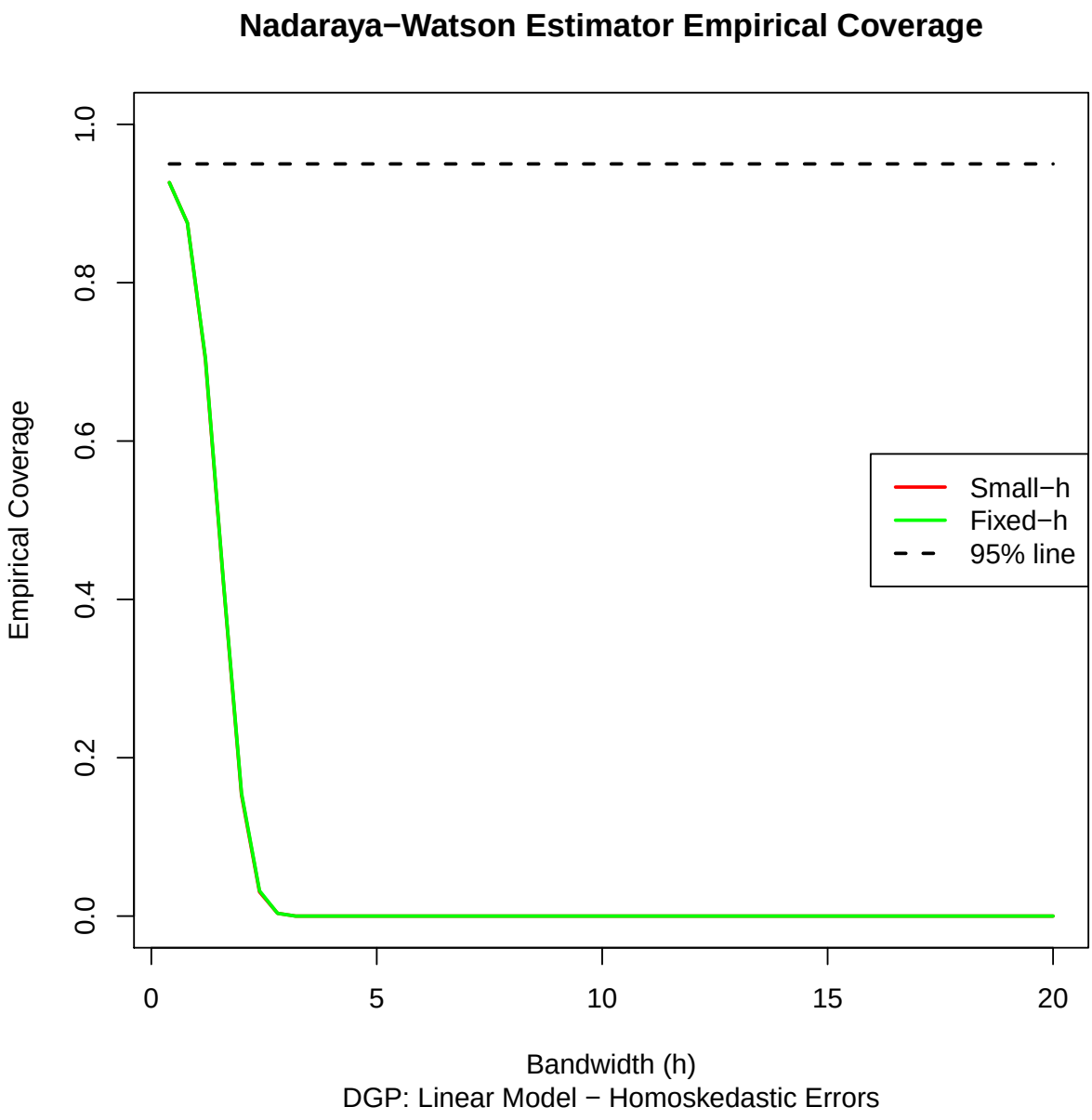


Figure B.3. Nadaraya-Watson Estimator - DGP: Linear - Bias Corrected - Homosk. Errors

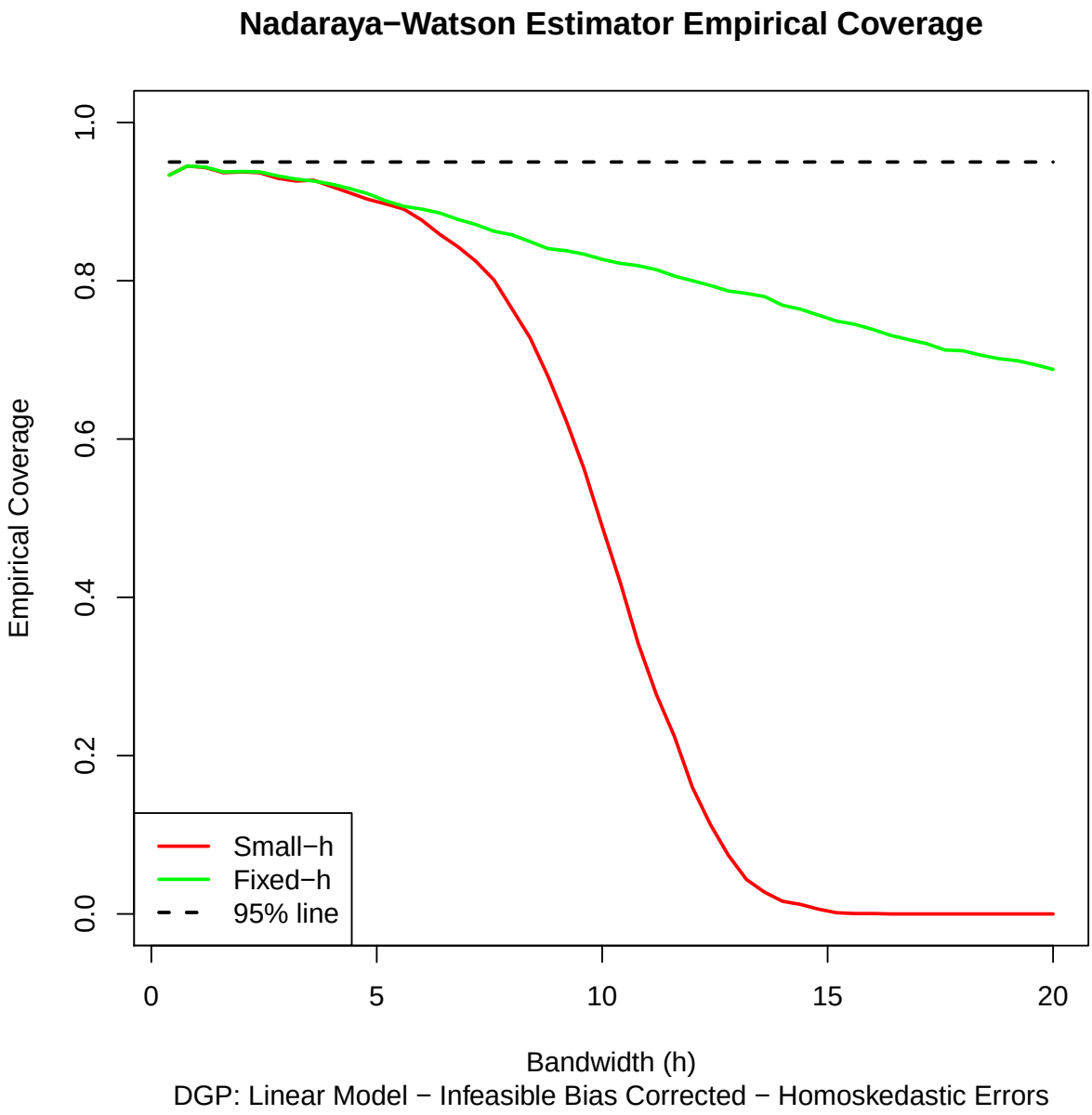
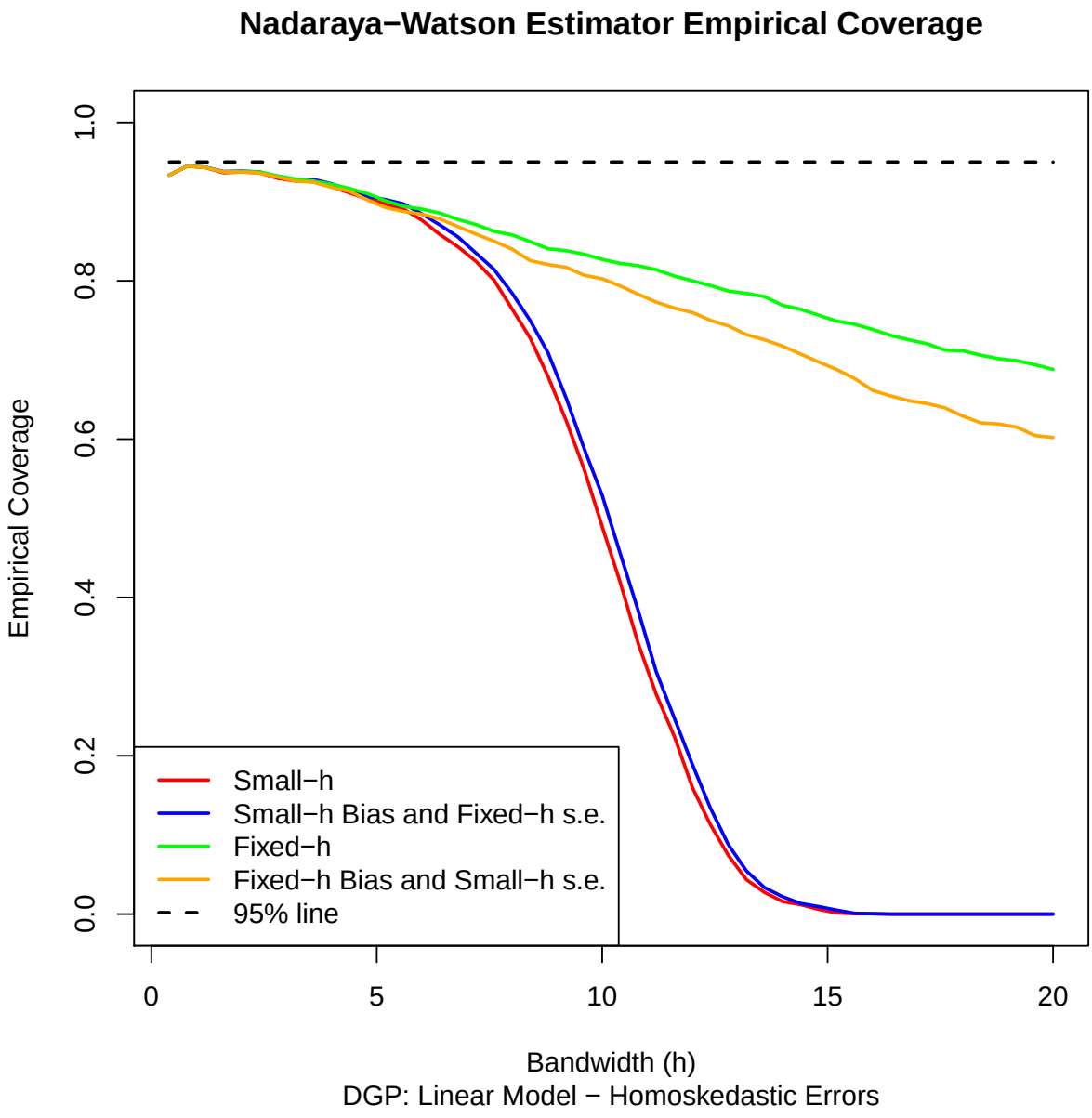


Figure B.4. Nadaraya-Watson Estimator - DGP: Linear - Comparison - Homosk. Errors



