Essays on a Discontinuity Test of Endogeneity

by

Maria Carolina Nizarala Martinez Caetano

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philisophy

in

Economics

in the

Gradute Division
of the
University of California, Berkeley

Committee in charge:
Professor James Powell, Chair
Professor Michael Jansson
Professor David Brillinger

Spring 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Essays on a Discontinuity Test of Endogeneity

by

Maria Carolina Nizarala Martinez Caetano

Abstract

Essays on a Discontinuity Test of Endogeneity

by

Maria Carolina Nizarala Martinez Caetano

Doctor of Philisophy in Economics

University of California, Berkeley

Professor James Powell, Chair

This dissertation develops a test of endogeneity without the need of instrumental variables. The test ensues from the novel observation that the potentially endogenous variable X is often of a nature such that the distribution of the unobservable Q conditional on X and covariates Z is discontinuous in X at a known value in its range. This relationship arises, for example, when X is subject to corner solutions, default contracts, social norms or law imposed restrictions, and may be argued using both economic theory and empirical evidence. The idea relies in that if X has a continuous effect on the dependent variable Y, any discontinuity of Y that is not accounted by the discontinuities in the covariates Z is evidence that Q and Y are dependent conditional on Z, i.e. it is evidence of the endogeneity of X.

The first part of this dissertation develops the test inside of a linear model where X is censored. In this case the test converges under the null hypothesis of the exogeneity of X at the rate $\sqrt{n}$. The part includes the identification of the parameter which will be used as a basis for the test statistics, the construction of the test statistics and derivation of an asymptotic theory of its behavior, and finally a Monte Carlo study where the test is compared to the score test applied to the model, which can be understood as an endogeneity test. The Monte Carlo study uses real data on the effects of maternal smoking in birth weight, and the different versions of the discontinuity test present identical size and power as the score test under the assumptions for the optimality of the latter. When Z is endogenous, the score test as previously defined is no longer optimal, and the discontinuity test performs significantly better, with gains of up to 100% more rejections than the score test for certain levels of correlation.

The second part of this dissertation develops a theory of the discontinuity test the endogeneity of X in the structural function f. In this case, X need not be censored, and f need not be linear. The parameter which serves as the basis of the test can be identified non-parametrically, and consists of the aggregation of the discontinuities of the $\mathbb{E}(Y \mid X, Z)$ over a measure of Z. The work develops the test

statistic of the first part as one of the cases, but then generalizes the test for the cases when $\mathbb{E}(Y \mid X, Z)$ is nonparametric in X and separably linear in Z, and when $\mathbb{E}(Y \mid X, Z)$ is nonparametric, but Z has finite support. In these two cases, the test statistic is shown to converge at the univariate nonparametric rate $\sqrt{nh}$. This part also discusses an undersized test of the endogeneity of X when the support of its distribution is not continuous. The part ends with a discussion of the applicability of the test, with examples of situations where the test assumptions can be argued naturally, and showing how this can be done in the case of the estimation of the effects of maternal smoking in birth weight.

The third part of this dissertation is a study of endogeneity in the problem of the estimation of the effect of maternal smoking on birth weight and on the probability of low birth weight (LBW). It presents a discussion of the difficulties faced by the randomized trials and instrumental variable approaches in the area. Then, it applies the discontinuity test for a partially linear specification (linear in Z), where Z is chosen to be the same as in the most exhaustive study using the selection on observables assumption in the literature, Almond et al. (2005). The test finds strong evidence of endogeneity in the structural function relating amounts smoked and birth weight, and very weak evidence in the structural function relating amounts smoked and the probability of LBW.

To Greg

Contents

List of Figures

List of Tables

Acknowledgments

I would like to begin by acknowledging my advisor in the Economics department, professor James Powell, who helped immensely in the formulation of the discontinuity test since the very beginning when it was only an intuitive idea. Jim dedicated many hours to me, and to this problem in particular, and without him this idea would not have gone nearly as far. I am taking a lot of Jim's wisdom with me as our path together comes to an end. Jim is widely loved and respected in the profession. He is calm and infinitely practical. He knows so much about so many areas of econometrics. I wish one day to be like him. Jim, thank you for your advice, mentoring, patience and kindness. You are awesome!

Professor David Brillinger came to this project late in my years at Berkeley. He so kindly agreed be my advisor in the M.A. in Statistics program. I was surprised at the amount of time and effort he put into this work, and I couldn't ask for anyone better as my M.A. advisor. He was constructive in his criticism, very precise in his writing, and he taught me a lot about Statistics' terminology, style and standards. David, this project improved so much because I had your help. I am really grateful.

I would also like to thank professors David Card, Kenneth Chay, Bin Chen, Sandrine Dudoit, David Freedman, Bryan Graham, Michael Jansson, Patrick Kline, Dennis Nekipelov, Demian Pouzo, Paul Ruud and Nese Yildiz for extremely helpful comments in different versions of this dissertation. Professors Douglas Almond, Kenneth Chay and David Lee graciously provided the data set used in the application. My colleagues and friends Felipe Barros, Maria Salas-Bixler, Mario Capizzani, Matias Cattaneo, Andres Donangelo, Shari Ely, Constanca Esteves, Gee Hee Hong, Ram Rajagopal and Muzhe Yang provided support and contributed immensely to this dissertation by listening about it, reading versions, proof editing, or simply by being there. I am most grateful to my dear friend Jeff Greenbaum, who shared so closely most of my years in the Ph.D. in Economics program, who listened and worried with me, and to whom I entrusted every possible task while I was away from Berkeley. Jeff, you are a true friend. Thank you.

I want to take this opportunity to acknowledge professor Mauricio Bugarin. I met him when I took his undergraduate Microeconomics course at the University of Brasilia. He made Economics so interesting that I decided to change fields and pursue an M.A. in Economics following my undergrad in International Relations. He was also the one to suggest, years later, that I came to the U.S. for a Ph.D. when I was already in the doctorate program at EPGE-FGV. Querido Bugarin, you are ever so kind, you truly care for your students, and this is why you make a difference. I know for me you did.

Professor Elon Lages Lima is a mathematician at IMPA, and he was my biggest mentor. I worked with him during all the years I spent in Rio in the M.A. in Economics program. I am trying to summarize everything I learned from him, but I can't. He taught me about math, about teaching (his biggest passion), and about life. He

believed in my potential better than anyone else, and because of him I did too. If I am an econometrician now, it is largely because of the love of math that he instigated in me.

I would also like to acknowledge my mother, Martha Berbert, a medical librarian of incredible competence. She helped with the research of the medical literature. However, more than this, she inspired in me the courage to leave home so early and pursue my dreams. Mom, I owe you my victories! My father Amado Nizarala de Avila, an amazing OB-GYN by whose hands more than 6000 children were born, was the source of many a prior information about the effects of maternal smoking. He was also the person who instigated in me the curiosity, the love of study and the eclectic taste. Dad, we are so much alike!

My family cheered me all the way. My younger sisters Cecilia Nizarala Martinez, Victoria Martinez Nizarala, Leila da Silva de Avila Nizarala, and brother Felipe Berbert are a constant source of amusement. My stepdad Frederico Berbert, who is always the coolest, and my stepmother Adielma, who is so very strong, my brother-in-law Tiberio Caetano, who is so intelligent and so rational, and my sister-in-law Camila Caetano, who is so much fun, you mean a lot to me. I must make a difference to my father-in-law Reginaldo Caetano and mother-in-law Ione Caetano. You are the sweetest people I know, you are my family indeed, you participated in every little victory, you cheered for me and you worried for me, you were proud of me like parents ought, and this is how I feel about you: I am lucky that you are my people.

Finally, I must speak of my husband Gregorio Caetano, who is also an economist. He joined the Ph.D. in Economics program at Berkeley one year before me. He was my source of strength and happiness all these years, and we came out of this even more united. He is the happiest, most positive person I know. He is also good to everybody, but infinitely generous to me. He discussed every little detail of this dissertation with me, and nobody knows it as he does. He helped in all aspects of this work, but he was especially needed in the application part, where I was out of my element, and where he excels. Because of the interactions with him I see many promising ramifications of this work. Amor, I admire you, and I have confidence in you. I would follow you anywhere, and I am glad of the path we took. I love you forever.

Maria Carolina Caetano, May 2, 2010

Part I

Discontinuity tests for endogeneity of a censored regressor

Chapter 1

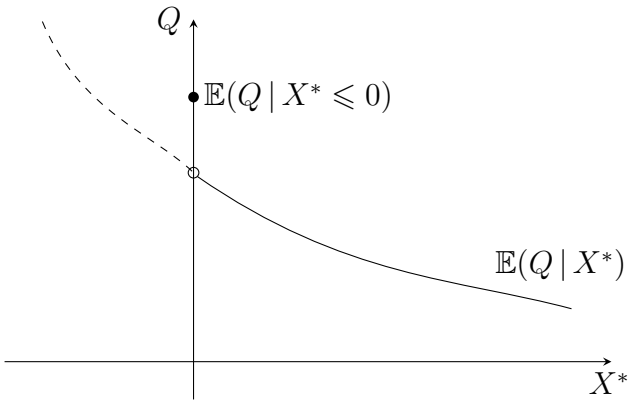
Introduction

Semi-parametric models of simultaneous equations with censored dependent variables were first treated by Heckman (1978), Amemiya (1979), and Lee (1979). Later, Newey (1987) considered models with endogenous explanatory variables, devising an efficient estimation method with instrumental variables. Smith and Blundell (1987) developed a test for exogeneity of a censored variable in a parametric model, and proposed an efficient estimator under these conditions. Vella (1993) used a generalized residuals approach (see Gourieroux et al. (1987)) in a very general model with censored (potentially endogenous) regressors, and Gaussian errors, from which he derived an endogeneity test and an estimator. Vella (1993) briefly considered the possibility of extending the testing results to non-Gaussian models by borrowing from the semi-parametric literature (Gallant and Nychka (1987)), and the diagnostic testing literature (Lee (1984) and Pagan and Vella (1989)). He suggested to capture the departures from normality by including powered up values of the generalized residuals in the moment condition. The properties of the resulting test are not discussed in Vella (1993)'s article, but its form appears to be similar to the score test, whose properties in the specific model treated in this part will be discussed in chapter 4.

This part suggests a different approach for testing the exogeneity of a censored regressor when this regressor appears in the structural equation in the censored form, as opposed to in the latent form. Such is the case of "corner solution" models. In a very simple setup, where the structural equation includes only a constant and the censored (potentially endogenous) regressor in the right hand side, the conditional expectation of the dependent variable given the regressor shouldn't be discontinuous at the point of the restriction. If such discontinuity is found, it could be evidence of the existence of an unobservable variable correlated to both the dependent variable and the latent form of the censored regressor.

In a more general context, it is the conditional expectation of the error term in the structural equation given the regressors that shouldn't be discontinuous at the censoring point. Moreover, there should exist no discontinuity at the censoring point

Figure 1.1: Discontinuity in Q generated by censoring



of the conditional expectation of the error term when it is interacted with any function of the controls. Section 2 presents a model in which this discontinuity is evidence of the endogeneity of the censored regressor.

In this context, this part introduces the discontinuity tests for endogeneity, which consistently estimate the expected value of the discontinuity at the censoring point of the conditional expectation of the error term interacted with a function of the controls, and test whether the true expected discontinuity is in fact zero. Any function of the controls can be chosen, provided it satisfies the necessary moment condition expressed in chapter 2.

The discontinuity tests are simple to implement, and don't rely on exclusion restrictions or distributional assumptions. The Monte Carlo simulations presented in chapter 4 provide evidence that, when the unobservables and the controls are correlated, the discontinuity tests are significantly more efficient than two other likely tests. When no such correlation exists, the discontinuity tests showed no loss of efficiency when compared with the same other likely tests.

Chapter 2

The model

The model consists of the following three equations:

$$Y_i = \beta X_i + Z_i^T \gamma + \delta Q_i + \varepsilon_i \tag{2.1}$$
$$X_i^* = Z_i^T \pi + Q_i \tag{2.2}$$
$$X_i = \max\{0, X_i^*\} \tag{2.3}$$

where Z_i is a $K \times 1$ vector, and all other are scalar variables. Endogeneity in this model is equivalent to $\delta \neq 0$. Provided $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$, and all other concerned expectations exist, it is possible to write:

$$\mathbb{E}(Y_i \,|\, X_i, Z_i) = (\beta + \delta) X_i + Z_i^T (\gamma - \pi\delta) + \delta \mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i) \mathbf{1}(X_i = 0). \tag{2.4}$$

From 2.4, for any nontrivial function $g : \mathbb{R}^K \to \mathbb{R}$, two equations follow:

1. $\mathbb{E}(g(Z_i)Y_i) = (\beta + \delta)\mathbb{E}(g(Z_i)X_i) + \mathbb{E}(g(Z_i)Z_i)^T(\gamma - \pi\delta) + \\ + \delta\mathbb{E}[g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)]$
2. $\mathbb{E}(g(Z_i)Y_i \,|\, X_i > 0) = (\beta + \delta)\mathbb{E}(g(Z_i)X_i \,|\, X_i > 0) + \mathbb{E}(g(Z_i)Z_i \,|\, X_i > 0)^T(\gamma - \pi\delta)$

which show that $\beta + \delta$, $\gamma - \pi\delta$ and $\delta\mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0))$ are identifiable, whereas δ alone is not unless further assumptions are made.

If g is such that

$$\mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)) \neq 0,$$

testing whether $\delta \neq 0$ is equivalent to testing whether

$$\delta\mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)) \neq 0.$$

In particular, if $P(X_i^* < 0) > 0$ and $g(Z_i) = 1$, the condition

$$\mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)) \neq 0$$

is satisfied, since $\mathbb{E}(\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)) = \mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0)P(X_i^* \leqslant 0) < 0$.

Chapter 3

Discontinuity tests of endogeneity

The discontinuity test using function g consists of the estimation of

$$\delta \mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)),$$

and subsequent testing of whether this term is in fact equal to zero. This is achieved in two steps. The first step uses only the observations for which $X_i > 0$, regressing Y_i on $W_i = (X_i, Z_i^T)^T$ so as to obtain $\hat{\psi} = (\widehat{\beta + \delta}, \widehat{(\gamma - \pi\delta)}^T)^T$. The test statistics is then calculated using only observations for which $X_i = 0$:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i)(Y_i - W_i^T\hat{\psi})\mathbf{1}(X_i = 0)$$

If $(Z_i^T, Q_i, \varepsilon_i)^T$ are i.i.d., and all necessary conditions of existence and invertibility of second moments are satisfied, it is simple to show that

$$\sqrt{n}(\hat{\theta} - \delta\mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0))) \xrightarrow{d} N\left(0, \sigma^2(p + \Delta) + \delta^2\omega^2\right) \quad (3.1)$$

where

$$\begin{aligned} p &= \mathbb{E}(g(Z_i)^2\mathbf{1}(X_i = 0), \\ \Delta &= \mathbb{E}(g(Z_i)W_i\mathbf{1}(X_i = 0))^T\mathbb{E}(W_iW_i^T\mathbf{1}(X_i > 0))^{-1}\mathbb{E}(g(Z_i)W_i\mathbf{1}(X_i = 0))^T \\ \omega^2 &= \mathbb{V}ar(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)). \end{aligned}$$

A formal statement and proof of this result is available in the appendix section A.1. The variance of the estimator under $H_0 : \delta = 0$ can be consistently estimated

by $\hat{\sigma}^2(\hat{p}+\hat{\Delta})$, where

$$\hat{\sigma}^2 = \frac{1}{n_A}\sum_{i\in A}(Y_i - W_i^T\hat{\psi})^2, \quad A = \{i = 1, ..., n; X_i > 0\},$$
$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\mathbf{1}(X_i = 0),$$
$$\hat{\Delta} = \left(\frac{1}{n}\sum_{i=1}^{n} g(Z_i)W_i\mathbf{1}(X_i{=}0)\right)^T\left(\frac{1}{n}\sum_{i=1}^{n} W_iW_i^T\mathbf{1}(X_i{>}0)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(Z_i)W_i\mathbf{1}(X_i{=}0)\right).$$

In matrix notation, let M be the $n\times 1$ vector whose i^{th} entry is $M_i = g(Z_i)\mathbf{1}(X_i = 0)$, W and $\tilde{\mathrm{W}}$ be the $n \times (K+1)$ matrices whose rows are respectively W_i^T and $W_i^T\mathbf{1}(X_i > 0)$. The discontinuity test statistic can be expressed as

$$\frac{\mathrm{Y}^T(I - \tilde{\mathrm{W}}(\tilde{\mathrm{W}}^T\mathrm{W})^{-1}\mathrm{W}^T)\mathrm{M}(\mathrm{M}^T(I + \mathrm{W}(\tilde{\mathrm{W}}^T\mathrm{W})^{-1}\mathrm{W}^T)\mathrm{M})\mathrm{M}^T(I - \mathrm{W}(\tilde{\mathrm{W}}^T\mathrm{W})^{-1}\tilde{\mathrm{W}}^T)\mathrm{Y}}{s},$$

where

$$s^2 := \frac{\mathrm{Y}^T(I - \tilde{\mathrm{W}}(\tilde{\mathrm{W}}^T\tilde{\mathrm{W}})^{-1}\tilde{\mathrm{W}}^T)\mathrm{Y}}{n - (K+1)}.$$

This value should be compared to the critical values of the distribution $\mathcal{F}_{1,n-(K+1)}$. This test statistic is equivalent to an IV test of whether the coefficient of a regressor η_i in a regression of Y_i on W_i and η_i is zero, using $W_i\mathbf{1}(X_i > 0)$ and $M_i = g(Z_i)\mathbf{1}(X_i = 0)$ as instruments. The choice of η_i is arbitrary, as long as $\mathrm{M}^T(I - \mathrm{W}(\tilde{\mathrm{W}}^T\mathrm{W})^{-1}\tilde{\mathrm{W}}^T)\eta$ is invertible for the available sample.

Chapter 4

Monte Carlo

The first design presented here intends to observe the behavior of the version of the discontinuity test with $g(Z_i) = 1$ in a real dataset. The discontinuity test using this function measures the average discontinuity of equation (2.4) at $X_i = 0$ after controlling for Z_i. The comparison benchmark is the score test.

The dataset chosen is the National Maternal and Infant Health Survey, 1988. The data (5461 observations) is assumed to satisfy equations (2.1) to (2.3), with Y_i being birth weight, X_i being cigarettes smoked per day during pregnancy (censored below zero), and Z_i a vector including a constant and 36 covariates. The covariates include mother's age, education, marital status, race, foreign status, father's age, education and race, number of prenatal visits, date of first prenatal care, first born dummy, number of previous live births, number of previous births where newborn died, interval since last live birth, alcohol use, residence in metropolitan area, delivery with doctor, delivery in hospital, and month of birth. The specification is based in Almond et al. (2005). The variates ε_i and Q_i are assumed to be independent and normally distributed with respective variances σ^2_ε and σ^2_q.

The model was then calibrated so that the parameters used in the design preserved the moments of the population. The design parameters π and σ^2_q where calculated using a tobit regression of X_i on Z_i. This also yielded an estimated $Q_i = X_i - Z_i^T\pi$ for all observations such that $X_i > 0$. To each level of correlation (ρ) between the unobservables in equation (2.1), $\delta Q_i + \varepsilon_i$, and the unobservables in equation (2.2), Q_i, corresponds a different value of $\delta = \delta(\rho)$. $\beta(\rho)$, $\gamma(\rho)$ and $\sigma^2_\varepsilon(\rho)$ were then calculated through an OLS regression of $Y_i - \delta(\rho)Q_i$ on X_i and Z_i, using only observations for which $X_i > 0$.