B.1.2 Local Linear Estimator

Figure B.5. Local Linear Estimator - DGP: Linear - Homosk. Errors

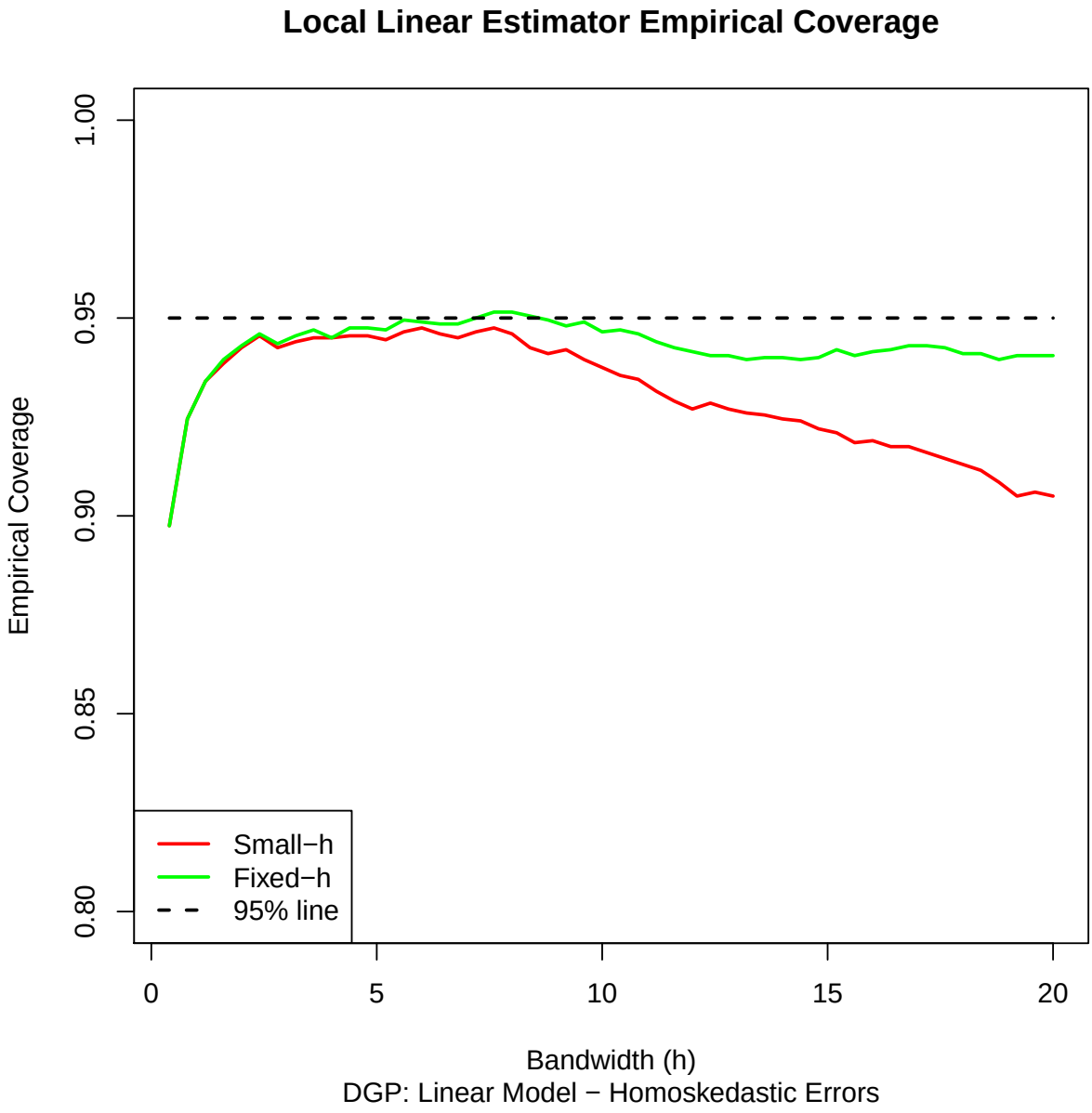


Figure B.6. Local Linear Estimator - DGP: Quadratic - Homosk. Errors

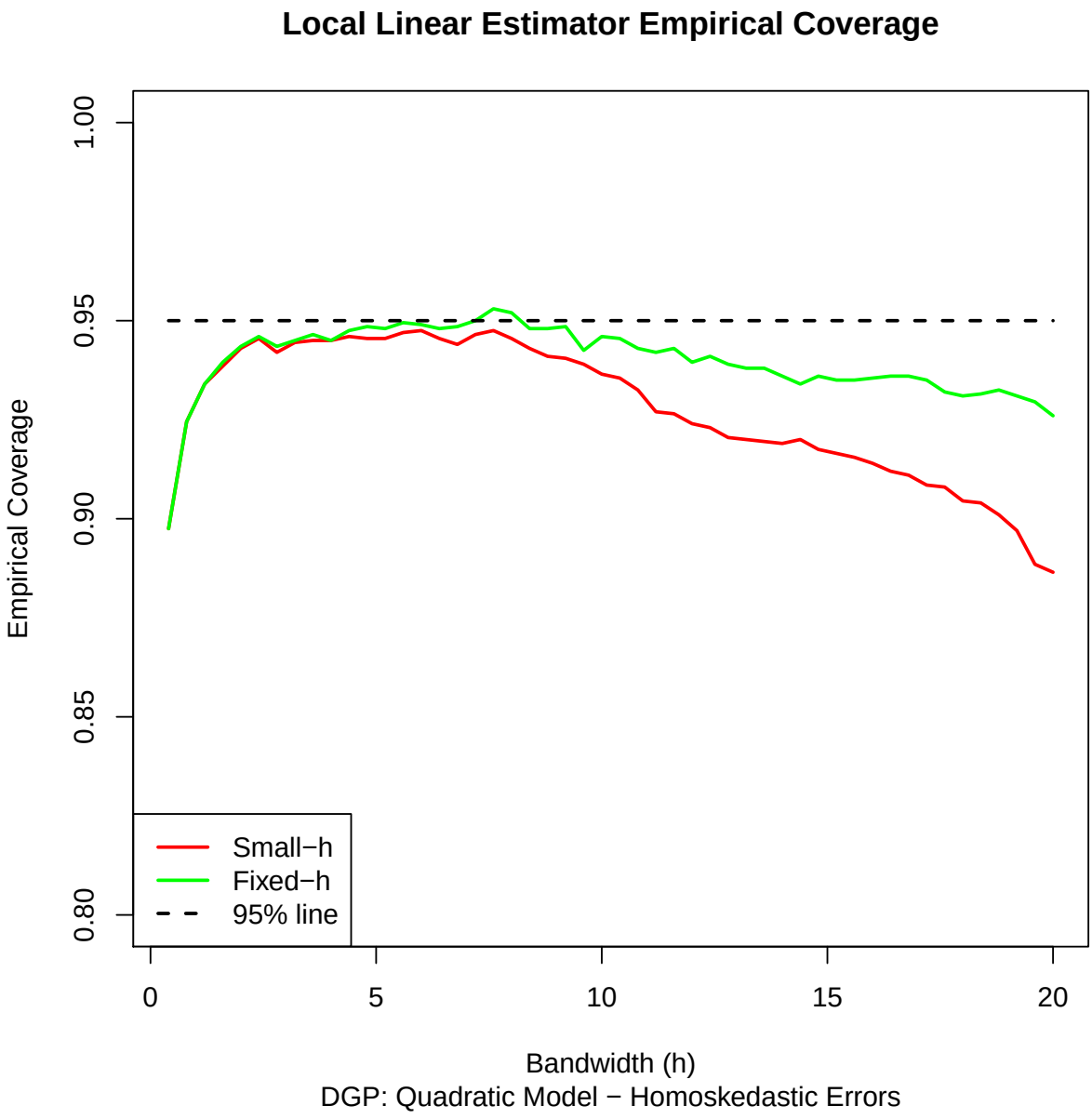


Figure B.7. Local Linear Estimator - DGP: Quadratic - Bias Corrected - Homosk. Errors

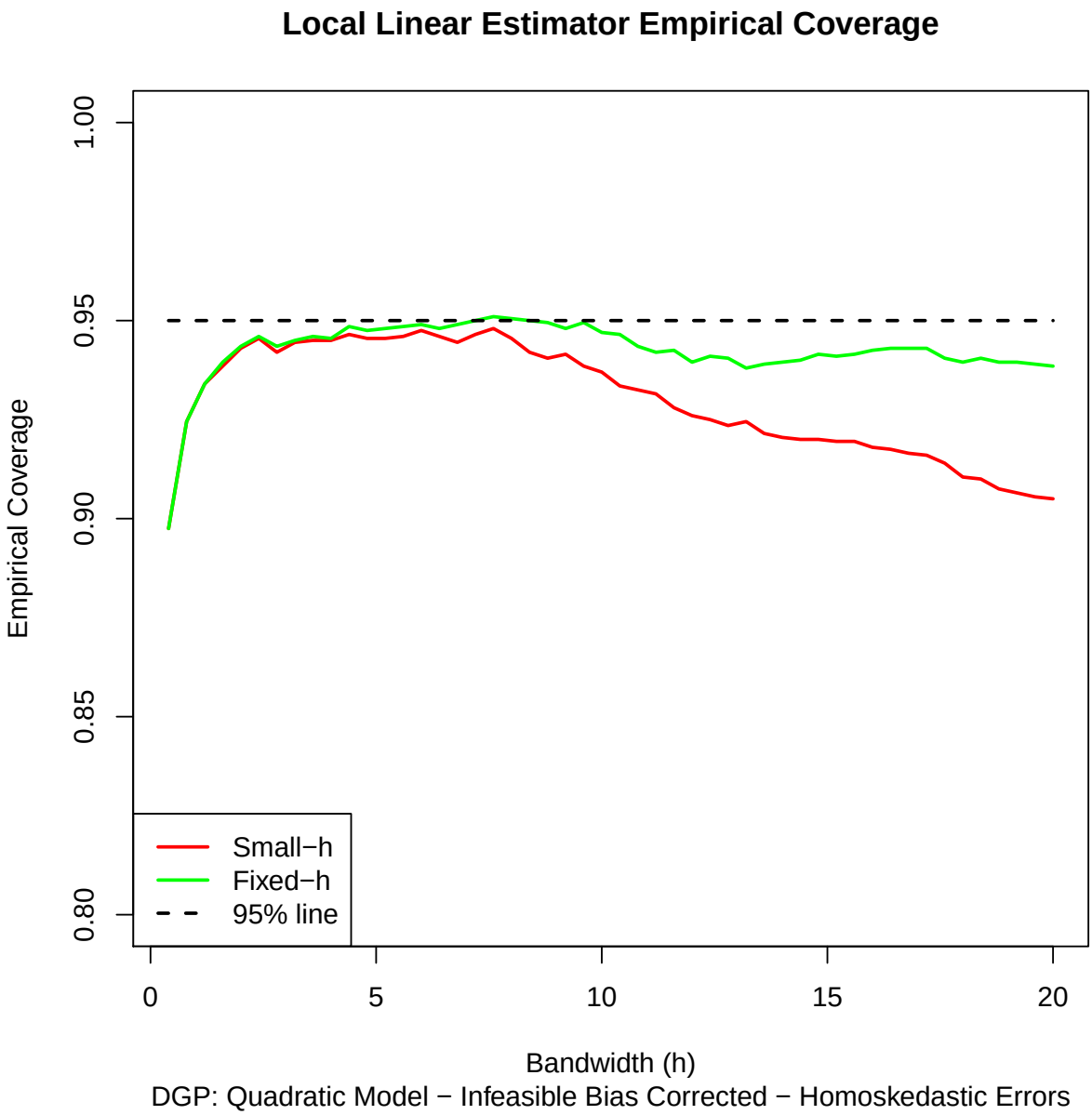
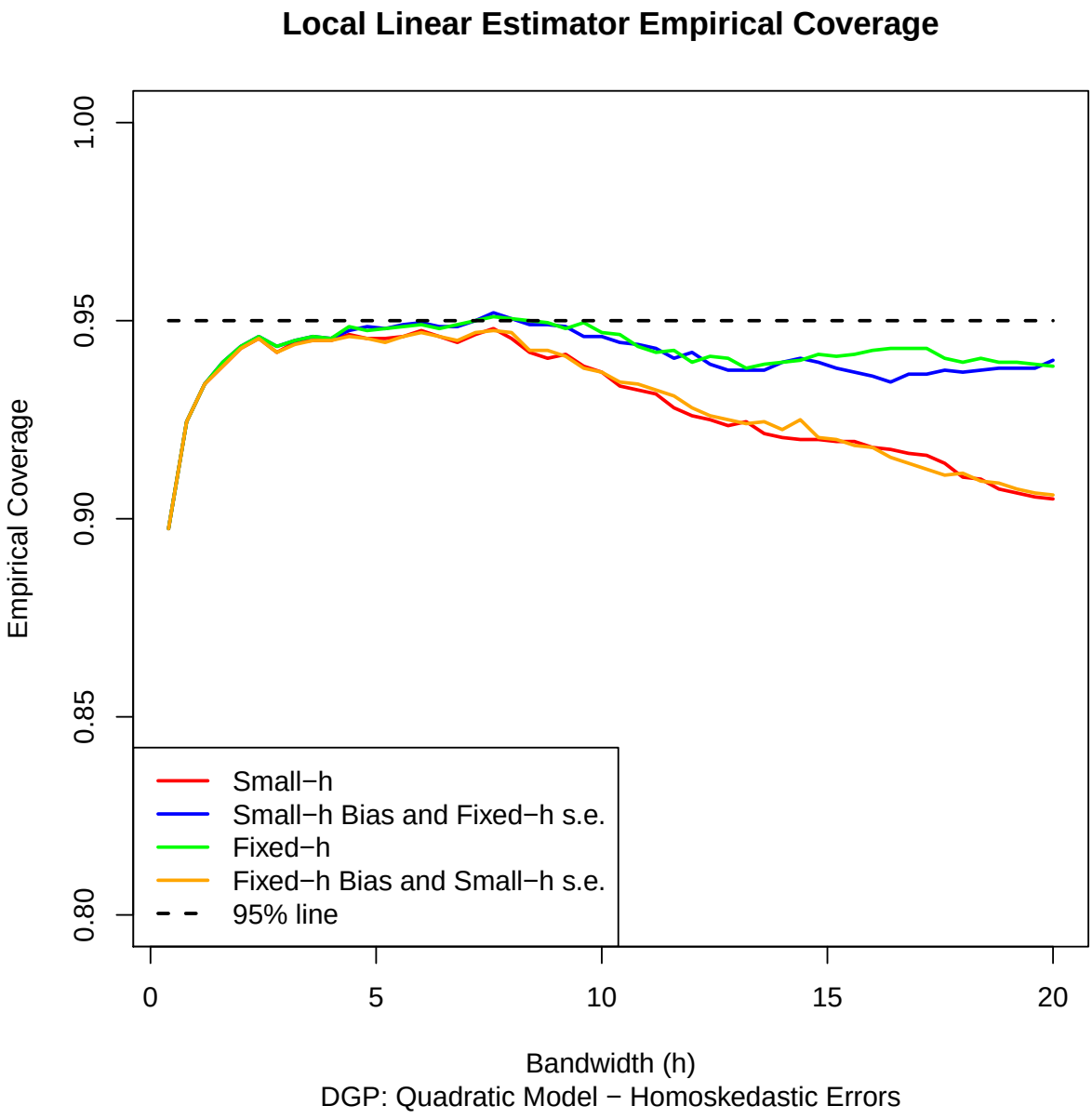


Figure B.8. Local Linear Estimator - DGP: Quadratic - Comparison - Homosk. Errors