Let

$$M_i = (X_i - Z_i^T\hat{\pi})1(X_i > 0) - \hat{\sigma}_q\lambda\left(\frac{-Z_i^T\hat{\pi}}{\hat{\sigma}_q}\right)\mathbf{1}(X_i = 0),$$

the score test in this design is equivalent to a t-test of whether the coefficient of the term M_i in an OLS regression of Y_i on X_i, Z_i and M_i is equal to zero. $\hat{\pi}$ and $\hat{\sigma}_q$ are obtained through a tobit regression of X_i on Z_i, and λ is the inverse Mills' Ratio.

For each sample size of 100, 200, 400 and 1000 observations, the values Z_i were chosen through resampling from the the original population. The values of ε_i and Q_i were taken from a joint normal distribution with variances given by the calibration process above, and no correlation. There were 10,000 repetitions. For each observation and repetition, Z_i, ε_i and Q_i generate all of the other variables through the design model. For small samples (notably 100 and 200 observations), it was necessary to correct for the degrees of freedom in each test, since both tests were oversized when using the critical values of the respective asymptotic distributions. The tables below show the proportion of rejections in the t-test versions of the discontinuity and the score tests for each correlation value ρ at a 5% significance level.

Table 4.1: **Rejections when Z_i and Q_i are independent**

Exp.1: 100 obs.		
ρ	*Disc.*	*Score*
0.00	4.9	5.0
0.10	5.1	5.1
0.25	6.3	6.8
0.50	12.6	13.4
0.75	36.4	35.8
0.90	79.0	74.4

Exp.2: 200 obs.		
ρ	*Disc.*	*Score*
0.00	5.0	5.0
0.10	5.5	5.6
0.25	9.5	9.7
0.50	29.4	29.8
0.75	78.7	78.0
0.90	99.5	99.3

Exp.3: 400 obs.		
ρ	*Disc.*	*Score*
0.00	5.1	5.0
0.10	6.7	6.8
0.25	16.6	16.8
0.50	59.1	58.5
0.75	98.7	98.6
0.90	100.0	100.0

Exp.4: 1000 obs.		
ρ	*Disc.*	*Score*
0.00	5.3	5.3
0.10	10.2	10.1
0.25	38.6	38.5
0.50	94.9	94.8
0.75	100.0	100.0
0.90	100.0	100.0

Proportion of rejections of the null hypothesis (H_0: $\delta = 0$) at the 5% significance level out of 10,000 repetitions.

The results show that the discontinuity test has the right (corrected) size. The power of the two tests is virtually identical at all correlation levels, even for small sample sizes.

If Q_i is correlated with Z_i, the score test shown above is no longer the true score test that would apply to the model. The appropriate score test would require that the form of the correlation between Q_i and Z_i be known. If such correlation is ignored, the version of the score test shown above may be significantly less efficient than the discontinuity test, as will be shown in the experiments bellow.

The Monte Carlo design that follows uses artificially generated data. The matrix Z, whose rows are the Z_i^T, has four columns, one of which is a constant, and 100 observations. Four experiments were implemented. In the first two (Experiments 5 and 6), the ε_i were simulated independent normally distributed. In the last two (Experiments 7 and 8), the ε_i were simulated independent beta distributed, left skewed, but with the same mean and variance as in the normal case. As to the correlation between Q_i and Z_i, the simulations in Experiments 5 and 7 have no correlation between Q_i and Z_i, while in the simulations in Experiments 6 and 8, Q_i and Z_i are correlated: the Q_i are assumed to be an arbitrary nonlinear function of some of the columns of Z and a random (normally distributed) variable. The correlations between Q_i and each of the three columns of Z_i are -0.04, -0.002, and -0.0002. Four tests were applied to the resulting datasets.

The first test is the discontinuity test with

$$g(Z_i) = 1,$$

and it will be identified in the tables as Disc1. The second test is a split-sample test with function

$$g(Z_i) = \lambda\left(\frac{-Z_i^T\hat{\pi}}{\hat{\sigma}_q}\right),$$

which will be identified as Disc2. The third test is the score test used in the real data simulations above, that assumes ε_i and Q_i to be jointly normal, and Q_i and Z_i to be uncorrelated. It will be identified as Score. Finally, the fourth test consists of a t-test of whether the coefficient of the dummy variable $\mathbf{1}(X_i = 0)$ in a regression of Y_i on X_i, Z_i and $\mathbf{1}(X_i = 0)$ is equal to zero. This test will be identified as the Dummy test. As with the simulation with real data (albeit less so), it was necessary to correct for the degrees of freedom in each test, since all were oversized when using the critical values of the respective asymptotic distributions.

The tables below show the proportion (out of 10,000 repetitions) of rejections at the 5% significance level. Table 5 shows the results of the first simulation, where Q_i and Z_i are independent, and ε_i is normally distributed. Table 6 shows the results of the second simulation, where Q_i and Z_i are correlated, and ε_i is normally distributed.

Table 5 shows that the four tests performed very similarly when Q_i and Z_i are independent. There are some gains of efficiency for the discontinuity tests at the correlation levels $\rho = 0.5, 0.75$, and 0.9. Larger gains of efficiency are obtained when Q_i and Z_i are correlated. For this example, Disc1 test performed marginally better

Table 4.2: **Artificial experiment, ε_i has a normal distribution**

	Exp.5: Q_i and Z_i uncorrelated				**Exp.6: Q_i and Z_i correlated**			
ρ	*Disc*1	*Disc*2	*Score*	*Dummy*	*Disc*1	*Disc*2	*Score*	*Dummy*
0.00	5.0	4.9	4.9	5.0	4.9	4.9	5.1	5.1
0.10	6.0	5.8	5.7	5.8	5.6	5.6	5.6	5.6
0.25	10.4	10.4	9.6	10.4	10.4	10.9	7.8	8.7
0.50	31.8	31.8	27.2	31.3	32.0	34.8	17.9	23.0
0.75	81.2	80.9	69.5	78.3	80.9	84.1	45.7	60.1
0.90	99.7	99.7	94.8	99.3	99.7	99.7	77.6	92.7
0.99	100.0	100.0	99.0	100.0	100.0	100.0	92.6	99.8

than Disc2 test. At the levels $\rho = 0.5, 0.75$, and 0.9, the dummy test exhibited a significant loss of efficiency compared to the discontinuity tests, and the score test rejected even less.

When ε_i is simulated as having a skewed distribution, the tests' comparative behavior remains the same. Table 7 shows the results of the third simulation, where Q_i and Z_i are independent, and ε_i has a skewed distribution. Table 8 shows the results of the fourth simulation, where Q_i and Z_i are correlated, and ε_i has a skewed distribution.

Table 4.3: **Artificial experiment, ε_i has a skewed distribution**

	Exp.7: Q_i and Z_i uncorrelated				**Exp.8: Q_i and Z_i correlated**			
ρ	*Disc*1	*Disc*2	*Score*	*Dummy*	*Disc*1	*Disc*2	*Score*	*Dummy*
0.00	5.1	5.2	5.2	5.3	5.1	5.0	5.2	5.1
0.10	5.7	5.6	5.9	5.8	5.4	5.5	5.6	5.6
0.25	9.5	9.7	9.8	10.6	9.5	10.2	7.8	8.9
0.50	31.0	30.7	27.1	31.3	30.8	33.7	17.8	22.5
0.75	81.9	81.5	69.2	78.2	81.9	84.7	45.2	59.5
0.90	99.8	99.8	94.8	99.4	99.7	99.8	77.2	92.9
0.99	100.0	100.0	99.0	100.0	100.0	100.0	92.5	99.9

Tables 4.2 and 4.3: Numbers shown are the proportion (over 10,000 repetitions) of rejections of the null hypothesis of no endogeneity at the 5% significance level, for actual correlation between the unobservables in the structural and in the latent equations equal to ρ. The ε_i have a skewed distribution. Specification with 3 covariates and one constant. Sample sizes of 100 obs.

The results are robust to different draws of the matrix Z. Changes in Z's generating process, yielded slight modifications in the relative efficiency of the Disc1 and Disc2 tests, with one or another performing better depending on the model chosen. However, the relative positions of both discontinuity tests and the other tests remained the same. The same happened to different choices of the model of correlation between Q and Z.

Part II

A discontinuity test of endogeneity

Chapter 5

Introduction

This part develops a test of the problem of endogeneity in a structural function. When a random variable (r.v.) Y has an expressed relationship to the r.v.'s X, Z, Q and ε, the structural function is defined as the function f such that

$$Y = f(X, Z, Q, \varepsilon). \tag{5.1}$$

Let X be a scalar observable r.v., Z be a vector of observable r.v.'s, Q be a vector of unobservable random variables that present some form of probabilistic dependence with X, and ε a vector of unobservable random variables which are independent of X. X is said to be "exogenous" in f if f is constant in Q, that is, when $f(X, Z, Q, \varepsilon) = f(X, Z, q, \varepsilon)$ everywhere but at a zero probability set, for any particular fixed value that Q can assume, denoted q. Conversely, X is said endogenous if such condition is not satisfied.

The problem of endogeneity in the structural function is fundamental for the identification of interesting properties of f through the conditional expectation of Y given the observable variables X and Z. For example, as is commonly the case, one may be interested in the expected derivative of f with respect to its first argument. If it is possible to interchange $\frac{\partial}{\partial X}$ and $\mathbb{E}$, then

$$\frac{\partial}{\partial X}\mathbb{E}(Y \mid X, Z) = \mathbb{E}\left(f_x(X, Z, Q, \varepsilon) + f_q(X, Z, Q, \varepsilon)\, \frac{dQ}{dX} \,\middle|\, X, Z \right),$$

where f_x and f_q are the derivatives with respect to the arguments X and Q respectively, then if X is exogenous, $f_q(X, Z, Q, \varepsilon) = 0$, and therefore $\mathbb{E}(f_1(X, Z, Q, \varepsilon) \mid X, Z)$ is identified.

Tests of endogeneity (which are in fact tests of the null hypothesis of the exogeneity of X) often assume that f has a specific structure. In these cases, it is impossible to disentangle whether a rejection was caused by endogeneity or by the misspecification of the structure of f. However, the two problems have entirely different solutions. Misspecification is solved by searching for the correct specification, while endogeneity

requires the use of instrumental variables when available, or the adoption of methods that account for endogeneity, such as "differences in differences," "correlated random effects," etc. when the data and the problem allow, or finally searching datasets where a wider array of covariates is observed. A test of endogeneity which does not impose a structure on f is immune to the problem of misspecification, and therefore a rejection in such a test can be inferred as evidence of endogeneity.

Nonparametric tests of endogeneity are not abundant in the literature. This is in part due to the recency of the research on nonparametric instrumental variable (IV) estimators. Blundell and Powell (2003) and Hall and Horowitz (2005) discuss the difficulties involved in such undertaking, due to the fact that the identifying condition is an "ill-posed inverse problem." Nonparametric IV estimators of the structural function have been proposed in Darolles et al. (2003), Blundell et al. (2007), Newey and Powell (2003), and Hall and Horowitz (2005). The available tests of endogeneity suppose either that the potentially omitted variables are observed (see for example Fan and Li (1996), Chen and Fan (1999) and Li and Racine (2007)), or that an instrumental variable exists and is observed (see Blundell and Horowitz (2007) and Horowitz (2009)). In both cases, f is identifiable and can be consistently estimated, and the test is useful in the decision of which estimation strategy to pursue. This is no small concern in nonparametric estimation, because the rates of convergence of the estimators decrease considerably if irrelevant covariates are included, and even more if an instrumental variable approach is used where it is not necessary. The potential efficiency losses are therefore much more substantial than in the parametric cases.

The test presented in this part does not require that the omitted variables be observable, nor that an instrumental variable exist, and to the author's knowledge, it is the first nonparametric test of endogeneity in the structural function where neither of these two conditions is necessary. Since most omitted variables are so because of being unobserved and since good instrumental variables are often not readily available, a test of endogeneity with no such requirements is of considerable interest. Its usefulness is in alerting about a problem of endogeneity before any measure to solve it is researched. Alternatively, the test can be used to validate a selection on observables approach when instrumental variables are not available. In the cases where selection on observables is acceptable, the test can be applied further as an omitted variable test, to aid the exclusion of further covariates, which greatly improves efficiency in nonparametric estimation.

The fundamental maintained assumption in this approach is that f is continuous in X. In that case, if X is exogenous, then $\mathbb{E}(Y \mid X, Z)$ has to be continuous in X. Hence, a test of the null hypothesis

$$H_0\colon\ X \text{ is exogenous,}$$

versus the alternative hypothesis

$$H_1\colon\ X \text{ is endogenous,}$$

can be built by estimating the discontinuity of $\mathbb{E}(Y \mid X, Z)$ with respect to X at a given point $X = x_0$.

The test has power against H_1 when the distribution of Q conditional on X and Z is discontinuous in X at a known point $X = x_0$. In this case, if X is endogenous, then f varies with Q, which in general implies that $\mathbb{E}(Y \mid X, Z)$ will be discontinuous in X at $X = x_0$. The cases where the distribution of Q conditional on X and Z is discontinuous in X at a known point $X = x_0$ define the spectrum of applications where this test can be used to validate a selection on observables approach.

In principle, the test could consist of the estimation of the discontinuity of $\mathbb{E}(Y \mid X = x, Z = z)$ in x at $x = x_0$ for a given z, but this would seem to require the nonparametric estimation of the limit of the multivariate function $\mathbb{E}(Y \mid X = x, Z = z)$ when x approaches x_0. The rate of convergence of such regressions is typically very slow, and hence the test would have little power. A much higher rate of convergence can be achieved through working with a characteristic aggregation over a measure of Z, i.e. by estimating, for example, the average discontinuity, or the correlation between the discontinuities and a function of Z, etc. For a wide variety of such tests, this part shows that the rate of convergence is the same as that of a univariate nonparametric regression (sections 7.2 and 7.3 in chapter 7). Therefore, the aggregation allows for controlling the influence of the observable covariates without loss of power due to slower rates of convergence. This is a property observed in the literature of partial means (see Newey (1994)), or marginal integration (see Linton and Nielsen (1995)).

The discontinuities are estimated in a similar fashion to what is done in the regression discontinuity literature (see Imbens and Lemieux (2008)), by estimating the one sided limits of the conditional expectation at a point. This entails nonparametric estimation at the boundary, and to minimize problems due to boundary bias, this part will make extensive use of local polynomial estimators. The estimators of the limits of G are very similar to those already seen in the regression discontinuity literature (see Porter (2003)), but this part proposes a new estimator of the variance of the boundary local polynomial estimator, which allows the density of X to be different at the right and left sides of the threshold x_0, and also uses boundary estimators for all the components of the variance. This approach to the estimation of the variance is more adequate to the kinds of situations where this test can be applied, but can also be useful in the context of the regression discontinuity design, and particularly useful in the newly developed regression kink design (Card et al. (2009)).

This part develops the discontinuity test of endogeneity under three testable assumptions about the $\mathbb{E}(Y \mid X, Z)$: when it is linear in X and Z, when it is nonparametric in X and additively linear in Z, and finally when no parametric assumptions are made about $\mathbb{E}(Y \mid X, Z)$, but there exists a finite subset of values of Z which has positive probability. The more restrictive cases yield naturally more precise test statistics, which converge at the rates $\sqrt{n}$ in the linear case and $\sqrt{nh}$ in both the partially linear and nonparametric cases. The test is not hard to implement. The linear case is trivial, and for the partially linear and fully nonparametric cases, all

that is required for the estimation of the test statistic and its variance is the computation of some local polynomial regressions at $X = x_0$ and some sample averages. The discretion requirements are only the choice of bandwidth, kernel type and the degree of the polynomial.

Chapter 8 discusses a test of endogeneity when the support of the distribution of X is not continuous. In this case, the comparison between $\mathbb{E}(Y \mid X = x_0, Z)$ and $\lim_{x \to x_0} \mathbb{E}(Y \mid X = x, Z)$ is irrelevant for uncovering the endogeneity of X. Instead, this section assumes that f has bounded variation, and builds an undersized test which detects endogeneity when Q causes $\mathbb{E}(Y \mid X = x, Z)$ to vary with x beyond what f is prescribed to vary with x.

Finally (in chapter 11), the partially linear version of the test is applied to the problem of the estimation of the effects of maternal smoking on birth weight, which is an example where the validity of the assumptions of the test can be argued. This is a problem where experiments cannot be performed for ethical reasons, and where a series of selection on observables approaches were attempted. Almond et al. (2005) is to the author's knowledge the most exhaustive of these studies, and in section 11 the discontinuity test is applied to the most complete specification in that paper for two outcome variables: birth weight, and probability of birth weight below 2500g. The test finds strong evidence of endogeneity in the birth weight equation, and very weak evidence in the probability of birth weight below 2500g equation.

The rest of this chapter has two sections. Section 5.1 presents the idea of the discontinuity test in a very simple linear example. Section 5.2 offers a detailed preview of all the parts of the paper.

5.1 A simple example

This section develops a simple model, with excessively restrictive conditions, with the purpose of clarifying the concepts and results that will be seen in the later chapters of this part. Consider the following model:

$$Y = \beta_X X + Z^T \beta_Z + \delta Q + \varepsilon \tag{5.2}$$

where ε is independent of X and Z and Q. If Q and X are correlated conditional on Z, then X is exogenous in (5.2) if $\delta = 0$, and endogenous if $\delta \neq 0$.

The structural function (5.2) is obviously continuous in X. Observe that

$$\mathbb{E}(Y \mid X, Z) = \beta_X X + Z^T \beta_Z + \delta \mathbb{E}(Q \mid X, Z).$$

Hence, if X is exogenous, $\delta = 0$, and $\mathbb{E}(Y \mid X, Z)$ is continuous in X. Define

$$\begin{aligned} \Delta(Z) &= \mathbb{E}(Y \mid X = x_0, Z) - \lim_{x \to x_0} \mathbb{E}(Y \mid X = x, Z) \\ &= \delta \left(\mathbb{E}(Q \mid X = x_0, Z) - \lim_{x \to x_0} \mathbb{E}(Q \mid X = x, Z) \right) \end{aligned}$$

for an x_0 in the support of the distribution of the X, supposing that these limits are well defined. Thus if X is exogenous, $\Delta(Z) = 0$, for all Z. Though this result is almost obvious due to the assumed linearity, the fundamental feature of (5.2) is continuity in X. Define

$$\theta = \mathbb{E}(\Delta(Z) \,|\, X = x_0), \tag{5.3}$$

then X exogenous implies $\theta = 0$, and the same would be true for any function G of the $\Delta(Z)$ for which $G(0) = 0$.

The following assumption, though extremely restrictive, explicits from where the power of the test derives.

Assumption 5.1.

1. $\mathbb{E}(Q \,|\, X, Z) = \alpha_X X + Z^T \alpha_Z + \alpha_Q \, g(Z)\mathbf{1}(X = x_0)$
2. $\alpha_Q \neq 0$
3. $E(g(Z)\mathbf{1}(X = x_0)) \neq 0$

Assumption 5.1 implies that $\theta = \delta \, \alpha_Q \, \mathbb{E}(g(Z) \,|\, X = x_0)$, and therefore if X is endogenous, $\delta \neq 0$, and hence $\theta \neq 0$. The fundamental fact implied by assumption 5.1 is that the distribution of Q conditional on X and Z is discontinuous in X at $X = x_0$, the linearity assumed in item (1) only helps to make the points clearer. The conclusion is that in this case θ is an appropriate parameter on which to base a test statistic, because not only it behaves as expected under the null hypothesis H_0 (that is, $\theta = 0$ when X is exogenous), but also it will yield non-trivial power under the alternative hypothesis H_1 (see theorem 7.3).