B.1.3 Heteroskedastic Errors

Figure B.9. Local Linear Estimator - DGP: Linear - Heterosk. Errors Case 1

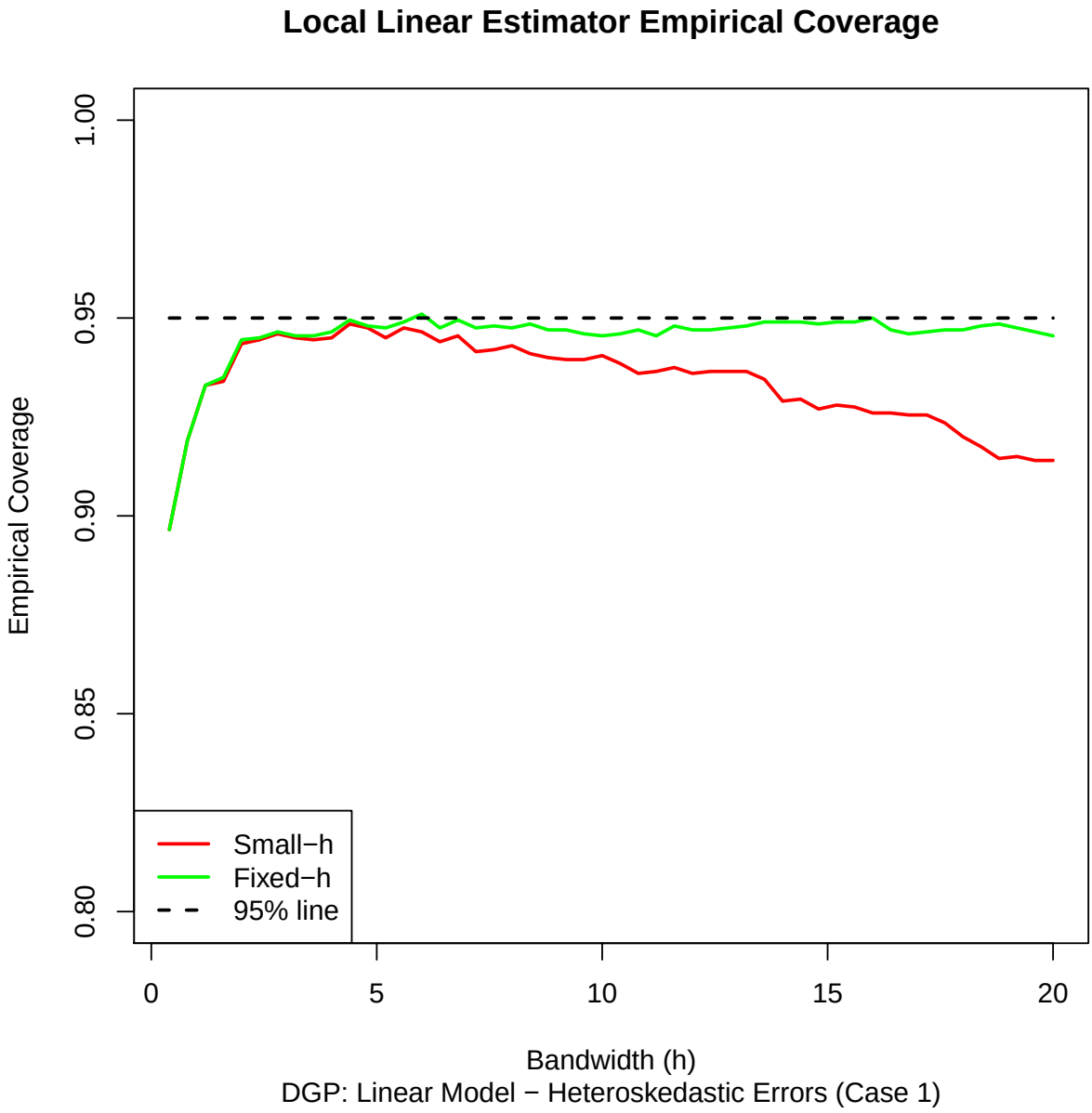


Figure B.10. Local Linear Estimator - DGP: Quadratic - Heterosk. Errors Case 1

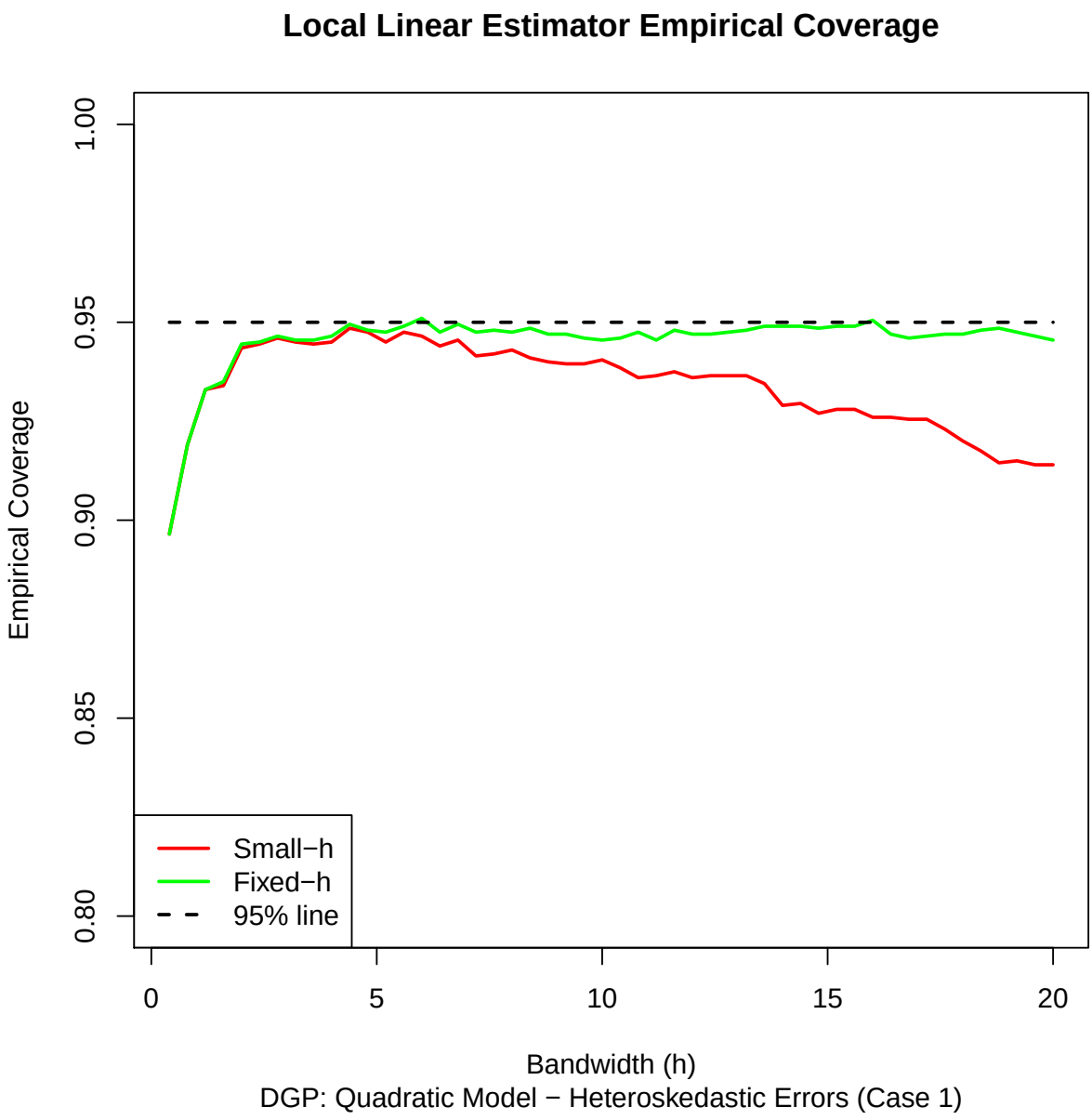


Figure B.11. Local Linear Estimator - DGP: Linear - Heterosk. Errors Case 2

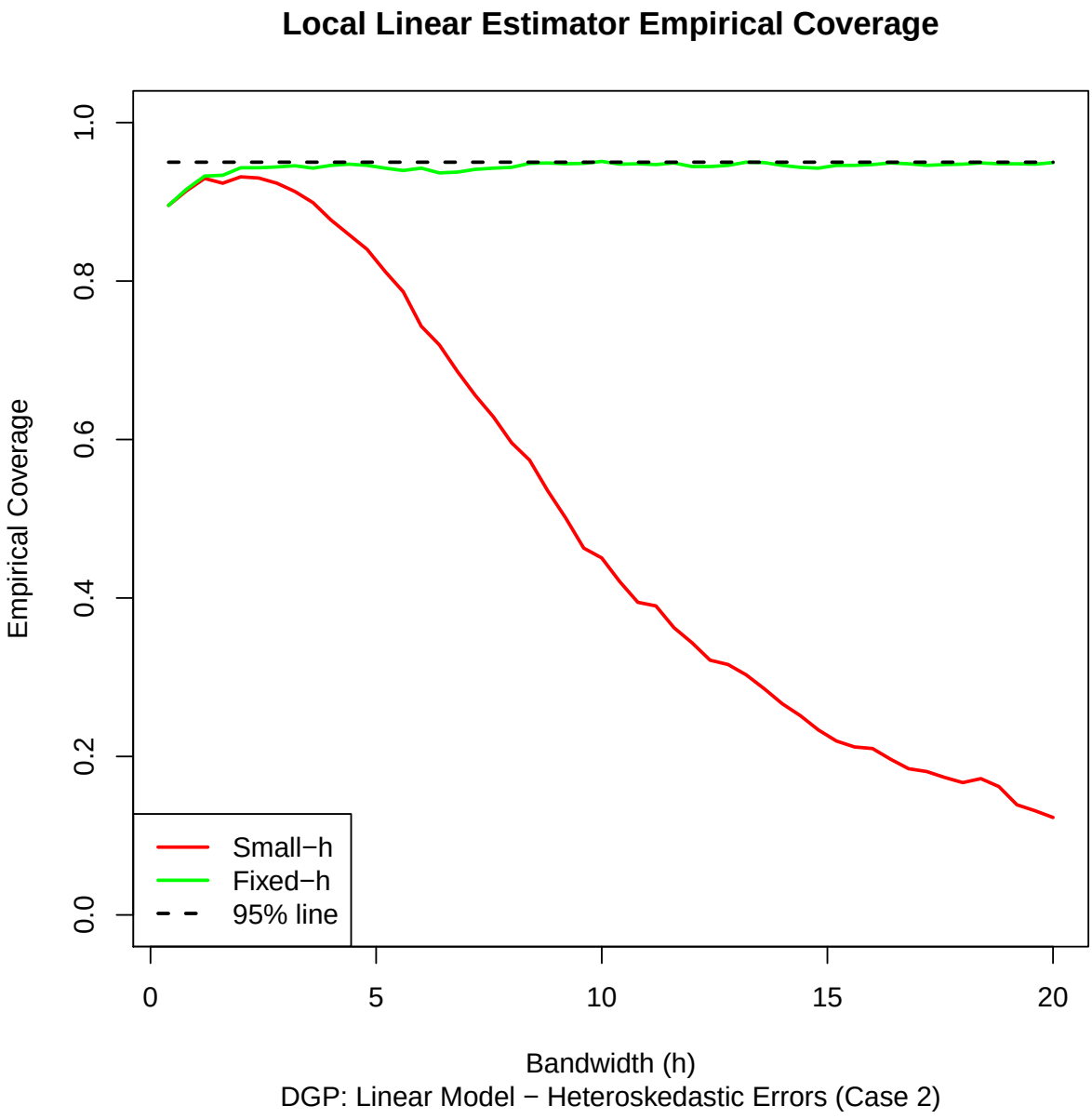
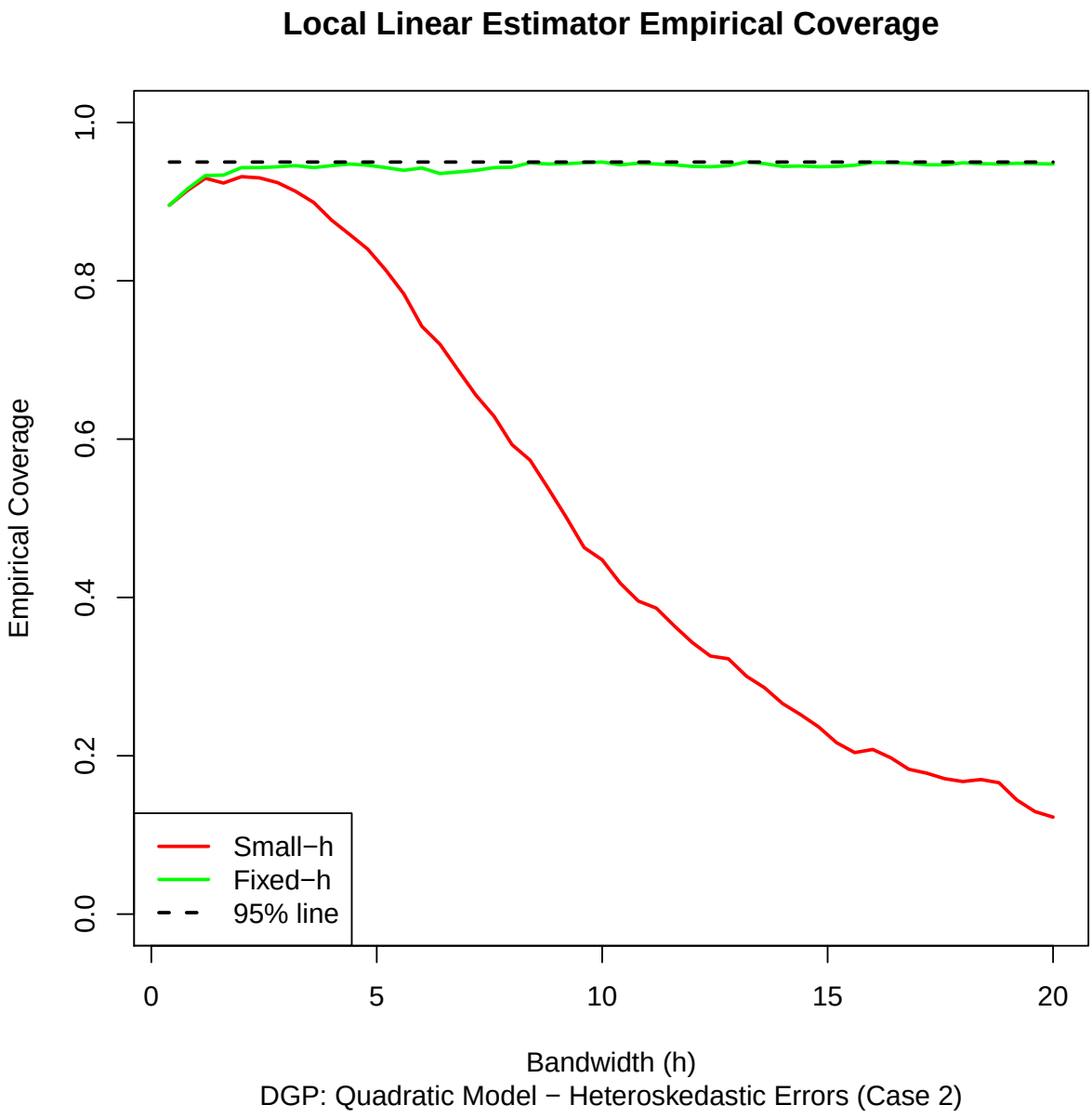


Figure B.12. Local Linear Estimator - DGP: Quadratic - Heterosk. Errors Case 2



B.2 Simulations for Feasible Inference

Figure B.13. Nadaraya-Watson Estimator - DGP: No X - Feasible - Homosk. Errors

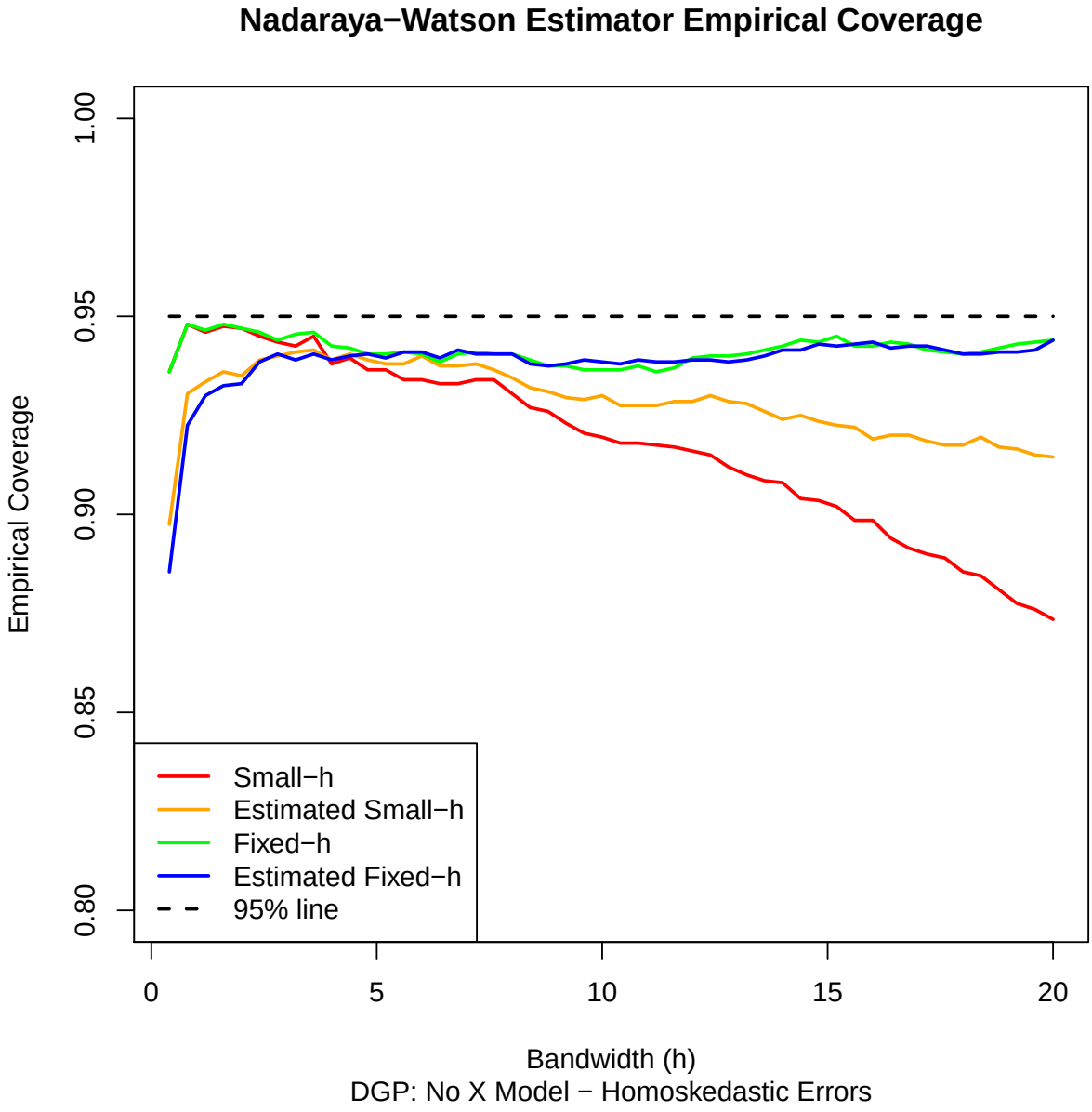
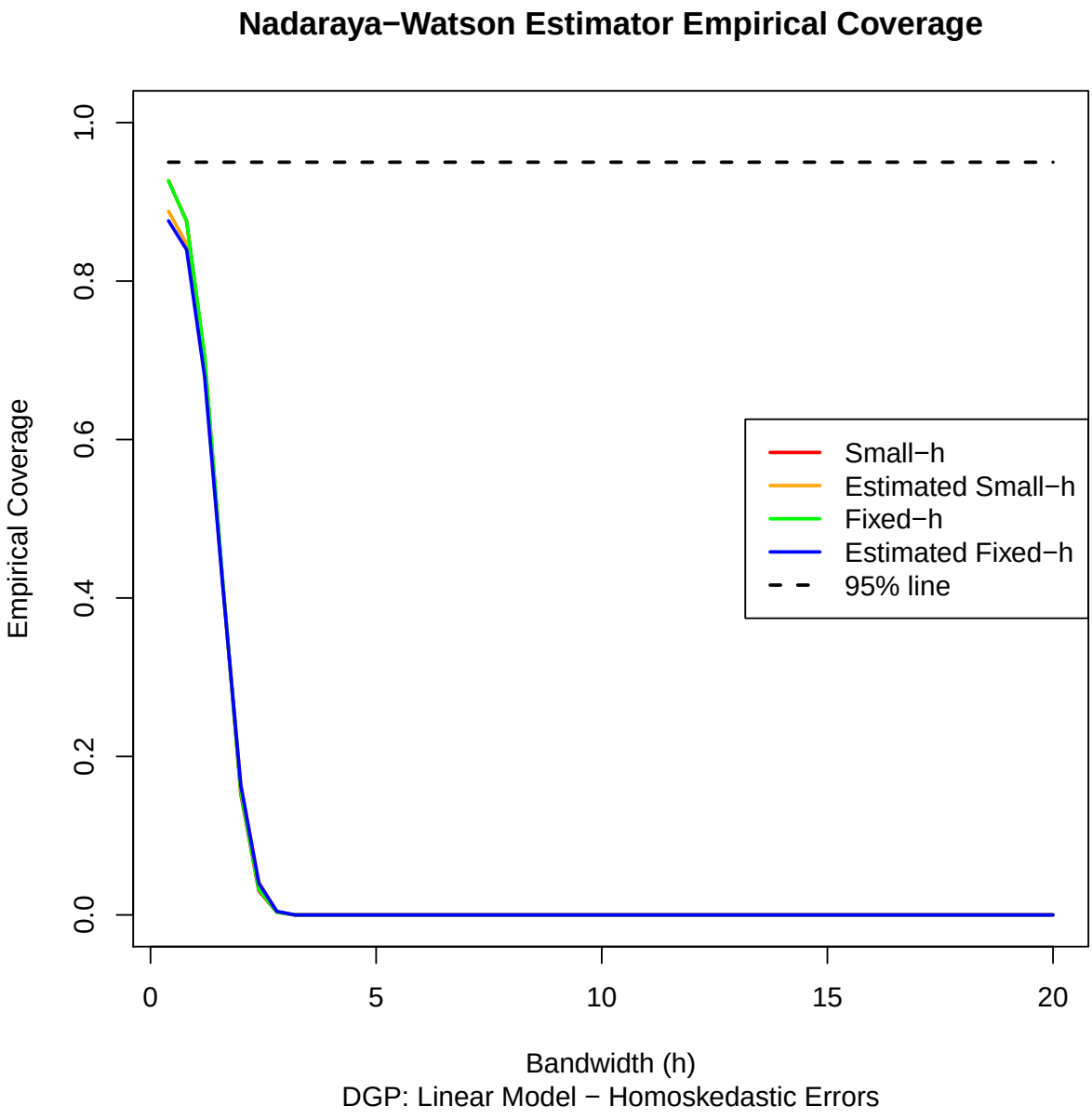


Figure B.14. Nadaraya-Watson Estimator - DGP: Linear - Feasible - Homosk. Errors



B.2.1 Local Linear Estimator

Figure B.15. Local Linear Estimator - DGP: Linear - Feasible - Homosk. Errors

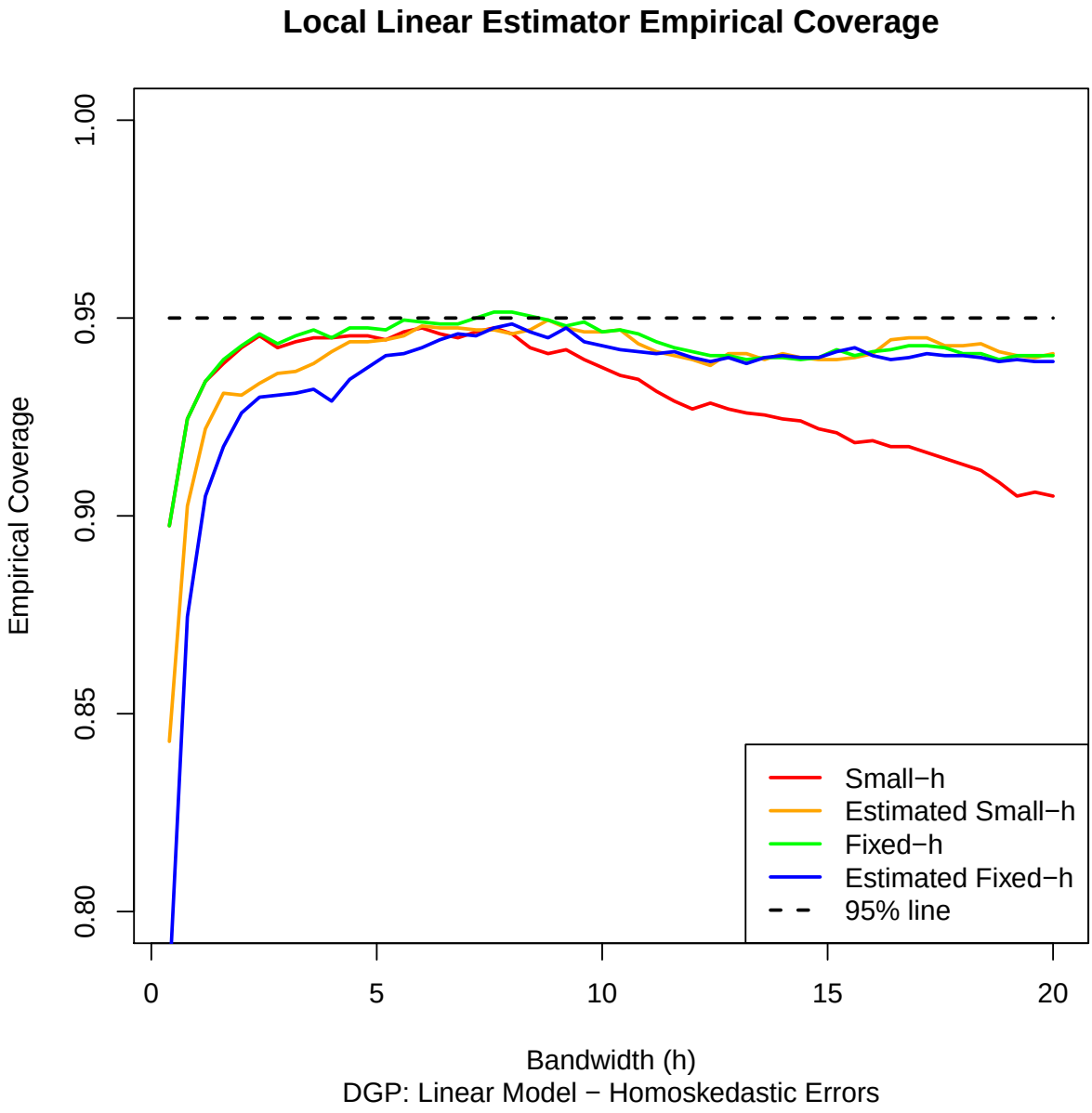
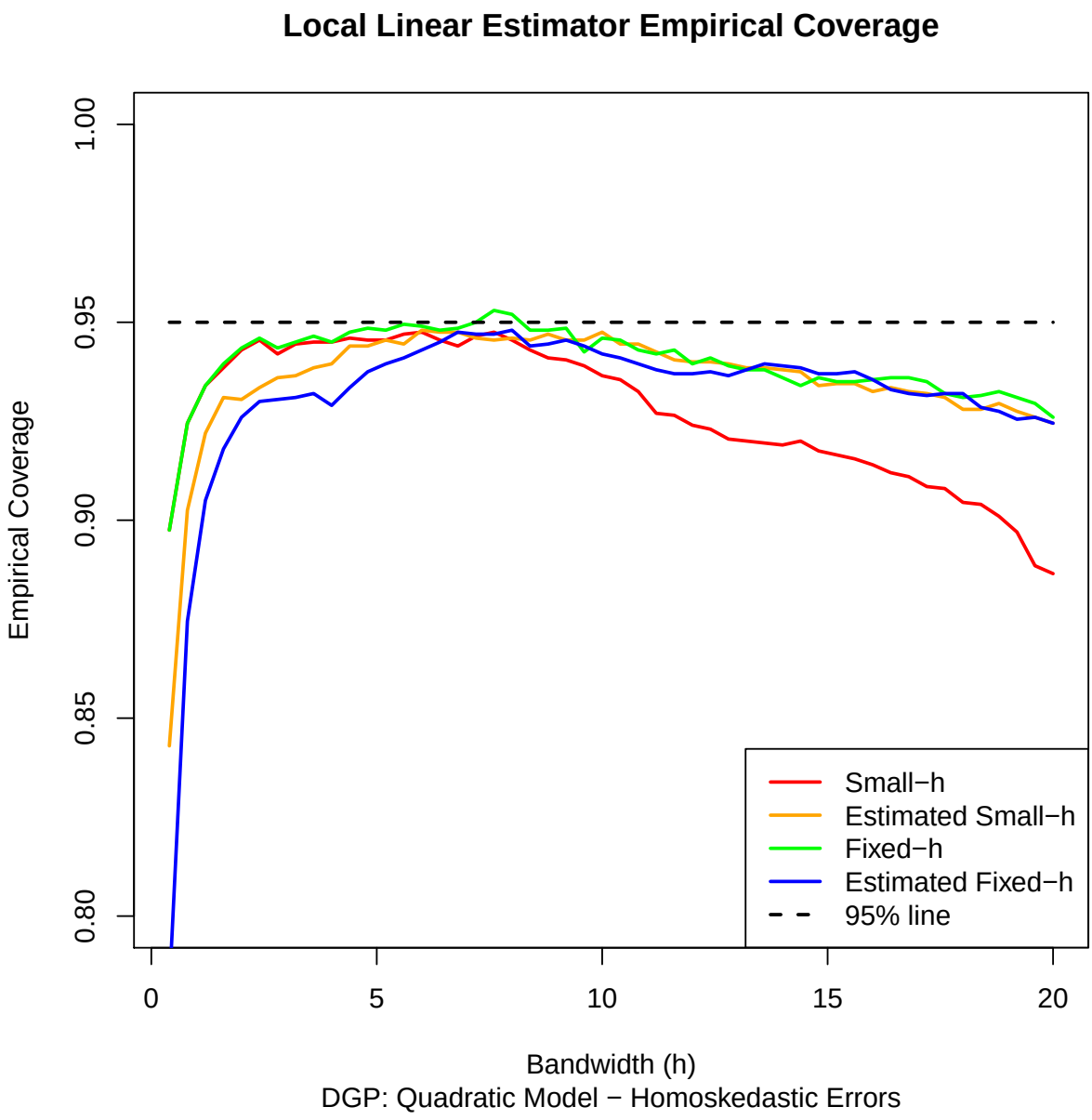


Figure B.16. Local Linear Estimator - DGP: Quadratic - Feasible - Homosk. Errors



B.2.2 Heteroskedastic Errors

Figure B.17. Local Linear Estimator - DGP: Linear - Feasible - Heterosk. Errors Case 1

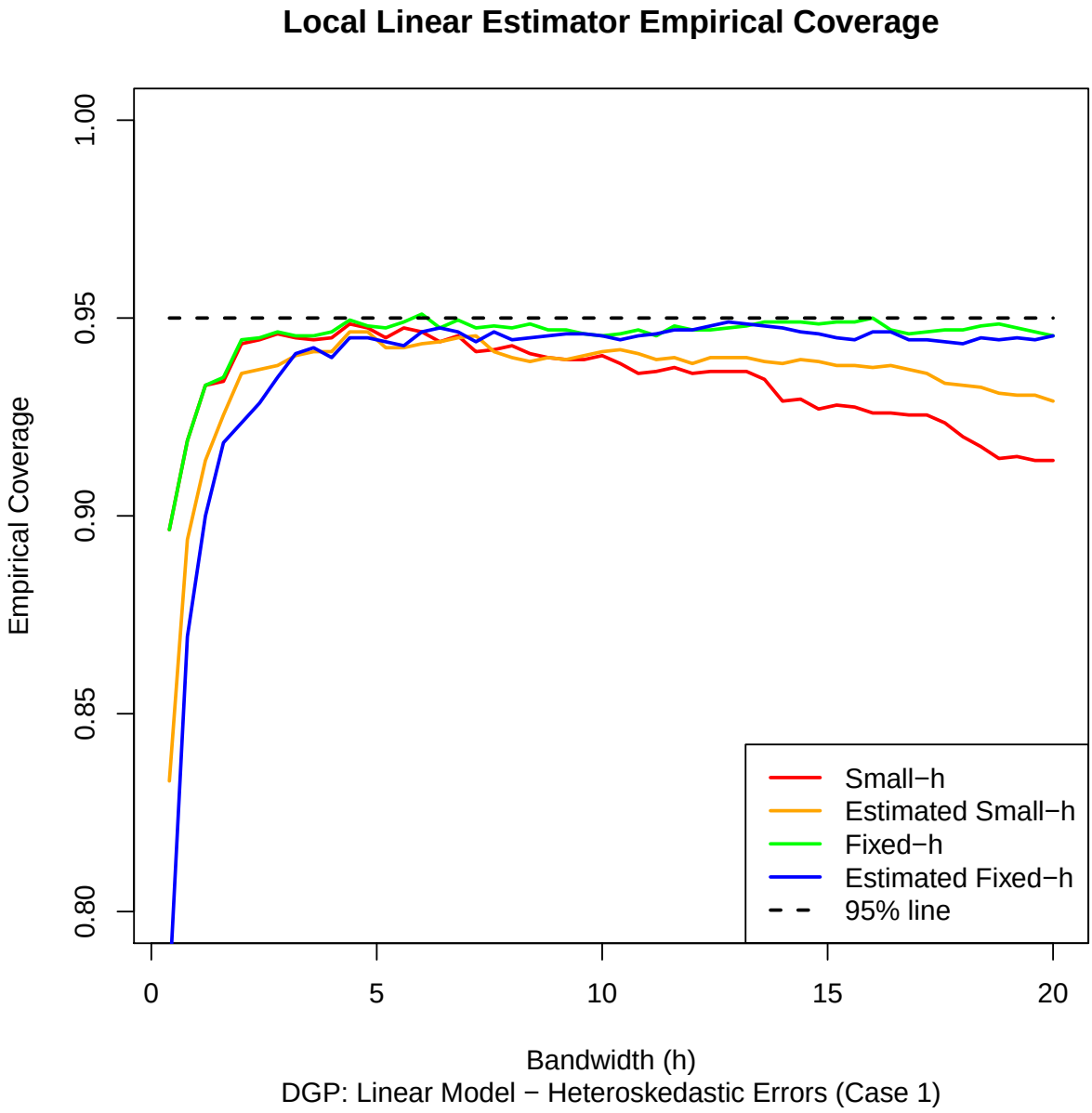


Figure B.18. Local Linear Estimator - DGP: Quadratic - Feasible - Heterosk. Errors Case 1

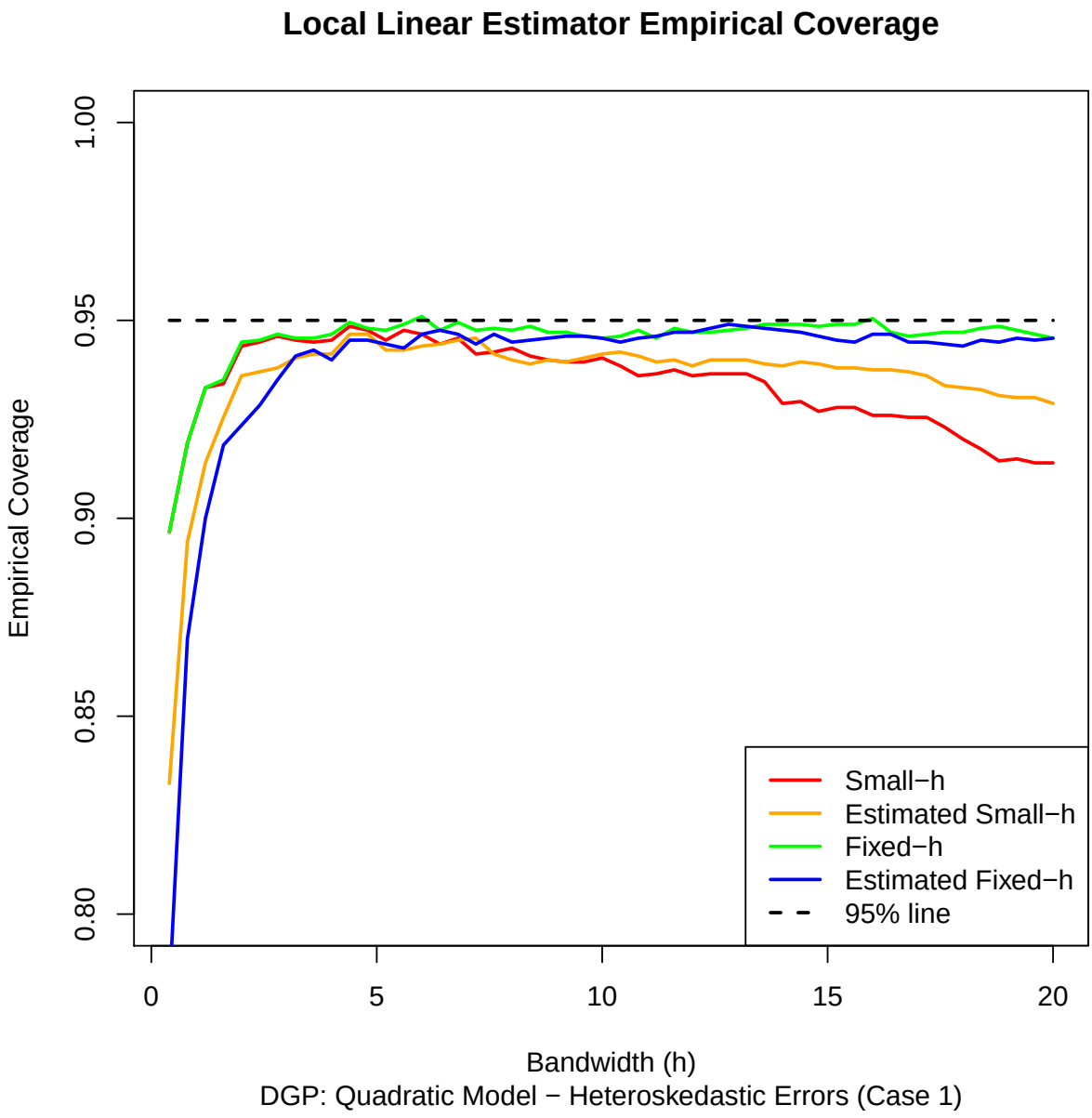


Figure B.19. Local Linear Estimator - DGP: Linear - Feasible - Heterosk. Errors Case 2