In order to build the test, observe that for $X \neq x_0$,

$$\mathbb{E}(Y \,|\, X, Z) = (\alpha_X + \beta_X)X + Z^T(\alpha_Z + \beta_Z). \tag{5.4}$$

Hence, from equation (5.3),

$$\begin{aligned}\theta &= \mathbb{E}(Y \,|\, X = x_0) - \mathbb{E}\left(\lim_{x \to x_0} \mathbb{E}(Y \,|\, X = x, Z) \,|\, X = x_0\right), \\ &= \mathbb{E}(Y \,|\, X = x_0) - \left[(\alpha_X + \beta_X)x_0 + \mathbb{E}(Z \,|\, X = x_0)^T(\alpha_Z + \beta_Z)\right].\end{aligned}$$

Suppose the data is composed of n random draws (Y_i, X_i, Z_i) of the variates (Y, X, Z). Equation (5.4) guarantees that $\gamma_X := \alpha_X + \beta_X$ and $\gamma_Z := \alpha_Z + \beta_Z$ are identified, and can be estimated reasonably by an OLS regression of the $Y_i\mathbf{1}(X_i \neq x_0)$ on $X_i\mathbf{1}(X_i \neq x_0)$ and $Z_i\mathbf{1}(X_i \neq x_0)$. $\mathbb{E}(Y \,|\, X = x_0)$ and $\mathbb{E}(Z \,|\, X = x_0)$ can be estimated using their empirical counterparts. The discontinuity test statistic is now

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = x_0)\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - \hat{\gamma}_X x_0 - Z_i^T \hat{\gamma}_Z\right] \mathbf{1}(X_i = x_0).$$

Assumption 5.2.

1. (Y_i, X_i, Z_i), $i = 1, \ldots, n$ *are i.i.d.*
2. $0 < \mathbb{P}(X_i = x_0) < 1$
3. $\mathbb{V}ar(\varepsilon_i \,|\, X_i, Z_i) = \sigma_\varepsilon^2 < \infty$
4. *Let* $\eta_i := Q - \mathbb{E}(Q_i \,|\, X_i, Z_i)$, *then* $\mathbb{V}ar(\eta_i \,|\, X_i, Z_i) = \sigma_\eta^2 < \infty$, $i = 1, \ldots, n$

   It is direct to decompose $\hat{\theta} - \theta$ *in matrix form as*

   $$\hat{\theta} - \theta = (\iota^T \iota)^{-1}(I + W(W^T DW)^{-1} W^T D)(\boldsymbol{\varepsilon} + \delta \boldsymbol{\eta}),$$

   where $\iota = (\mathbf{1}(X_1 = x_0), \ldots, \mathbf{1}(X_n = x_0))^T$, I *is the identity* $n \times n$ *matrix,* W *is a matrix with rows equal to* (X_i, Z_i^T), $D = Diag\{\mathbf{1}(X_1 \neq x_0), \ldots, \mathbf{1}(X_n \neq x_0)\}$ *and* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{\eta}$ *are the column vectors with rows* ε_i *and* η_i *respectively, and supposing that* $W^T DW$ *is invertible. Assume further that*
5. $\mathbb{E}\left(\frac{W^T DW}{n}\right)$ *is invertible*

Theorem 5.1. *Given the model and assumptions 5.1 and 5.2, then*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{P}(X = x_0)^{-1}(\sigma_\varepsilon^2 + \delta^2 \sigma_\eta^2)(1 + v)\right),$$

where

$$v = \begin{bmatrix} x_0 \, \mathbb{E}(Z \,|\, X = x_0) \end{bmatrix} \begin{bmatrix} \mathbb{E}(X^2 \mathbf{1}(X \neq x_0)) & \mathbb{E}(XZ^T \mathbf{1}(X \neq x_0)) \\ \mathbb{E}(ZX \mathbf{1}(X \neq x_0)) & \mathbb{E}(ZZ^T \mathbf{1}(X \neq x_0)) \end{bmatrix}^{-1} \begin{bmatrix} x_0 \\ \mathbb{E}(Z \,|\, X = x_0) \end{bmatrix}.$$

This result is very similar to theorem 7.1. Under H_0,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\varepsilon^2 \mathbb{P}(X = x_0)^{-1}(1 + v)\right),$$

and the method of moments estimator of σ_ε^2 is given by

$$\hat{\sigma}_\varepsilon^2 = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \neq x_0)\right)^{-1} \frac{1}{n}\sum_{i=1}^n \left[Y_i - \hat{\gamma}_X x_0 - Z_i^T \hat{\gamma}_Z\right]^2 \mathbf{1}(X_i \neq x_0).$$

The other elements can be estimated by substituting the population by the sample counterpart.

5.2 Overview of Part II

This section presents a detailed overview of what is covered in each chapter of this part. Since the objective is to develop a test of endogeneity, the work begins by defining a parameter that will be the basis of the test statistics, then it builds the actual test statistics using that parameter as reference. For different test statistics it provides different asymptotic results and power considerations. Then it goes on to discuss the applicability of the test, using many examples and also showing how to argue that the test is applicable in a given situation. Finally, it applies the test to the case of the effects of maternal smoking.

Chapter 6 covers the development of the parameter that is the basis of the discontinuity test. Let the discontinuities of $\mathbb{E}(Y \mid X = x, Z)$ in x at x_0 be denoted $\Delta(Z)$, $\theta = \int G(\Delta(Z), Z) d\nu(Z)$ is the aggregation of known functions G of such discontinuities over a known measure ν over the range of the Z. In this chapter it is shown (theorem 6.1) that X exogenous implies that $\theta = 0$ under a set of assumptions discussed there (assumption 6.1), of which the most important is that $\mathbb{E}(Y \mid X = x, Z, Q)$ is continuous in X for all Z and Q. The power of the test derives from $\theta \neq 0$ when X is endogenous. The same chapter proposes a condition (assumption 6.2) that guarantees non-trivial power in general.

Chapter 6 also discusses that the function $G(\Delta(Z), Z) = \Delta(Z)g(Z)$ and the measure $\nu(Z) = F(Z \mid X = x_0)$ are desirable, because the resulting θ is simpler to estimate. In that case, $\theta = \mathbb{E}(Y g(Z) \mid X = x_0) - \mathbb{E}(\lim_{x \to x_0} \mathbb{E}(Y \mid X = x, Z) g(Z) \mid X = x_0)$. The first term can be estimated via a simple sample average, and the second can be estimated by a sample average of the plugin estimators of the $\lim_{x \to x_0} \mathbb{E}(Y \mid X = x, Z)$ interacted with $g(Z)$.

Chapter 7 develops the discontinuity test statistic from the parameter discussed in chapter 6. It begins by proposing the empirical equivalent of θ, $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \big[Y_i - b_n(x, Z_i) \big] \mathbf{1}(X_i = x_0)$, where $b_{n(x,Z_i)}$ is an estimator of $\lim_{x \to x_0} \mathbb{E}(Y \mid X = x, Z)$. The form of b_n, as well as the empirical properties of $\hat{\theta}$ depend on the assumptions over $\mathbb{E}(Y \mid X, Z)$. This part explores three possibilities. The first (section 7.1) assumes that $\mathbb{E}(Y \mid X = x, Z)$ is linear when $x \neq x_0$. This is a testable assumption, and if satisfied, b_n can be chosen as the result of a simple OLS regression, as was shown in the simple example in section 5.1.

The second possibility (section 7.2) assumes that $\mathbb{E}(Y \mid X = x, Z) = \tau(x) + Z^T \gamma$ whenever $x \neq x_0$. This is testable assumption, and b_n in this case depends on the estimation of both γ and $\lim_{x \to x_0} \tau(x)$. This is done restricting the sample for the observations such that $X \neq x_0$ and estimating γ in the way it is traditionally done in the partially linear models literature (see Robinson (1988)). The nonparametric term $\lim_{x \to x_0} \tau(x)$ is then estimated by a local polynomial regression of the $Y_i - Z_i^T \hat{\gamma}$ on X_i at x_0. The resulting $\hat{\theta}$ is shown converge at the rate $\sqrt{nh}$ to a normally distributed random variable, and the asymptotic variance is the same as that of a local polynomial

regression of $Y_i - Z_i^T\gamma$ on X_i at x_0, which can be found for example in Porter (2003).

The third possibility (section 7.3) allows the $\mathbb{E}(Y \mid X, Z)$ to be arbitrary, though some differentiability conditions are necessary given the estimators used. The fundamental condition in this case is that the distribution of the Z have finite support. $\hat{\theta}$ is then the average of the $\hat{\theta}_m$, which are the discontinuity test statistic if the sample is restricted to the observations such that $Z_i = z^m$. The $\hat{\theta}_m$ are weighted by the empirical probability that $Z_i = z^m$. $\hat{\theta}$ converges at the rate $\sqrt{nh}$ to a normally distributed random variable with variance equivalent to the average of local polynomial regressions of $(Y_i - Z_i^T\gamma)\mathbf{1}(Z_i = z^m)$ on X_i at x_0, weighted by the probability that $Z_i = z^m$.

Both the identification and estimation strategies used in sections 6 and 7 depend on assumption 6.1 (1), which requires that the distribution of X have a continuous support, or that the support have at least one interval subset containing x_0. When this assumption fails, it is still possible to test endogeneity, though the approach has to change. Chapter 8 focuses on this problem, and supposes that when X varies, the variation of f is bounded, and the bounds are known. From this condition it is possible to define a parameter θ which is valued between the same bounds of variation of f whenever X is exogenous. The test built form such parameter is naturally undersized, but it has non-trivial power against H_1 when the variation caused by the endogeneity of X pushes θ outside of the established bounds.

Finally, chapter 9 discusses teh applicability of the discontinuity test, more specifically when the condition that $dF(Q \mid X, Z)$ is discontinuous in X at $X = x_0$. The main point of the chapter is that example that satisfy this condition are usually found when x_0 is a mass point in the data distribution.