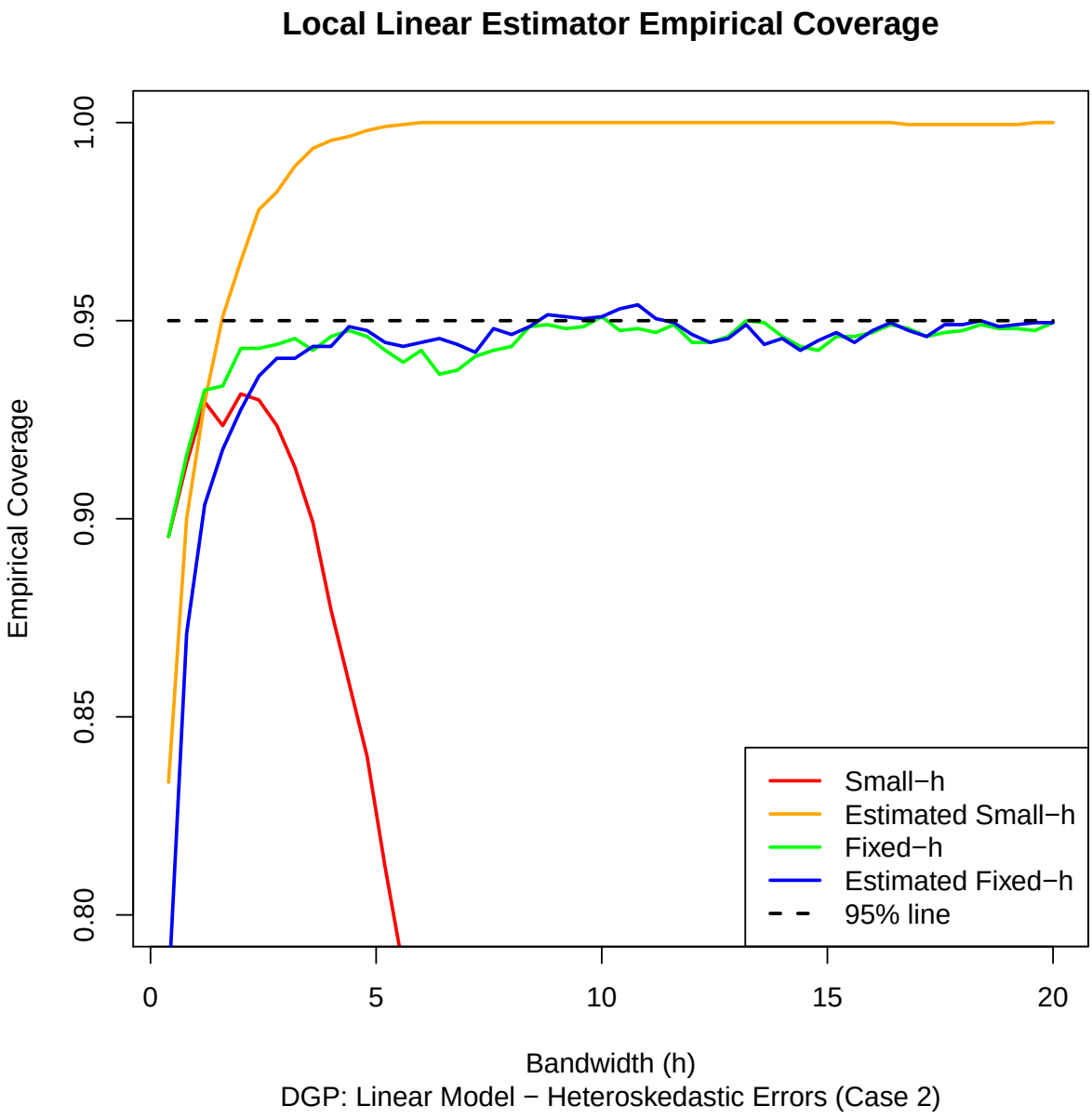
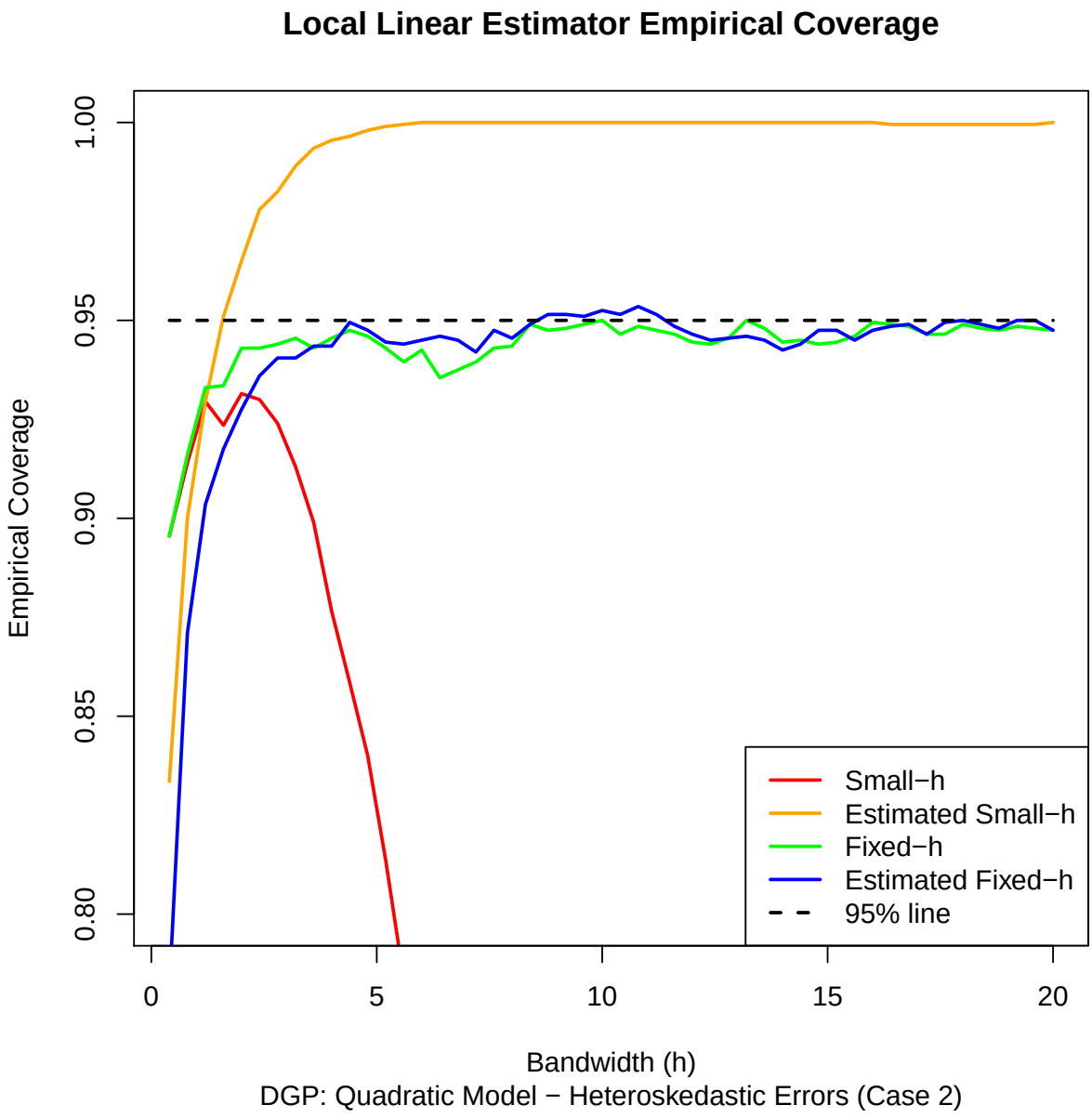


Figure B.20. Local Linear Estimator - DGP: Quadratic - Feasible - Heterosk. Errors Case 2



B.2.3 Bandwidth Choice for $\widehat{f}_o(\overline{x})$

Figure B.21. Small-h Sensitivity to Density Bandwidth - DGP: Linear

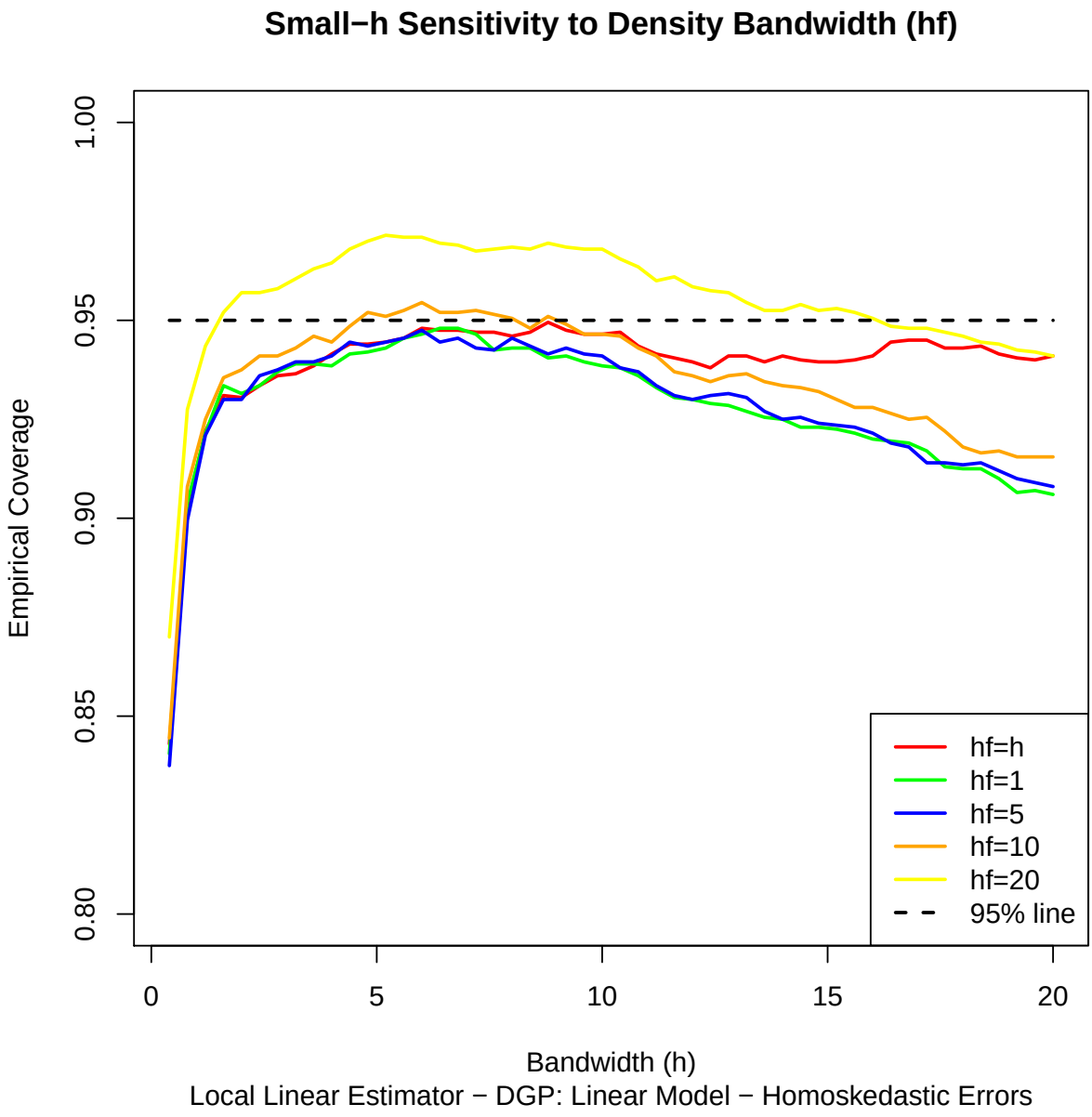
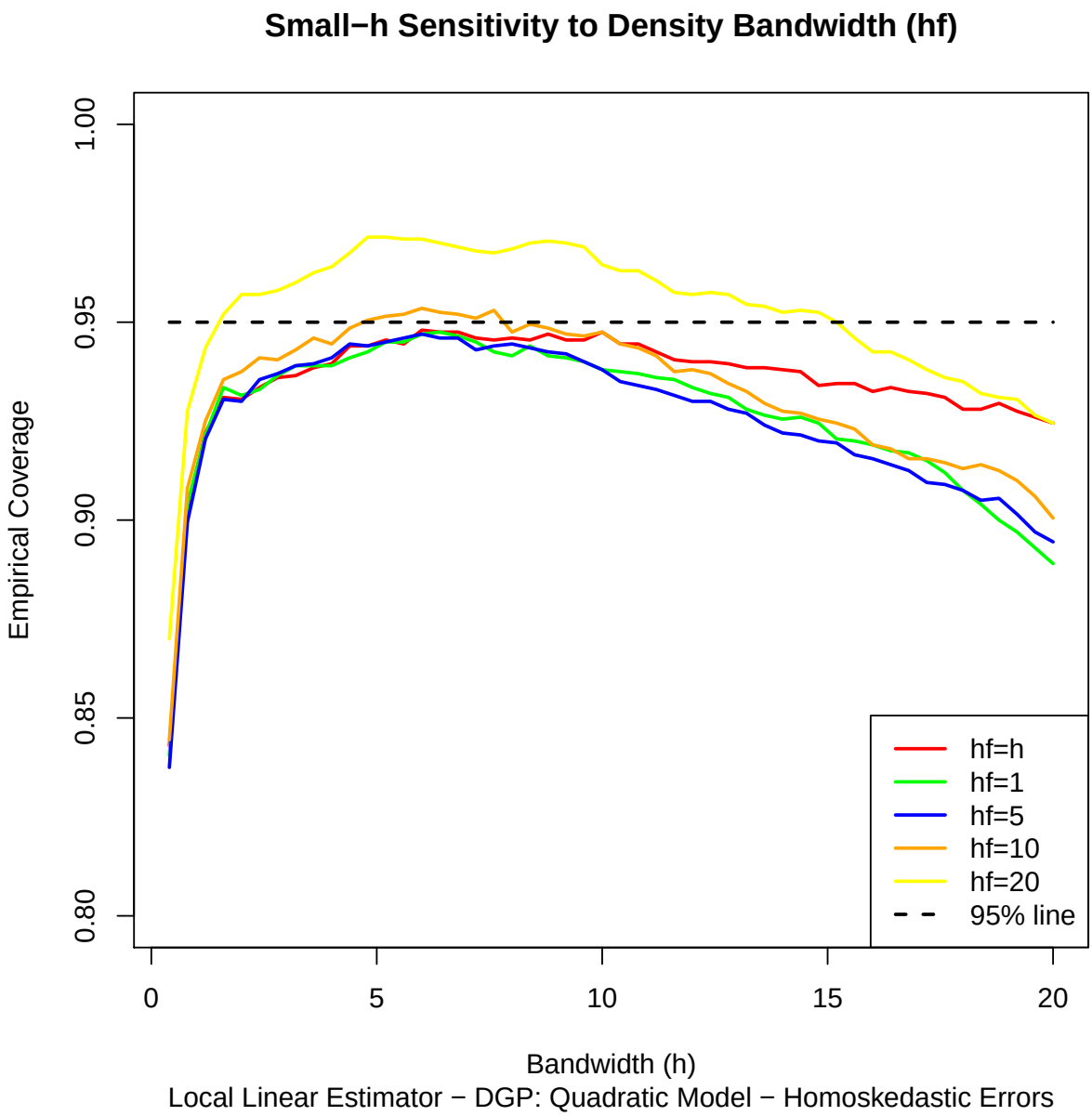


Figure B.22. Small-h Sensitivity to Density Bandwidth - DGP: Quadratic



APPENDIX C

Proofs to "Fixed Bandwidth Asymptotics for Regression Discontinuity Designs"