Chapter 6

Identification

This chapter is concerned with the parameter θ on which the discontinuity test statistic is based. The chapter begins by presenting the formal definition of the endogeneity of X. Then in assumption 6.1, it presents conditions that guarantee that if X is exogenous, then $\mathbb{E}(Y \mid X = x, Z)$ is continuous. The fundamental requirement among them is that f is continuous in X (assumption 6.1 (2)). However, the discontinuity test will not be based on $\mathbb{E}(Y \mid X = x, Z)$ because of power concerns (see remark 6.1 below), as discussed in chapter 5, so the parameter θ is introduced as more desirable. The main identification theorem can then be considered (theorem 1). It consists of that if X is exogenous, then $\theta = 0$. The theorem is followed by assumption 6.2, which requires, among other technical details, that the distribution of Q conditional on $X = x$ and Z be discontinuous in x at $x = x_0$. A consequence of this assumption is that in general if X is endogenous, then $\mathbb{E}(Y \mid X = x, Z)$ is discontinuous in x at $x = x_0$. Result 1 then states that this in general translates into $\theta \neq 0$ when X is endogenous. The parameter θ has therefore the property that if X is exogenous, then $\theta = 0$, and if X is endogenous, then in general $\theta \neq 0$.

Definition 1. *Let Y, X, Z, Q be defined as in chapter 5. If Q and X are dependent conditional on Z, then X is exogenous in f if*

$$\mathbb{P}\left\{\mathbb{E}(Y \mid X, Z, Q) = \mathbb{E}(Y \mid X, Z)\right\} = 1.$$

Otherwise, X is endogenous.

Assumption 6.1. *Let $\mathcal{X}$ be the support of the distribution of X, and $\mathcal{Z}_x$ be the support of the distribution of the Z conditional on $X = x$. Let F denote the distribution function corresponding to the probability function $\mathbb{P}$. Then,*

1. *X is real-valued, $\mathbb{P}(X = x_0) > 0$, and there exists a neighborhood $\mathcal{N}$ of x_0 such that $\mathcal{N} \subset \mathcal{X}$.*

2. *The sets $\mathcal{Z}_x$ are identical, $\forall x \in \mathcal{N}$.*

3. $\mathbb{E}(Y \mid X = x, Z, Q)$ *exists for all the values of* x, Z *and* Q, *and is continuous in* x *at* $x = x_0$ *for all the values of* Z *and* Q.

4. *If for all* $x \in \mathcal{X}$, $x_0 \leqslant x$, *define* $\lim_{x \uparrow x_0} dF(Q \mid X = x, Z) = dF(q \mid x_0, z)$, *and if for all* $x \in \mathcal{X}$, $x_0 \geqslant x$, *define* $\lim_{x \downarrow x_0} dF(Q \mid X = x, Z) = dF(q \mid x_0, z)$. *Assume that* $\lim_{x \downarrow x_0} dF(Q \mid X = x, Z)$ *and* $\lim_{x \uparrow x_0} dF(Q \mid X = x, Z)$ *exist for all the values of* Z.

Assumption 6.1 (1) implies that $0 < \mathbb{P}(X_i = x_0) < 1$. It also implies that $\mathcal{X}$ contains at least one interval subset $\mathcal{N}$, and therefore X is locally a continuous random variable. For $0 \leqslant \alpha \leqslant 1$, define the quantity

$$\Delta(Z) = \mathbb{E}(Y \mid X = x_0, Z) - \left(\alpha \lim_{x \downarrow x_0} \mathbb{E}(Y \mid X = x, Z) + (1 - \alpha) \lim_{x \uparrow x_0} \mathbb{E}(Y \mid X = x, Z) \right),$$

where if x_0 is a lower boundary point in $\mathcal{N}$, then $\alpha = 1$, and if x_0 is at the upper boundary, $\alpha = 0$. $\Delta(Z)$ is the weighted right and left discontinuity of $\mathbb{E}(Y \mid X = x, Z)$ at $x = x_0$. Let θ be defined as

$$\theta = \int G(\Delta(Z), Z) d\nu(Z) \tag{6.1}$$

for a known function G and a known measure ν on the range of Z, and supposing that the integral exists.

This part presents a test of the null hypothesis

$$H_0\text{: } X \text{ is exogenous,}$$

against the alternative hypothesis,

$$H_1\text{: } X \text{ is endogenous.}$$

Implementation of the test requires the estimation of the parameter θ. It will be soon stated that X exogenous implies $\theta = 0$. This is fundamental to establishing that tests based on θ are well defined, in the sense that they have the correct asymptotic size under H_0.

Theorem 6.1. *If* ν *is identifiable,* $G(0, z) = 0$, $\forall z$ *and assumptions 6.1 and B.1 (see remark 6.2) are satisfied, then* θ *is identifiable and is equal to zero if* X *is exogenous.*

The proof is provided in appendix B.1.1. The argument consists of: if X exogenous, then $\mathbb{E}(Y \mid X = x, Z, Q) = \mathbb{E}(Y \mid X = x, Z)$ almost surely for all x and Z. This implies that $\mathbb{E}(Y \mid X = x, Z)$ is also continuous due to assumption B.1, and so $\Delta(Z) = 0$. $\theta = 0$ follows because $G(0, z) = 0$. It will be seen in chapter 7 that θ is in fact estimable (pending more conditions), and that its estimator (defined in chapter

7) is the discontinuity test statistic. An example of an identified ν is the distribution of Z. It yields $\theta = \mathbb{E}(G(\Delta(Z), z))$. If, for example, $G(\Delta(Z), Z) = \Delta(Z)^2$, then $\theta = \mathbb{E}(\Delta(Z)^2)$, which is the average square of the discontinuities $\Delta(Z)$.

Of particular interest is the case of $G(\Delta(Z), Z) = \Delta(Z)g(Z)$, for some real-valued function g in the domain of Z, and $\nu(Z) = F(Z \mid X = x_0)$. Here, from equation (6.1),

$$\begin{aligned}\theta =& \ \mathbb{E}(\Delta(Z)g(Z) \mid x = x_0) \\ =& \ \mathbb{E}(y\, g(Z) \mid x = x_0) - \Bigg[\alpha E\left(\lim_{x \downarrow x_0} \mathbb{E}(Y \mid X = x, Z)\, g(Z) \,\middle|\, x = x_0\right) \\ &+ (1-\alpha)\mathbb{E}\left(\lim_{x \uparrow x_0} \mathbb{E}(Y \mid X = x, Z)\, g(Z) \,\middle|\, x = x_0\right)\Bigg], \quad (6.2)\end{aligned}$$

via the law of iterated expectations, supposing that all the moments in (6.2) exist. This parameter will be useful because its estimation does not require the estimation of $\mathbb{E}(y \mid x = x_0, z)$, and because $\mathbb{E}(y\, g(Z) \mid x = x_0)$ can be estimated at the rate $\sqrt{n}$ if $P(x = x_0) > 0$, as will be shown in chapter 7.

Theorem 6.1 showed that under H_0, $\theta = 0$. Since the discontinuity test statistic is based on an estimator of θ, the discontinuity test has non-trivial power against H_1 if under H_1, $\theta \neq 0$. The following assumption determines cases in which the discontinuity test has non-trivial power against H_1.

Assumption 6.2. *Suppose assumption 6.1. Define, for a given $\alpha \in [0, 1]$,*

$$\begin{aligned}\zeta(Q, Z) :=& \ dF(Q \mid X = x_0, Z) \\ &- \left(\alpha \lim_{x \downarrow x_0} dF(Q \mid X = x, Z) + (1-\alpha) \lim_{x \uparrow x_0} dF(Q \mid X = x, Z)\right).\end{aligned}$$

Then, $\mathbb{P}(\zeta(Q, Z) \neq 0 \mid x) > 0$, for all the values of X in a neighborhood of x_0.

Assumption 6.2 implies that $dF(Q \mid X = x, Z)$ is discontinuous at x_0. The discontinuity may be different from the right or left hand side, or even exist for only one of the sides. The assumption also stipulates that the discontinuity is one sided if x_0 is at the boundary of $\mathcal{X}$, that is, if either $x_0 \leqslant x$, $\forall x \in \mathcal{X}$ or $x_0 \geqslant x$, $\forall x \in \mathcal{X}$. If x_0 is an interior point, knowledge of the particular problem can be used to choose α. If the right and left limits of $dF(Q \mid X = x, Z)$ are the same, the choice of α is irrelevant, and then $\zeta(Q, Z) := dF(Q \mid X = x_0, Z) - \lim_{x \to x_0} dF(Q \mid X = x, Z)$.

Assumption 6.2 defines x_0. In other words, the discontinuity test will have non-trivial power in the situations where there exists an x_0 in $\mathcal{X}$ such that assumption 6.2 is reasonable. This is established in the following result.

Result 1. *If assumptions 6.1, 6.2 and B.1 hold, then in general, if X is endogenous, $\theta \neq 0$.*

Result 1 is not a theorem, as Assumption 6.2 alone cannot guarantee that under H_1, $\theta \neq 0$. To understand this result, observe that by assumption B.1,

$$\begin{aligned}\theta = &\int G\Big(\Big[\int \mathbb{E}(Y \mid X = x_0, Z, Q) dF(Q \mid X = x_0, Z) - \\ &- \alpha \int \mathbb{E}(Y \mid X = x_0, Z, Q) \lim_{x \downarrow x_0} dF(Q \mid X = x, Z) \\ &- (1-\alpha) \int \mathbb{E}(Y \mid X = x_0, Z, Q) \lim_{x \uparrow x_0} dF(Q \mid X = x, Z) \Big], Z \Big) d\nu(z) \\ = &\int G\Big(\int \mathbb{E}(Y \mid X = x_0, Z, Q) \zeta(Q, Z), Z \Big) d\nu(Z) \end{aligned} \tag{6.3}$$

Assumption 6.2 guarantees that $\zeta(Q, Z) \neq 0$ with positive probability. However, it cannot be guaranteed that $\theta \neq 0$ unless further requirements are made concerning each of G, ν, $\mathbb{E}(Y \mid X, Z)$ and the shape of $\zeta(Q, Z)$, and this is why result 1 is said to hold "in general." Section 6.1 presents an example where conditions are given such that result 1 holds always. The results concerning the power of the test (theorems 7.3, 7.6 and 7.9) hold if result 1 holds.

Remark 6.1. *The discontinuity test could simply consist of forming an estimate of $\Delta(Z)$ for some value of Z, and then testing whether $\Delta(Z)$ is zero. However, such a test may have little power because $\Delta(Z)$ can in general be estimated only at very low convergence rates. In the interest of the accuracy of the estimation, and to avoid the problems that an incorrect choice of Z could occasion, it is preferable to aggregate the discontinuities. This leads to the definition of the parameter θ.*

Remark 6.2. *Condition B.1 in appendix B.1.1 requires the interchangeability of the integral and the limits in the following specification. For each Z, consider a sequence $x_n \downarrow x_0$. Define $\psi_n(Q) = \mathbb{E}(Y \mid X = x_n, Z, Q)$, $\psi(Q) = \mathbb{E}(Y \mid X = x_0, Z, Q)$, $\mu_n(Q) = F(Q \mid X = x_n, Z)$ and $\mu(Q) = d \lim_{n\to\infty} F(Q \mid X = x_n, Z)$, when they exist. By assumption 6.1 (1), $\psi_n \to \psi$ pointwise in Q. By the definition of the Riemann-Stieltjes integral,*

$$\lim_{n\to\infty} \int \psi_n \, d\mu_n = \lim_{n\to\infty} \lim_{\mathrm{d}Q \to 0} \sum f_n(Q^c) \mu_n(\mathrm{d}Q),$$

where Q^c is any point in the intervals of length $\mathrm{d}Q$. Then, assumption B.1 can be expressed as $\lim_{n\to\infty} \int \psi_n \, d\mu_n = \int \psi \, d\mu$. Conditions for this can be established with measure theory convergence theorems concerning changing the order of the limits and by requiring that the support of $dF(Q)$ be compact.

Remark 6.3. *The parameter used in the case with no covariates Z (described in chapter 5) cannot be used in the case where the Z are present. In that case, $\mathbb{E}(Y \mid X = x_0)$ is compared with the limit $\lim_{x\to x_0} \mathbb{E}(Y \mid X = x)$. Observe that the parameter*

θ controls the distribution of Z, because it uses the fixed measure ν to weight the different Z. In the simple comparison of $\lim_{x\to x_0}\mathbb{E}(Y\,|\,X=x)$ and $\mathbb{E}(Y\,|\,X=x_0)$, the distribution of Z, which is often discontinuous at $x=x_0$ can be responsible for a difference even when X is exogenous. To see this, notice that if X is exogenous, $\mathbb{E}(Y\,|\,X=x,Z,Q)=\mathbb{E}(Y\,|\,X,Z)$, and provided the limit can exchange places with the integral sign,

$$\begin{aligned}\lim_{x\to x_0}\mathbb{E}(Y\,|\,X=x) &= \lim_{x\to x_0}\int \mathbb{E}(Y\,|\,X=x,Z)dF(Z\,|\,X=x)\\ &= \int \mathbb{E}(Y\,|\,X=x_0,Z)\lim_{x\to x_0} dF(Z\,|\,X=x)\end{aligned}$$

and

$$\mathbb{E}(Y\,|\,X=x_0)=\int \mathbb{E}(Y\,|\,X=x_0,Z)\,dF(Z\,|\,X=x_0)$$

Since $\lim_{x\to x_0} dF(Z\,|\,X=x)$ and $dF(Z\,|\,X=x_0)$ can be and often are different, $\mathbb{E}(Y\,|\,X=x)$ can be discontinuous at x_0 even when X is exogenous, and therefore this comparison is useless for the detection of endogeneity.

Remark 6.4. *Theorem 6.1 allows for other random variables to enter the model. Suppose for example that*

$$Y=f(X,Z,Q,\varepsilon)$$

where ε is statistically independent of X, Z and Q. Provided f is continuous in X at x_0, it is direct to show that $\mathbb{E}(Y\,|\,X=x,Z,Q)$ will also be continuous at x_0. This will be done in appendix B.1.2.

6.1 A censoring example

Basic conditions for assumptions 6.1 and 6.2 can be set down for a model in which X is censored. Such situations could occur, for example, when X is the result of a cornered optimization problem. This is the case of the effects of maternal smoking example referred to in chapter 5 and analysed in chapter 11. There, smoking is a choice variable that cannot take on negative values. The following model and the suggested assumptions are not the weakest for identification of θ. They are meant to illustrate the point in an intuitive way.

Suppose an unobservable variable X^* is observed in its censored form $X=\max\{X^*,0\}$. Suppose ε is independent of the variables X, Z and Q, and that the variables Y, X, X^*, Z, Q and ε are related via the structural equations

$$Y=f_1(X,Z,Q)+\varepsilon \quad\text{and}\quad X^*=f_2(Z,Q).$$

Then

$$\mathbb{E}(Y\,|\,X,Z)=\begin{cases} f_1(X,Z,f_2^{-1}(X;Z)), & \text{if } X>0,\\ \mathbb{E}(f_1(0,Z,f_2^{-1}(X^*;Z))\,|\,X^*\leqslant 0,Z), & \text{if } X=0.\end{cases} \tag{6.4}$$

In this model, instead of assumptions 6.1 and 6.2, consider the following assumption:

Assumption 6.3.

1. *f_1 is continuous in X at $X = 0$, and if f_1 varies on Q, it is continuous and increasing in Q.*
2. *f_2 is strictly decreasing in X^*, and $f_2(\cdot; Z)^{-1}$ is continuous in X^*, $\forall Z$.*
3. $0 < \mathbb{P}(X^* < 0) < 1$.

Given the model and assumption 6.3, one has

$$\Delta(Z) = \mathbb{E}(f_1(0, Z, f_2^{-1}(X^*; Z)) \,|\, X^* \leqslant 0, Z) - f_1(0, Z, f_2^{-1}(0, Z)) > 0$$

if and only if f_1 varies in Q. Hence, if $G(\Delta(Z), Z) = \Delta(Z)g(Z)$, for a strictly positive function g, and suppose ν is not zero everywhere, then $\theta = \int \Delta(Z)g(Z)\, d\nu(Z) > 0$ if and only if X is endogenous.

In the context of the smoking example, suppose that Y is birthweight, X is smoking, X^* is "intended" smoking, and the Z are a set of covariates. If Y, X and Z satisfy the model and equation 6.4, then assumption 6.3 implies that even if the covariates are held constant, the average birthweight of babies born to nonsmoker mothers will be discontinuously higher than the birthweight of babies born to mothers that smoked positive if and only if smoking is endogenous.

Chapter 7

A discontinuity test of endogeneity

This chapter develops the discontinuity test statistic $\hat{\theta}$. It depends on the estimation of $\lim_{x \to x_0} \mathbb{E}(Y \,|\, X = x, Z)$. For that, diverse assumptions can be made about the nature of $\mathbb{E}(Y \,|\, X, Z)$. After a general discussion of the aspects of $\hat{\theta}$ that are common to all cases, this chapter contains three sections. Section 7.1 assumes that when $X \neq x_0$, $\mathbb{E}(Y \,|\, X, Z)$ is linear in X and Z, and develops an estimator that takes this assumption into account. Section 7.2 assumes that $\mathbb{E}(Y \,|\, X, Z)$ is partially linear (linear in Z), while section 7.3 allows $\mathbb{E}(Y \,|\, X = x, Z)$ to have a nonparametric structure. The three sections provide results about the asymptotic behavior of $\hat{\theta}$ in each case, as well as variance estimators and power considerations. The three sections stand alone.

The discontinuity test consists of the estimation of θ and testing whether it is equal to zero. A natural approach would be to adopt, for some $0 \leqslant \alpha \leqslant 1$,

$$\hat{\theta} = \int G(\hat{\Delta}(Z), z) d\hat{\nu}(z),$$

where

$$\hat{\Delta}(Z) = \hat{\mathbb{E}}(Y \,|\, X = x_0, Z) - \left(\alpha \hat{\mathbb{E}}(Y \,|\, x_0, Z)^{\downarrow} + (1 - \alpha) \hat{\mathbb{E}}(Y \,|\, x_0, Z)^{\uparrow}\right),$$

and $\hat{\mathbb{E}}(Y \,|\, x_0, Z)^{\downarrow}$ is an estimator of $\lim_{x \downarrow x_0} \mathbb{E}(Y \,|\, X = x, Z)$ and $\hat{\mathbb{E}}(Y \,|\, x_0, Z)^{\uparrow}$ is an estimator of $\lim_{x \uparrow x_0} \mathbb{E}(Y \,|\, X = x, Z)$ which will be defined differently in each section 7.1, 7.2 and 7.3.

As explained in chapter 6, the tests where $G(\Delta(Z), Z) = \Delta(Z)\, g(Z)$, for some g, and $\nu(Z) = F(Z \,|\, X = x_0)$ are of particular interest. They eliminate one step in the estimation of θ. Hence, the rest of this section will develop the discontinuity test for $\theta = \mathbb{E}(\Delta(Z)\, g(Z) \,|\, X = x_0)$. Let the data satisfy

Assumption 7.1. *Suppose*

1. *The observations* (Y_i, X_i, Z_i), $i = 1, ..., n$ *are i.i.d.,* $Z_i = (Z_i^1, \ldots, Z_i^d)^T$.

2. $0 < \mathbb{P}(X_i = x_0) < 1$.

3. $\mathbb{E}(|\Delta(Z_i)\, g(Z_i)|^{2+\xi_1} \mathbf{1}(X_i = x_0)) < \infty$ *for some* $\xi_1 > 0$.