Proof. [Proof of Theorem 6] The local polynomial estimator is given by

$$\widehat{\alpha}_p = \widehat{\alpha}_{p+} - \widehat{\alpha}_{p-}$$

note that,

$$\begin{aligned}
\widehat{\alpha}_{p+} &= e_1' \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i Z_i' \right]^{-1} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i y_i \right] \\
&= e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i y_i \right]
\end{aligned}$$

with $D_{n+} = \left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i Z_i'\right]^{-1}$ and,

$$\widehat{\alpha}_{p-} = e_1' D_{n-} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) (1 - d_i) Z_i y_i \right]$$

with $D_{n-} = \left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_iZ_i'\right]^{-1}$. Then,

$$\begin{aligned}
\widehat{\alpha}_{p+} &= e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_i\left[m(x_i)+\alpha d_i+\varepsilon_i\right]\right] \\
&= e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_im(x_i)\right] + \\
&\quad +\alpha e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_i\right] + \\
&\quad +e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_i\varepsilon_i\right]
\end{aligned}$$

note that $Z_i = Z_iZ_i'e_1$, $e_1'e_1 = 1$ then

$$e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_i\right] = e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_iZ_i'\right]e_1 = 1$$

and

$$\widehat{\alpha}_{p+}-\alpha = e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_im(x_i)\right] + e_1'D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_i\varepsilon_i\right]$$

similarly

$$\widehat{\alpha}_{p-} = e_1'D_{n-}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_im(x_i)\right] + e_1'D_{n-}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_i\varepsilon_i\right]$$

Then

$$\begin{aligned}
\sqrt{nh}(\widehat{\alpha}_p-\alpha) &= \sqrt{nh}(\widehat{\alpha}_{p+}-\alpha-\widehat{\alpha}_{p-}) \\
&= e_1'D_{n+}\sqrt{nh}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_im(x_i)\right] + \\
&\quad +e_1'D_{n+}\left[\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_iZ_i\varepsilon_i\right] - \\
&\quad -e_1'D_{n-}\left\{\begin{array}{c} \sqrt{nh}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_im(x_i)\right] + \\ +\left[\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_i\varepsilon_i\right] \end{array}\right\}
\end{aligned}$$

For the denominator terms D_{n+} and D_{n-},

$$D_{n+}^{-1} = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i Z_i'$$

and each element of this matrix is given by

$$\left[D_{n+}^{-1}\right]_{l,j} = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}$$

which has asymptotic variance converging to zero since

$$\begin{aligned}
Var\left(\left[D_{n+}^{-1}\right]_{j,l}\right) &= \frac{1}{(nh)^2} Var\left(\sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}\right) \\
&= \frac{1}{nh^2} Var\left(k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}\right) \\
&\leq \frac{1}{nh} E\left[\frac{1}{h} k^2\left(\frac{x - \overline{x}}{h}\right) d \left(\frac{x - \overline{x}}{h}\right)^{2(j+l-2)}\right] \\
&= \frac{1}{nh}\int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k^2\left(\frac{x - \overline{x}}{h}\right)\left(\frac{x - \overline{x}}{h}\right)^{2(j+l-2)} f_o(x)dx
\end{aligned}$$

Note that the terms in the integral and the integral itself are $O(1)$ and $\frac{1}{nh} = o(1)$. Hence, $Var\left(\left[D_{n+}^{-1}\right]_{l,j}\right) \to 0$. Now,

$$\begin{aligned}
\left[D_{n+}^{-1}\right]_{l,j} &= E\left\{\left[D_{n+}^{-1}\right]_{l,j}\right\} + o_p(1) \\
&= \frac{1}{nh} E\left[\sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}\right] + o_p(1) \\
&= E\left[\frac{1}{h} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}\right] + o_p(1) \\
&= \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k\left(\frac{x - \overline{x}}{h}\right)\left(\frac{x - \overline{x}}{h}\right)^{j+l-2} f_o(x)dx + o_p(1) \\
&= \int_{0}^{\infty} k(u)\, u^{j+l-2} f_o(\overline{x} + uh)du + o_p(1)
\end{aligned}$$

Let, $\gamma_j^+ = \int_0^\infty k(u)\, u^j f_o(\overline{x}+uh)du$ and Γ_+^* is the $(p+1)\times(p+1)$ matrix with (j,l) element γ_{j+l-2}^+ for $j,l = 1,...,p+1$. Then,

$$D_{n+} \xrightarrow{p} \left(\Gamma_+^*\right)^{-1}$$

Similarly,

$$D_{n-} \xrightarrow{p} \left(\Gamma_{-}^{*}\right)^{-1}$$

where Γ_{-}^{*} is the $(p+1)\times(p+1)$ matrix with (j,l) element γ_{j+l-2}^{-} for $j,l=1,...,p+1$, and $\gamma_{j}^{-}=(-1)^{j}\int_{0}^{\infty}k\left(u\right)u^{j}f_{o}(\overline{x}-uh)du$.

Now we will derive the asymptotic distribution of $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{x_i-\overline{x}}{h}\right)d_iZ_i\varepsilon_i$. Following Porter (2003) I use the Cramer-Wold device to derive the asymptotic distribution. Let λ be a nonzero, finite vector. Then,

$$\begin{aligned}
& E\left[\left|\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{x_i-\overline{x}}{h}\right)d_i\lambda'Z_i\varepsilon_i\right|^{2+\zeta}\right] \\
= & \sum_{i=1}^{n}\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\frac{1}{nh}E\left[\left|k\left(\frac{x_i-\overline{x}}{h}\right)\right|^{2+\zeta}d_i\left|\lambda'Z_i\right|^{2+\zeta}\left|\varepsilon_i\right|^{2+\zeta}\right] \\
= & \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\frac{1}{h}E\left[\left|k\left(\frac{x-\overline{x}}{h}\right)\right|^{2+\zeta}d\left|\sum_{l=1}^{p}\lambda_l\left(\frac{x-\overline{x}}{h}\right)^{l}\right|^{2+\zeta}E\left[\left|\varepsilon\right|^{2+\zeta}\mid X=x\right]\right] \\
\leq & \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\frac{1}{h}E\left[\left|k\left(\frac{x-\overline{x}}{h}\right)\right|^{2+\zeta}d\left(\sum_{l=1}^{p}\left|\lambda_l\left(\frac{x-\overline{x}}{h}\right)^{l}\right|\right)^{2+\zeta}E\left[\left|\varepsilon\right|^{2+\zeta}\mid X=x\right]\right] \\
\leq & \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\frac{1}{h}\sup_{x\in\aleph}E\left[\left|\varepsilon\right|^{2+\zeta}\mid X=x\right]E\left[\left|k\left(\frac{x-\overline{x}}{h}\right)\right|^{2+\zeta}d\sum_{l=1}^{p}\left|\lambda_l\left(\frac{x-\overline{x}}{h}\right)^{l}\right|^{2+\zeta}\right] \\
= & \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\frac{1}{h}\sup_{x\in\aleph}E\left[\left|\varepsilon\right|^{2+\zeta}\mid X=x\right]\int_{\overline{x}}^{\overline{x}+h}\left|k\left(\frac{x-\overline{x}}{h}\right)\right|^{2+\zeta}\sum_{l=1}^{p}\left|\lambda_l\left(\frac{x-\overline{x}}{h}\right)^{l}\right|^{2+\zeta}dF_o(x) \\
= & \left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}O(1)O(1)=O\left(\left(\frac{1}{nh}\right)^{\frac{\zeta}{2}}\right)=o(1)
\end{aligned}$$

then, $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{x_i-\overline{x}}{h}\right)d_iZ_i\varepsilon_i$ follows Liapunov's CLT. Note that,

$$\begin{aligned}
& E\left[\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{x_i-\overline{x}}{h}\right)d_iZ_i\varepsilon_i\right] \\
= & E\left[\frac{\sqrt{n}}{\sqrt{h}}k\left(\frac{x_i-\overline{x}}{h}\right)d_iZ_iE\left[\varepsilon\mid X=x\right]\right]=0
\end{aligned}$$

and

$$\begin{aligned}
& Var\left[\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i\right] \\
= & \frac{1}{h} Var\left[k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i\right] \\
= & \frac{1}{h} E\left[k^2\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i Z_i' \varepsilon_i^2\right] \\
= & \frac{1}{h} E\left[k^2\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i Z_i' E\left[\varepsilon_i^2 \mid X=x\right]\right] \\
= & \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right) Z Z' \sigma^2(x) f_o(x) dx
\end{aligned}$$

It helps to remember that $Z_i Z_i'$ is a function of the x,

$$Z_i Z_i' = \begin{bmatrix} 1 & \left(\frac{x_i-\overline{x}}{h}\right) & \cdots & \left(\frac{x_i-\overline{x}}{h}\right)^p \\ \left(\frac{x_i-\overline{x}}{h}\right) & \left(\frac{x_i-\overline{x}}{h}\right)^2 & \cdots & \left(\frac{x_i-\overline{x}}{h}\right)^{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{x_i-\overline{x}}{h}\right)^p & \left(\frac{x_i-\overline{x}}{h}\right)^{p+1} & \cdots & \left(\frac{x_i-\overline{x}}{h}\right)^{2p} \end{bmatrix}$$

Let $\delta_j^+ = \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right)\left(\frac{x-\overline{x}}{h}\right)^j \sigma^2(x) f_o(x) dx = \int_0^\infty k^2(u)\, u^j \sigma^2(\overline{x}+uh) f_o(\overline{x}+uh) du$ and Δ_+^* is the $(p+1)\times(p+1)$ matrix with (j,l) element δ_{j+l-2}^+ for $j,l = 1,...,p+1$. Then,

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i \xrightarrow{p} N(0, \Delta_+^*)$$

Similarly we can show that

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) (1-d_i) Z_i \varepsilon_i \xrightarrow{p} N(0, \Delta_-^*)$$

where Δ_-^* is the $(p+1)\times(p+1)$ matrix with (j,l) element δ_{j+l-2}^- for $j,l = 1,...,p+1$, and $\delta_j^- = \int_{\overline{x}-h}^{\overline{x}} \frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right)\left(\frac{x-\overline{x}}{h}\right)^j \sigma^2(x) f_o(x) dx = (-1)^j \int_0^\infty k^2(u)\, u^j \sigma^2(\overline{x}-uh) f_o(\overline{x}-uh) du$.

The bias term is given by

$$\sqrt{nh} e_1' \left\{ D_{n+}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i m(x_i)\right] - D_{n-}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i) Z_i m(x_i)\right]\right\}$$

Notice that if the rectangular kernel is used this is nothing else than the difference between the intercepts estimated by the linear projection of $m(x)$ on Z, above and below the cutoff point using only the data inside the bandwidth.

Note that,

$$\begin{aligned} E\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i m(x_i)\right] &= E\left[\frac{1}{h}k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i m(x_i)\right] \\ &= \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h}k\left(\frac{x-\overline{x}}{h}\right) Z(x)m(x)f(x)dx \\ &= \int_{0}^{\infty} k\left(u\right) Z(\overline{x}+uh)m(\overline{x}+uh)f(\overline{x}+uh)du \end{aligned}$$

and similarly,

$$\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i) Z_i m(x_i) \xrightarrow{p} \int_{\overline{x}-h}^{\overline{x}} \frac{1}{h}k\left(\frac{x-\overline{x}}{h}\right) Z(x)m(x)dx$$

Hence, the bias term can be approximated by,

$$e_1' \left\{ \begin{array}{l} \left(\Gamma_+^*\right)^{-1}\left[\int_0^\infty k\left(u\right) Z(\overline{x}+uh)m(\overline{x}+uh)f(\overline{x}+uh)du\right] - \\ -\left(\Gamma_-^*\right)^{-1}\left[\int_0^\infty k\left(u\right) Z(\overline{x}-uh)m(\overline{x}-uh)f(\overline{x}-uh)du\right] \end{array} \right\}$$

■

Proof. [Proof of Corollary 2] First, note that, if $h \to 0$,

$$\begin{aligned} \gamma_j^+ &= \lim_{h\to 0}\int_0^\infty k\left(u\right) u^j f_o(\overline{x}+uh)du \\ &= f_o(\overline{x})\int_0^\infty k\left(u\right) u^j du \\ &= f_o(\overline{x})\gamma_j \end{aligned} \tag{C.1}$$

and

$$\begin{aligned} \delta_j^+ &= \lim_{h\to 0}\int_0^\infty k^2\left(u\right) u^j \sigma^{2+}(\overline{x}+uh) f_o(\overline{x}+uh)du \\ &= \sigma^{2+}(\overline{x}) f_o(\overline{x})\int_0^\infty k^2\left(u\right) u^j du \\ &= \sigma^{2+}(\overline{x}) f_o(\overline{x})\delta_j \end{aligned} \tag{C.2}$$

and similarly for γ_j^- and δ_j^-. Then, for the variance,

$$\begin{aligned}
&\lim_{h\to 0} \left(\Gamma_+^*\right)^{-1} \Delta_+^* \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^* \left(\Gamma_-^*\right)^{-1} \\
= \;& (f_o(\overline{x})\Gamma)^{-1} \left[\sigma^{2+}(\overline{x}) f_o(\overline{x})\Delta\right] (f_o(\overline{x})\Gamma)^{-1} + (f_o(\overline{x})\Gamma)^{-1} \left[\sigma^{2-}(\overline{x}) f_o(\overline{x})\Delta\right] (f_o(\overline{x})\Gamma)^{-1} \\
= \;& \frac{\sigma^{2+}(\overline{x}) + \sigma^{2-}(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1
\end{aligned}$$

For the bias, if we approximate $m(\overline{x}+uh) = m(x)$ just above $m(\overline{x})$:

$$m(x) = LP^+(m(x) \text{ on } Z(x)) + \frac{1}{(p+1)!} m^{(p+1)+}(\overline{x}) (x-\overline{x})^{p+1} + o\left(h^{p+1}\right)$$

and similarly for approximating $m(x)$ just below the cutoff. When we evaluate $LP^+(m(x)$ on $Z(x))$ at $\overline{x}$, we get the intercept $m(\overline{x})$ and the "residual" as described above. A helpful fact is that, by the definition of $Z(x)$,

$$\int_0^\infty k(u) Z(\overline{x}+uh) u^{p+1} du = \int_0^\infty k(u) \begin{bmatrix} 1 \\ \vdots \\ u^p \end{bmatrix} u^{p+1} du = \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix} \tag{C.3}$$

$$\int_0^\infty k(u) Z(\overline{x}-uh) u^{p+1} du = \int_0^\infty k(u) \begin{bmatrix} 1 \\ \vdots \\ (-u)^p \end{bmatrix} u^{p+1} du = \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ (-1)^p \gamma_{2p+1} \end{bmatrix} \tag{C.4}$$

Note that $\Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}$ is equal both above and below the cutoff. The bias formula in theorem 6 is given by

$$e_1' \left\{ \begin{array}{l} \left(\Gamma_+^*\right)^{-1} \left[\int_0^\infty k(u) Z(\overline{x}+uh) m(\overline{x}+uh) f_o(\overline{x}+uh) du\right] - \\ -\left(\Gamma_-^*\right)^{-1} \left[\int_0^\infty k(u) Z(\overline{x}-uh) m(\overline{x}-uh) f_o(\overline{x}-uh) du\right] \end{array} \right\} \tag{C.5}$$

as discussed in section 2.5 the main term is just the difference between the intercepts of the linear projections of $k(u) m(x)$ on $k(u) Z(x)$ in the bandwidth above below the cutoff, which is equal to the linear projections evaluated at $\overline{x}$. Hence, plugging the bias formula for the linear projection, formula (C.5) can be written as (plus an $o\left(h^{p+1}\right)$ term),

$$\begin{aligned}
& e_1' \left[\left(\Gamma_+^*\right)^{-1} \int_0^\infty k(u) Z(\overline{x}+uh) \left(\frac{1}{(p+1)!} m^{(p+1)+}(\overline{x}) (uh)^{p+1}\right) f_o(\overline{x}+uh) du\right] \\
& -e_1' \left[\left(\Gamma_-^*\right)^{-1} \int_0^\infty k(u) Z(\overline{x}-uh) \left(\frac{(-1)^{p+1}}{(p+1)!} m^{(p+1)-}(\overline{x}) (uh)^{p+1}\right) f_o(\overline{x}-uh) du\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{h^{p+1}}{(p+1)!} e_1' \left[\left(\Gamma_+^*\right)^{-1} \int_0^\infty k(u) Z(\overline{x}+uh) m^{(p+1)+}(\overline{x}) u^{p+1} f_o(\overline{x}+uh) du \right] \\
&\quad - \frac{h^{p+1}}{(p+1)!} e_1' (-1)^{p+1} \left[\left(\Gamma_-^*\right)^{-1} \int_0^\infty k(u) Z(\overline{x}-uh) m^{(p+1)-}(\overline{x}) u^{p+1} f_o(\overline{x}-uh) du \right] \\
&= \frac{h^{p+1}}{(p+1)!} e_1' m^{(p+1)+}(\overline{x}) \left[\left(\Gamma_+^*\right)^{-1} \int_0^\infty k(u) Z(\overline{x}+uh) u^{p+1} f_o(\overline{x}+uh) du \right] \\
&\quad - \frac{h^{p+1}}{(p+1)!} e_1' (-1)^{p+1} m^{(p+1)-}(\overline{x}) \left[\left(\Gamma_-^*\right)^{-1} \int_0^\infty k(u) Z(\overline{x}-uh) u^{p+1} f_o(\overline{x}-uh) du \right]
\end{aligned}
$$

where $\begin{pmatrix} m(\overline{x}) \\ m_{p+(-)} \end{pmatrix}$, is the vector of coefficients of the linear projection of $m(x)$ on $Z(x)$ is the bandwidth above (below) the cutoff. If $h \to 0$, using the equalities in equations (C.3), (C.4), (C.2) and (C.1),

$$
= \frac{\lim_{h\to 0}\left(h^{p+1}\right)}{(p+1)!} \left[m^{(p+1)+}(\overline{x}) - (-1)^{p+1} m^{(p+1)-}(\overline{x}) \right] e_1' \Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}
$$

■

Proof. [Proof of Corollary 3] First, note that, if $h > 0$ and, in the bandwidth around the cutoff, $f_o(x) = f_o(\overline{x})$, $\sigma^2(x) = \sigma^2(\overline{x})$ and

$$
m(x) = m(\overline{x}) + m'^{+}(\overline{x})(x-\overline{x}) + ... + \frac{1}{p!} m^{(p)+}(\overline{x})(x-\overline{x})^p + \frac{1}{(p+1)!} m^{(p+1)+}(\overline{x})(x-\overline{x})^{p+1}
$$

then,

$$
\begin{aligned}
\gamma_j^+ &= \int_0^\infty k(u) u^j f_o(\overline{x}+uh) du \\
&= f_o(\overline{x}) \int_0^\infty k(u) u^j du \\
&= f_o(\overline{x}) \gamma_j
\end{aligned} \tag{C.6}
$$

and

$$
\begin{aligned}
\delta_j^+ &= \int_0^\infty k^2(u) u^j \sigma^2(\overline{x}+uh) f_o(\overline{x}+uh) du \\
&= \sigma^{2+}(\overline{x}) f_o(\overline{x}) \int_0^\infty k^2(u) u^j du \\
&= \sigma^{2+}(\overline{x}) f_o(\overline{x}) \delta_j
\end{aligned} \tag{C.7}
$$

and similarly for γ_j^- and δ_j^-. Then, for the variance,

$$\begin{aligned}
& \left(\Gamma_+^*\right)^{-1}\Delta_+^*\left(\Gamma_+^*\right)^{-1}+\left(\Gamma_-^*\right)^{-1}\Delta_-^*\left(\Gamma_-^*\right)^{-1} \\
= \; & (f_o(\overline{x})\Gamma)^{-1}\left[\sigma^{2+}(\overline{x})f_o(\overline{x})\Delta\right](f_o(\overline{x})\Gamma)^{-1}+(f_o(\overline{x})\Gamma)^{-1}\left[\sigma^{2-}(\overline{x})f_o(\overline{x})\Delta\right](f_o(\overline{x})\Gamma)^{-1} \\
= \; & \frac{\sigma^{2+}(\overline{x})+\sigma^{2-}(\overline{x})}{f_o(\overline{x})}e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1
\end{aligned}$$

For the bias, the strategy is basically the same as in the proof of corollary 2:

$$m(x)=LP^+(m(x)\text{ on }Z(x))+\frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})\,(x-\overline{x})^{p+1}$$

and similarly for approximating $m(x)$ just below the cutoff. When we evaluate $LP^+(m(x)$ on $Z(x))$ at $\overline{x}$, we get the intercept $m(\overline{x})$ and the "residual" as described above. Once again, using formulas (C.3) and (C.4), the bias formula in theorem 6 is given by

$$e_1'\left\{\begin{array}{l}\left(\Gamma_+^*\right)^{-1}\left[\int_0^\infty k\left(u\right)Z(\overline{x}+uh)m(\overline{x}+uh)f_o(\overline{x}+uh)du\right]- \\ -\left(\Gamma_-^*\right)^{-1}\left[\int_0^\infty k\left(u\right)Z(\overline{x}-uh)m(\overline{x}-uh)f_o(\overline{x}-uh)du\right]\end{array}\right\}$$

as discussed in section 2.5 the main term is just the difference between the intercepts of the linear projections of $k\left(u\right)m(x)$ on $k\left(u\right)Z(x)$ in the bandwidth above below the cutoff, which is equal to the linear projections evaluated at $\overline{x}$. Hence, plugging the bias formula for the linear projection:

$$\begin{aligned}
& e_1'\left[\left(\Gamma_+^*\right)^{-1}\int_0^\infty k\left(u\right)Z(\overline{x}+uh)\left(\frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})\,(uh)^{p+1}\right)f_o(\overline{x}+uh)du\right] \\
& -e_1'\left[\left(\Gamma_-^*\right)^{-1}\int_0^\infty k\left(u\right)Z(\overline{x}-uh)\left(\frac{(-1)^{p+1}}{(p+1)!}m^{(p+1)-}(\overline{x})\,(uh)^{p+1}\right)f_o(\overline{x}-uh)du\right] \\
= \; & \frac{h^{p+1}}{(p+1)!}e_1'\left[\left(\Gamma_+^*\right)^{-1}\int_0^\infty k\left(u\right)Z(\overline{x}+uh)m^{(p+1)+}(\overline{x})u^{p+1}f_o(\overline{x}+uh)du\right] \\
& -\frac{h^{p+1}}{(p+1)!}e_1'(-1)^{p+1}\left[\left(\Gamma_-^*\right)^{-1}\int_0^\infty k\left(u\right)Z(\overline{x}-uh)m^{(p+1)-}(\overline{x})u^{p+1}f_o(\overline{x}-uh)du\right] \\
= \; & \frac{h^{p+1}}{(p+1)!}e_1'm^{(p+1)+}(\overline{x})\left[\left(\Gamma_+^*\right)^{-1}\int_0^\infty k\left(u\right)Z(\overline{x}+uh)u^{p+1}f_o(\overline{x}+uh)du\right] \\
& -\frac{h^{p+1}}{(p+1)!}e_1'\left(-1\right)^{p+1}m^{(p+1)-}(\overline{x})\left[\left(\Gamma_-^*\right)^{-1}\int_0^\infty k\left(u\right)Z(\overline{x}-uh)u^{p+1}f_o(\overline{x}-uh)du\right]
\end{aligned}$$

where $\begin{pmatrix} m(\overline{x}) \\ m_{p+(-)} \end{pmatrix}$, is the vector of coefficients of the linear projection of $m(x)$ on $Z(x)$ is the bandwidth above (below) the cutoff. Using $f_o(x) = f_o(\overline{x})$ and the equalities in formulas (C.3), (C.4), (C.7) and (C.6),

$$= \frac{h^{p+1}}{(p+1)!} \left[m^{(p+1)+}(\overline{x}) - (-1)^{p+1} m^{(p+1)-}(\overline{x}) \right] e_1' \Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}$$

■

Proof. [Proof of Theorem 8] To obtain the Covariance term for the asymptotic variance of the Fuzzy Regression Discontinuity estimator, note that the covariance will be determined by the expectation of the product of the stochastic terms.

The covariance between the estimators for the outcome of interest and the treatment probability will be given by two independent terms, one for each side of the threshold. The upper side is given by

$$E \left\{ e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i y_i \right] \left[\frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i t_i \right] D_{n+} e_1 \right\}$$

Where t_i is the dummy variable indicating that the observation has received treatment. In obtaining the asymptotic covariance, the bias term of the estimator can be ignored, hence

$$\begin{aligned} & E \left[e_1' D_{n+} \left(\frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i \varepsilon_i \right) \left(\sum_{i=1}^{n} k \left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i \eta_i \right)' D_{n+} e_1 \right] \\ = \; & E \left[e_1' D_{n+} \left(\frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{n} k \left(\frac{x_i - \overline{x}}{h} \right) k \left(\frac{x_j - \overline{x}}{h} \right) d_i d_j Z_i Z_j' \varepsilon_i \eta_j \right) D_{n+} e_1 \right] \end{aligned}$$

$$
\begin{aligned}
&= E\left[e_1' D_{n+}\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)^2 d_i Z_i Z_i' E\left[\varepsilon_i\eta_i \mid X=x\right]\right) D_{n+}e_1\right] \\
&= E\left[e_1' D_{n+}\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right)^2 d_i Z_i Z_i' \sigma_{\varepsilon\eta}(x_i)\right) D_{n+}e_1\right] \\
&= E\left[e_1' D_{n+}\left(\frac{1}{h} k\left(\frac{x_i-\overline{x}}{h}\right)^2 d_i Z_i Z_i' \sigma_{\varepsilon\eta}(x_i)\right) D_{n+}e_1\right] \\
&= e_1' D_{n+}\left[\int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k\left(\frac{x-\overline{x}}{h}\right)^2 ZZ'\sigma_{\varepsilon\eta}(x) f_o(x) dx\right] D_{n+}e_1
\end{aligned}
$$

where I used the assumption that $E\left[\varepsilon_i\eta_j \mid X=x\right]=0$ for $j\neq i$.

Similarly for the second term,

$$
e_1' D_{n-}\left[\int_{\overline{x}-h}^{\overline{x}} \frac{1}{h} k\left(\frac{x-\overline{x}}{h}\right)^2 ZZ'\sigma_{\varepsilon\eta}(x) f_o(x) dx\right] D_{n-}e_1
$$

Let $\rho_j^+ = \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h}k^2\left(\frac{x-\overline{x}}{h}\right)\left(\frac{x-\overline{x}}{h}\right)^j \sigma_{\varepsilon\eta}(x) f_o(x) dx = \int_0^\infty k^2(u)\, u^j \sigma_{\varepsilon\eta}(\overline{x}+uh) f_o(\overline{x}+uh) du$, Δ_+^ρ is the $(p+1)\times(p+1)$ matrix with (j,l) element ρ_{j+l-2}^+ for $j,l=1,...,p+1$, $\rho_j^- = \int_{\overline{x}-h}^{\overline{x}} \frac{1}{h}k^2\left(\frac{x-\overline{x}}{h}\right)\left(\frac{x-\overline{x}}{h}\right)^j \sigma_{\varepsilon\eta}(x) f_o(x) dx = (-1)^j \int_0^\infty k^2(u)\, u^j \sigma_{\varepsilon\eta}(\overline{x}-uh) f_o(\overline{x}-uh) du$ and Δ_-^ρ is the $(p+1)\times(p+1)$ matrix with (j,l) element ρ_{j+l-2}^- for $j,l=1,...,p+1$ Then the asymptotic covariance is given by

$$
C_{\alpha\theta} = e_1'\left[\left(\Gamma_+^*\right)^{-1}\Delta_+^\rho\left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1}\Delta_-^\rho\left(\Gamma_-^*\right)^{-1}\right] e_1
$$

The asymptotic covariance for the Nadaraya-Watson estimator will be given by the special case when $p=0$.

$$
C_{\alpha\theta} = \int_0^\infty k^2(u)\left(\frac{\sigma_{\varepsilon\eta}(\overline{x}+uh) f_o(\overline{x}+uh)}{\left(\int_0^\infty k(u) f_o(\overline{x}+uh) du\right)^2} + \frac{\sigma_{\varepsilon\eta}(\overline{x}-uh) f_o(\overline{x}-uh)}{\left(\int_0^\infty k(u) f_o(\overline{x}-uh) du\right)^2}\right) du
$$

■

Proof. [Proof of Corollary 4] Using the results in equations C.1 and noting that, if $h\to 0$

$$
\begin{aligned}
\rho_j^+ &= \lim_{h\to 0}\int_0^\infty k^2(u)\, u^j \sigma_{\varepsilon\eta}(\overline{x}+uh) f_o(\overline{x}+uh) du \\
&= \sigma_{\varepsilon\eta}^+(\overline{x}) f_o(\overline{x}) \delta_j
\end{aligned}
$$

and similarly for ρ_j^-. Then,

$$\begin{aligned}
&\lim_{h\to 0} e_1' \left[\left(\Gamma_+^*\right)^{-1} \Delta_+^\rho \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^\rho \left(\Gamma_-^*\right)^{-1} \right] e_1 \\
= \; & e_1' \left[(f_o(\overline{x})\Gamma)^{-1} \sigma_{\varepsilon\eta}^+(\overline{x}) f_o(\overline{x}) \Delta \left(f_o(\overline{x})\Gamma\right)^{-1} + (f_o(\overline{x})\Gamma)^{-1} \sigma_{\varepsilon\eta}^-(\overline{x}) f_o(\overline{x}) \Delta \left(f_o(\overline{x})\Gamma\right)^{-1} \right] e_1 \\
= \; & \frac{\sigma_{\varepsilon\eta}^+(\overline{x}) + \sigma_{\varepsilon\eta}^-(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1
\end{aligned}$$

■

Proof. [Proof of Corollary 5] The proof follows very closely corollary 4. Using the results in equations C.1 and noting that, if $h > 0$ and $f_o(x) = f_o(\overline{x})$ and $\sigma_{\varepsilon\eta}(x) = \sigma_{\varepsilon\eta}(\overline{x})$ for any x in the range around the cutoff

$$\begin{aligned}
\rho_j^+ &= \int_0^\infty k^2(u)\, u^j \sigma_{\varepsilon\eta}(\overline{x} + uh) f_o(\overline{x} + uh) du \\
&= \sigma_{\varepsilon\eta}^+(\overline{x}) f_o(\overline{x}) \delta_j
\end{aligned}$$

and similarly for ρ_j^-. Then,

$$\begin{aligned}
& e_1' \left[\left(\Gamma_+^*\right)^{-1} \Delta_+^\rho \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^\rho \left(\Gamma_-^*\right)^{-1} \right] e_1 \\
= \; & e_1' \left[(f_o(\overline{x})\Gamma)^{-1} \sigma_{\varepsilon\eta}^+(\overline{x}) f_o(\overline{x}) \Delta \left(f_o(\overline{x})\Gamma\right)^{-1} + (f_o(\overline{x})\Gamma)^{-1} \sigma_{\varepsilon\eta}^-(\overline{x}) f_o(\overline{x}) \Delta \left(f_o(\overline{x})\Gamma\right)^{-1} \right] e_1 \\
= \; & \frac{\sigma_{\varepsilon\eta}^+(\overline{x}) + \sigma_{\varepsilon\eta}^-(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1
\end{aligned}$$

■

APPENDIX D

Proofs to "Asymptotic Properties of Quantile Regression for Standard Stratified Samples"

Proof. [Proof of Corollary 6] The general form of the variance follows directly from Newey and McFadden (1994) theorem 7.1. The specific formulas for A_w and B_w, are obtained by checking that the proof used by Wooldridge (2001) still holds for the estimator that minimizes the objective function given by equation 3.2. I follow his procedure below.

Since within each stratum we have a i.i.d. sequence $\left\{w_{ij} : i = 1, 2, ..., N_j\right\}$ for each j, a CLT for i.i.d. observations can be applied for each stratum.

$$N_j^{-\frac{1}{2}} \sum_{i=1}^{N_j} \left[s_{ij}\left(\beta_{\tau o}\right) - \mu_j \right] \xrightarrow{d} N(0, B_j)$$

where $s_{ij}\left(\beta_{\tau o}\right) \equiv \left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g}_i$, with $\overset{\bullet}{g}_i \equiv \frac{\partial g\left(x_{ij}, \beta\right)}{\partial \beta} \mid_{\beta = \beta_{\tau o}}$, $\mu_j \equiv E\left[s_{ij}\left(\beta_{\tau o}\right)\right] = E\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g}_i | w \in W_j\right]$, and $B_j \equiv Var\left[s_{ij}\left(\beta_{\tau o}\right)\right] = Var\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g}_i | w \in W_j\right]$. As seen in equation

3.4,

$$\sum_{j=1}^{J} Q_j \mu_j = 0$$

The score of the objective function, multiplied by $N^{\frac{1}{2}}$ and evaluated at β_{τ_o} can be written as

$$\begin{aligned} N^{-\frac{1}{2}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \left[\sum_{i=1}^{N_j} s_{ij}\left(\beta_{\tau_o}\right) \right] &= N^{-\frac{1}{2}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \left[\sum_{i=1}^{N_j} s_{ij}\left(\beta_{\tau_o}\right) - \mu_j \right] \\ &= \sum_{j=1}^{J} \frac{Q_j}{H_j^{\frac{1}{2}}} \left[N_j^{-\frac{1}{2}} \sum_{i=1}^{N_j} s_{ij}\left(\beta_{\tau_o}\right) - \mu_j \right] \xrightarrow{d} N(0, B_w) \end{aligned}$$

with $B_w = \sum_{j=1}^{J} \frac{Q_j^2}{H_j} B_j = \sum_{j=1}^{J} \frac{Q_j^2}{H_j} Var\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau_o}\right) \leq 0 \right] \right) \overset{\bullet}{g}_i | w \in W_j \right]$. Where I used the fact that sampling is random within stratum and the observations are independent across strata. In the linear case this formula simplifies to $B_w = \sum_{j=1}^{J} \frac{Q_j^2}{H_j} B_j = \sum_{j=1}^{J} \frac{Q_j^2}{H_j} Var\left[\left(\tau - 1\left[y_{ij} - x_{ij}' \beta_{\tau_o} \leq 0 \right] \right) x_{ij} | w \in W_j \right]$.

For the outer part of the variance matrix it is enough to note that

$$\begin{aligned}
A_w &= \nabla_\beta \sum_{j=1}^{J} Q_j E[\left(\tau - 1\left[y - g\left(x, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g} | w \in W_j] \\
&= \sum_{j=1}^{J} Q_j \nabla_\beta E[\left(\tau - 1\left[y - g\left(x, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g} | w \in W_j] \\
&= \sum_{j=1}^{J} Q_j \nabla_\beta E[\left(\tau - 1\left[y - g\left(x, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g} | w \in W_j] \\
&= \sum_{j=1}^{J} Q_j \nabla_\beta E\left[E\left[\tau - 1\left[y - g\left(x, \beta_{\tau o}\right) \leq 0\right] | x, w \in W_j\right] \overset{\bullet}{g} | w \in W_j\right] \\
&= \sum_{j=1}^{J} Q_j \nabla_\beta E\left[\left(\tau - F_{y|x,w \in W_j}\left(g\left(x, \beta_{\tau o}\right)\right)\right) \overset{\bullet}{g} | w \in W_j\right] \\
&= \sum_{j=1}^{J} Q_j E[f_{y|x,w \in W_j}\left(g\left(x, \beta_{\tau o}\right)\right) \overset{\bullet}{g}\overset{\bullet}{g}' | w \in W_j] \\
&= E\left[f_{y|x}\left(g\left(x, \beta_{\tau o}\right)\right) \overset{\bullet}{g}\overset{\bullet}{g}'\right]
\end{aligned}$$

And that the Jacobian of $\sum_{j=1}^{J} Q_j \left[\frac{1}{N_j} \sum_{i=1}^{N_j} \left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_\tau\right) \leq 0\right]\right) \overset{\bullet}{g}_i\right]$ converges in probability uniformly to A_w. In the linear case this formula simplifies to $A_w = \sum_{j=1}^{J} Q_j E[f_{y|x,w \in W_j}\left(x'\beta_{\tau o}\right) xx' | w \in W_j] = E\left[f_{y|x}(x'\beta_{\tau o}) xx'\right]$. ■

Proof. [Proof of Corollary 7] The general form of the variance follows directly from Newey and McFadden (1994) theorem 7.1. Under exogenous stratification, the following changesneed to be made to the definitions used to prove corollary 6: (a) $\mu_j \equiv E\left[s_{ij}\left(\beta_{\tau o}\right)\right] = E\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g}_i | x \in \chi_j\right] = 0$ for every stratum j; (b) $B_j \equiv Var\left[s_{ij}\left(\beta_{\tau o}\right)\right] = Var\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau o}\right) \leq 0\right]\right) \overset{\bullet}{g}_i | x \in \chi_j\right]$. Then, the

score of the objective function, multiplied by $N^{\frac{1}{2}}$ and evaluated at β_{τ_o} can be written as

$$\begin{aligned} N^{-\frac{1}{2}}\sum_{i=1}^{N} s_{ij}\left(\beta_{\tau_o}\right) &= \sum_{j=1}^{J} H_j^{\frac{1}{2}}\left[N_j^{-\frac{1}{2}}\sum_{i=1}^{N_j} s_{ij}\left(\beta_{\tau_o}\right)\right] \\ &= \sum_{j=1}^{J} H_j^{\frac{1}{2}}\left[N_j^{-\frac{1}{2}}\sum_{i=1}^{N_j} s_{ij}\left(\beta_{\tau_o}\right)\right] \xrightarrow{d} N(0, B_u) \end{aligned}$$

with $B_u = \sum_{j=1}^{J} \overline{H}_j B_j = \sum_{j=1}^{J} \overline{H}_j Var\left[\left(\tau - 1\left[y_{ij} - g\left(x_{ij}, \beta_{\tau_o}\right) \leq 0\right]\right)\overset{\bullet}{g}_i | x \in \chi_j\right]$. Where I used the fact that sampling is random within stratum and the observations are independent across strata. In the linear case this formula simplifies to $B_u = \sum_{j=1}^{J} \overline{H}_j B_j = \sum_{j=1}^{J} \overline{H}_j Var\left[\left(\tau - 1\left[y_{ij} - x_{ij}'\beta_{\tau_o} \leq 0\right]\right)x_{ij} | x \in \chi_j\right]$.

Note that in this case, since $E\left[1\left[y_{ij} - x_{ij}'\beta_{\tau_o} \leq 0\right]|x\right]$ is the same to every stratum,

$$Var\left[\left(\tau - 1\left[y_{ij} - x_{ij}'\beta_{\tau_o} \leq 0\right]\right)\overset{\bullet}{g}_i | x \in \chi_j\right] = \tau(1-\tau)E\left[\overset{\bullet}{g}_i\overset{\bullet}{g}_i' | x \in \chi_j\right]$$

and $B_u = \sum_{j=1}^{J} \overline{H}_j \tau(1-\tau)E\left[\overset{\bullet}{g}_i\overset{\bullet}{g}_i' | x \in \chi_j\right] = \tau(1-\tau)\sum_{j=1}^{J} \overline{H}_j E\left[\overset{\bullet}{g}_i\overset{\bullet}{g}_i' | x \in \chi_j\right]$. In the in the linear CQF case, $B_u = \sum_{j=1}^{J} \overline{H}_j \tau(1-\tau)E\left[\overset{\bullet}{g}_i\overset{\bullet}{g}_i' | x \in \chi_j\right] = \tau(1-\tau)\sum_{j=1}^{J} \overline{H}_j E\left[xx' | x \in \chi_j\right]$.

For the outer part of the variance matrix it is enough to note that

$$\begin{aligned} A_u &= \sum_{j=1}^{J} \overline{H}_j \nabla_\beta E[\left(\tau - 1\left[y - g\left(x, \beta_{\tau_o}\right) \leq 0\right]\right)\overset{\bullet}{g} | x \in \chi_j] \\ &= \sum_{j=1}^{J} \overline{H}_j \nabla_\beta E\left[E\left[\tau - 1\left[y - g\left(x, \beta_{\tau_o}\right) \leq 0\right]|x\right]\overset{\bullet}{g} | x \in \chi_j\right] \\ &= \sum_{j=1}^{J} \overline{H}_j \nabla_\beta E\left[\left(\tau - F_{y|x}\left(g\left(x, \beta_{\tau_o}\right)\right)\right)\overset{\bullet}{g} | x \in \chi_j\right] \\ &= \sum_{j=1}^{J} \overline{H}_j E[f_{y|x}\left(g\left(x, \beta_{\tau_o}\right)\right)\overset{\bullet}{g}\overset{\bullet}{g}' | x \in \chi_j] \end{aligned}$$

In the linear case this formula simplifies to $A_u = \sum_{j=1}^{J} \overline{H}_j E[f_{y|x}\left(x_{ij}'\widehat{\beta}_{\tau_o}\right)xx' | x \in \chi_j]$. ■

Proof. [Proof of Corollary 8] The asymptotic multivariate normality result follows directly from the use of a standard Cramer-Wold device argument for the vector of the scores for each quantile applied separately for each stratum. Let $s_{ij}(\beta_\tau) = \left[s_{1_{ij}}(y,x,\beta_{\tau_1})', s_{2_{ij}}(y,x,\beta_{\tau_2})', \ldots, s_{p_{ij}}(y,x,\beta_{\tau_p})'\right]$ and $\mu_j \equiv E\left[s_{ij}(\beta_\tau)\right] = E\left[s_{1_{ij}}(y,x,\beta_{\tau_1})', s_{2_{ij}}(y,x,\beta_{\tau_2})', \ldots, s_{p_{ij}}(y,x,\beta_{\tau_p})'\right]$ then,

$$N_j^{-\frac{1}{2}} \sum_{i=1}^{N_j} \left[s_{ij}(\beta_\tau) - \mu_j\right] \xrightarrow{d} N(0, B_j)$$

with B_j a $p \times p$ variance covariance matrix with typical element,

$$\begin{aligned} B_{j_{l,k}} &\equiv Cov\left[s_{ij}\left(\beta_{\tau_l}\right), s_{ij}\left(\beta_{\tau_k}\right)\right] \\ &= Cov\left[\left(\tau_l - 1\left[y - g(x,\beta_{\tau_l}) \leq 0\right]\right)\overset{\bullet}{g}_l, \left(\tau_k - 1\left[y - g(x,\beta_{\tau_k}) \leq 0\right]\right)\overset{\bullet}{g}_k | w \in W_j\right] \end{aligned}$$

Then,

$$\begin{aligned} N^{-\frac{1}{2}} \sum_{j=1}^{J} \frac{Q_j}{H_j}\left[\sum_{i=1}^{N_j} s_{ij}(\beta_\tau)\right] &= N^{-\frac{1}{2}} \sum_{j=1}^{J} \frac{Q_j}{H_j}\left[\sum_{i=1}^{N_j} s_{ij}(\beta_\tau) - \mu_j\right] \\ &= \sum_{j=1}^{J} \frac{Q_j}{H_j^{\frac{1}{2}}}\left[N_j^{-\frac{1}{2}} \sum_{i=1}^{N_j} s_{ij}(\beta_\tau) - \mu_j\right] \xrightarrow{d} N(0, B_s) \end{aligned}$$

where $B_s = \left\{B_{s_{l,k}}\right\}_{l,k=1,\ldots p}$ with typical element

$$B_{s_{l,k}} = \left[\sum_{j=1}^{J} \frac{Q_j^2}{\overline{\overline{H}}_j} B_{j_{l,k}}\right]$$

and the outer part of each term is given by the $A_{w_l} \equiv E\left[f_{y|x}(g(x,\beta_{\tau l}))\overset{\bullet}{g}_l\overset{\bullet}{g}_l'\right]$ and $A_{w_k} \equiv E\left[f_{y|x}(g\left(x,\beta_{\tau_k}\right))\overset{\bullet}{g}_k\overset{\bullet}{g}_k'\right]$ as argued in Buchinsky (1998).

Hence, $\sqrt{N}\left(\widehat{\beta}_\tau - \beta_\tau\right) \overset{a}{\sim} N(0, \Lambda_\tau)$, where $\Lambda_\tau = \left\{\Lambda_{\tau_{l,k}}\right\}_{l,k=1,\ldots p}$ with typical element defined as

$$\Lambda_{\tau_{l,k}} = E\left[f_{y|x}(g\left(x,\beta_{\tau_l}\right))\overset{\bullet}{g}_l\overset{\bullet}{g}_l'\right]^{-1}\left[\sum_{j=1}^{J} \frac{Q_j^2}{\overline{\overline{H}}_j} B_{j_{l,k}}\right] E\left[f_{y|x}(g\left(x,\beta_{\tau_k}\right))\overset{\bullet}{g}_k\overset{\bullet}{g}_k'\right]^{-1}$$

and, in the special case of the linear CQF, $g(X,\beta) = x'\beta$

$$\Lambda_{\tau_{l,k}} = E\left[f_{y|x}\left(x'\beta_{\tau_l}\right)xx'\right]^{-1}\left[\sum_{j=1}^{J}\frac{Q_j^2}{\overline{H}_j}B_{j_{l,k}}\right]E\left[f_{y|x}\left(x'\beta_{\tau_k}\right)xx'\right]^{-1}$$

■

BIBLIOGRAPHY

BIBLIOGRAPHY

S. Ahn and P. Schmidt. A separability result for gmm estimation, with applications to gls prediction and conditional moment tests. *Econometric reviews*, 14(1):19–34, 1995.

J. Albrecht, A. Björklund, and S. Vroman. Is there a glass ceiling in sweden? *Journal of Labor Economics*, 21(1):145–177, 2003.

D. Andrews. Empirical process methods in econometrics. In R. Engle and D. McFadden, editors, *Handbook of econometrics*, volume 4, pages 2247–2294. Elsevier, 1994.

J. Angrist and J. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009.

J. Angrist, V. Chernozhukov, and I. Fernández-Val. Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563, 2006.

M. Buchinsky. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources*, 33(1):88–126, 1998.

M. Buchinsky. Quantile regression with sample selection: Estimating women's return to education in the us. *Empirical Economics*, 26(1):87–113, 2001.

M. Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010.

X. Chen, O. Linton, and I. Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003.

X. Chen, H. Hong, and A. Tarozzi. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

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

V. Chernozhukov and C. Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.

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

H. Daniels. The asymptotic efficiency of a maximum likelihood estimator. In *Proceedings of the fourth Berkeley symposium on mathematical statististics and probability*, volume 1, pages 151–163. Berkeley: University of California Press, 1961.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman & Hall/CRC, 1996.

Y. Fan. Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory*, 14(5):604–621, 1998.

J. Hahn, P. Todd, and W. Van der Klaauw. Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. *NBER Working Paper Series*, 1999.

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

K. Hirano, G. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

K. Hitomi, Y. Nishiyama, and R. Okui. A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. *Econometric Theory*, 24(06):1717–1728, 2008.

P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–33, 1967.

G. Imbens and J. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

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

B. James. *Probabilidade: Um curso em nível intermediário*. IMPA, Rio de Janeiro, third edition, 2004.

R. Koenker. *Quantile regression*. Cambridge University Press, 2005.

D. Lee and T. Lemieux. Regression discontinuity designs in economics. Technical report, National Bureau of Economic Research, 2009.

W. Lee. Robust tests of hypotheses in models with m-estimation. *Working Paper, Department of economics, National Chung Cheng University*, 2008.

J. Machado and J. Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, 20(4):445–465, 2005.

P. Martins and P. Pereira. Does education reduce wage inequality? quantile regression evidence from 16 countries. *Labour Economics*, 11(3):355–371, 2004.

B. Melly. Decomposition of differences in distribution using quantile regression. *Labour Economics*, 12(4):577–590, 2005.

H. Neave. An improved formula for the asymptotic variance of spectrum estimates. *The Annals of Mathematical Statistics*, 41(1):70–77, 1970.

W. Newey and D. McFadden. Large sample estimation and hypothesis testing. In R. Engle and D. McFadden, editors, *Handbook of econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

A. Pagan and A. Ullah. *Nonparametric econometrics*. Cambridge University Press, 1999.

A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057, 1989.

D. Pollard. New ways to prove central limit theorems. *Econometric Theory*, 1(3): 295–313, 1985.

J. Porter. Estimation in the regression discontinuity model. *Unpublished manuscript, Department of economics, University of Wisconsin at Madison*, 2003.

J. Powell. Estimation of monotonic regression models under quantile restrictions. In W. Barnett, J. Powell, and G. Tauchen, editors, *Nonparametric and semiparametric methods in econometrics and statistics*, pages 357–384, 1991.

A. Prokhorov and P. Schmidt. Gmm redundancy results for general missing data problems. *Journal of Econometrics*, 151(1):47–55, 2009.

H. Qian and P. Schmidt. Improved instrumental variables and generalized method of moments estimators. *Journal of Econometrics*, 91(1):145–169, 1999.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25, 1982.

H. White. *Asymptotic theory for econometricians*. New York: Academic Press, New York, New York, 1984.

H. White. *Estimation, inference and specification analysis*. Cambridge University Press, 1996.

J. Wooldridge. Asymptotic properties of weighted m-estimators for standard stratified samples. *Econometric Theory*, 17(2):451–470, 2001.

J. Wooldridge. *Econometric analysis of cross section and panel data*. First edition, 2002a.

J. Wooldridge. Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1(2):117–139, 2002b.

J. Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301, 2007.