Define $\epsilon_i = Y_i - \mathbb{E}(Y_i \,|\, X_i, Z_i)$, and define $V_A = \mathbb{V}ar(\Delta(Z_i)\, g(Z_i) \,|\, X_i = x_0)$. Let $\hat{p}_{x_0} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i = x_0)$ denote the method of moments estimator of $P(X_i = x_0)$. When $\hat{p}_{x_0} > 0$, the suggested estimator of θ is then, from equation (6.2),

$$\hat{\theta} = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^n \left[Y_i - \alpha \hat{\mathbb{E}}(Y_i \,|\, x_0, Z_i)^{\downarrow} - (1-\alpha) \hat{\mathbb{E}}(Y_i \,|\, x_0, Z_i)^{\uparrow} \right] g(Z_i)\, \mathbf{1}(X_i = x_0). \quad (7.1)$$

Define

$$\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\downarrow} = \lim_{x \downarrow x_0} \mathbb{E}(Y_i \,|\, X_i = x, Z_i), \qquad \hat{\Gamma}(z)^+ = \hat{\mathbb{E}}(Y \,|\, x_0, Z)^{\downarrow} - \mathbb{E}(Y \,|\, x_0, Z)^{\downarrow}$$
$$\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\uparrow} = \lim_{x \uparrow x_0} \mathbb{E}(Y_i \,|\, X_i = x, Z_i), \qquad \hat{\Gamma}(z)^- = \hat{\mathbb{E}}(Y \,|\, x_0, Z)^{\uparrow} - \mathbb{E}(Y \,|\, x_0, Z)^{\uparrow}.$$

Next write

$$\hat{\theta} - \theta = A_n - B_n$$

where

$$A_n = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^n \Delta(Z_i) g(Z_i) \mathbf{1}(X_i = x_0) - \mathbb{E}(\Delta(Z_i) g(Z_i) \,|\, X_i = x_0) \quad (7.2)$$

$$B_n = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^n [\alpha \hat{\Gamma}(Z_i)^+ + (1-\alpha)\hat{\Gamma}(Z_i)^-] g(Z_i) \mathbf{1}(X_i = x_0). \quad (7.3)$$

Under the null hypothesis that X_i is exogenous, $\Delta(Z_i) = 0$, and therefore $A_n = 0$. Results are developed for the case $A_n \neq 0$ for power considerations. Since A_n does not depend on the estimators of $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\downarrow}$ or $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\uparrow}$, neither does its asymptotic distribution. Assumption 7.1 item (1) and the LLN (p. 124 in Chow and Teicher (1997)) imply that $\hat{p}_{x_0} \xrightarrow{p} \mathbb{P}(X_i = x_0)$, and items (1) and (3) and the CLT (theorem 3.3.7 in Amemiya (1985)) imply that $\sqrt{n}\big(\frac{1}{n}\sum_{i=1}^n \Delta(Z_i) g(Z_i) \mathbf{1}(X_i = x_0) - \mathbb{E}(\Delta(Z_i) g(Z_i) \mathbf{1}(X_i = x_0))\big)$ is asymptotically normally distributed. Finally, item (2), the continuous mapping theorem (theorem 3.2.5 in Amemiya (1985)) and Slutsky's theorem (theorem 3.2.7 in Amemiya (1985)) imply that

$$\sqrt{n} A_n \xrightarrow{d} \mathcal{N}(0, V_A) \quad (7.4)$$

as $n \to \infty$. The asymptotic behavior of B_n, and hence of $\hat{\theta}$, depends on the assumptions one is willing to make on the nature of $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\downarrow}$ and $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\uparrow}$, and the related choice of estimators. This is done in the following sections.

7.1 The linear case

Suppose that for $X > x_0$, the conditional expectation satisfies

$$\mathbb{E}(Y \mid X, Z) = \beta^+ X + Z^T \gamma^+, \tag{7.5}$$

and for $X < x_0$, the conditional expectation satisfies

$$\mathbb{E}(Y \mid X, Z) = \beta^- X + Z^T \gamma^-.$$

If x_0 is the left boundary of $\mathcal{N}$, then $\beta^- = 0$ and $\gamma^- = 0$. If x_0 is the right boundary of $\mathcal{N}$, then $\beta^+ = 0$ and $\gamma^+ = 0$.

Example 1. *(Censoring) Equation* (7.5) *can be derived inside the censoring model presented in chapter 6.1. Suppose*

$$f(X, Z, Q) = \alpha_X X + Z^T \alpha_Z + \alpha_Q Q$$
$$g(Z, Q) = Z^T \pi_Z + Q$$

then, substituting into the conditional expectation equation (6.4) *for* $x > 0$,

$$\mathbb{E}(Y \mid X, Z) = (\alpha_x + \alpha_q)\, X + Z^T (\alpha_Z - \alpha_Q \pi_Z),$$

which translates into equation (7.5) *if* $\beta^+ := \alpha_X + \alpha_Q$ *and* $\gamma^+ := \alpha_Z - \alpha_Q \pi_Z$.

In the linear (in X and Z) case, $\mathbb{E}(Y_i \mid x_0, Z_i)^{\downarrow} = \beta^+ x_0 + Z_i^T \gamma^+$, and $\mathbb{E}(Y_i \mid x_0, Z_i)^{\uparrow} = \beta^- x_0 + Z_i^T \gamma^-$. The coefficients β^+ and γ^+ can be estimated by simply regressing Y_i on X_i and Z_i using only the observations for which $X_i > x_0$. Then $\hat{\mathbb{E}}(Y_i \mid x_0, Z_i)^{\downarrow} = \hat{\beta}^+ x_0 + Z_i^T \hat{\gamma}^+$. The result for $\hat{\mathbb{E}}(Y_i \mid x_0, Z_i)^{\uparrow}$ is analogous.

Next, let

$$W_i = (X_i, Z_i^T)^T, \quad \delta^+ = (\beta^+, \gamma^{+T})^T, \quad \delta^- = (\beta^-, \gamma^{-T})^T,$$

then if x_0 is an interior point of $\mathcal{N}$, the proposed method of moments estimators of γ^+ and γ^- are

$$\hat{\delta}^+ = \left(\sum_{i=1}^n W_i W_i^T \mathbf{1}(X_i > x_0) \right)^{-1} \sum_{i=1}^n W_i\, Y_i\, \mathbf{1}(X_i > x_0),$$
$$\hat{\delta}^- = \left(\sum_{i=1}^n W_i W_i^T \mathbf{1}(X_i < x_0) \right)^{-1} \sum_{i=1}^n W_i\, Y_i\, \mathbf{1}(X_i < x_0).$$

If x_0 is the left boundary of $\mathcal{N}$, then $\hat{\delta}^- = 0$. If x_0 is the right boundary of $\mathcal{N}$, then $\hat{\delta}^+ = 0$.

Let the estimators of $\mathbb{E}(g(Z_i) \,|\, X_i = x_0)$ and $\mathbb{E}(g(Z_i)Z_i \,|\, X_i = x_0)$ be defined as

$$\hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0) = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i = x_0) g(Z_i),$$
$$\hat{\mathbb{E}}(g(Z_i)Z_i \,|\, X_i = x_0) = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i = x_0) g(Z_i)Z_i,$$

then from equation (7.3),

$$B_n = \begin{bmatrix} \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \hat{\mathbb{E}}(g(Z_i)Z_i \,|\, X_i = x_0) \end{bmatrix}^T \left(\alpha(\hat{\delta}^+ - \delta^+) + (1-\alpha)(\hat{\delta}^- - \delta^-)\right)$$

Assumption 7.2. *Suppose*

1. $\mathbb{E}(|g(Z_i)| \,|\, X_i = x_0) < \infty$ *and* $\mathbb{E}(||g(Z_i)Z_i|| \,|\, X_i = x_0) < \infty$.
2. $\mathbb{V}ar(\epsilon_i \,|\, X_i, Z_i) = \sigma^2 < \infty$. *(See remark 7.1 below about relaxing this condition.)*
3. $\mathbb{E}(W_i W_i^T \mathbf{1}(X_i > x_0)) < \infty$ *is positive definite, and* $\mathbb{E}(W_i W_i^T \mathbf{1}(X_i < x_0)) < \infty$ *is positive definite.*

Theorem 7.1. *If assumptions 6.1, 7.1 and 7.2 hold, then*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V_A + V_B) \tag{7.6}$$

where

$$V_B = \sigma^2 \begin{bmatrix} \mathbb{E}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \mathbb{E}(g(Z_i)Z_i \,|\, X_i = x_0) \end{bmatrix}^T \big[\alpha^2 \mathbb{E}(W_i W_i^T \mathbf{1}(X_i > x_0))^{-1} + \\ + (1-\alpha)^2 \mathbb{E}(W_i W_i^T \mathbf{1}(X_i < x_0))^{-1}\big] \begin{bmatrix} \mathbb{E}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \mathbb{E}(g(Z_i)Z_i \,|\, X_i = x_0) \end{bmatrix}.$$

The proof is similar to the classical proofs of the asymptotic properties of the OLS estimator. The absence of a term to account for the correlation of A_n and B_n follows because $\left(\alpha(\hat{\delta}^+ - \delta^+) + (1-\alpha)(\hat{\delta}^- - \delta^-)\right)$ is independent of A_n and of $\hat{\mathbb{E}}(g(Z_i)Z_i \,|\, X_i = x_0)$, since the two latter only use observations for which $X_i = x_0$, while the former has zero mean and only uses observations for which $X_i \neq x_0$. The absence of a cross term in V_B happens because $\hat{\delta}^+$ and $\hat{\delta}^+$ are built using different parts of the sample, and are therefore independent. See the proof in detail in the appendix B.2.1.

Theorem 7.2. *Under H_0: X_i is exogenous, $\theta = 0$ and $\sqrt{n}\,\hat{\theta} \xrightarrow{d} \mathcal{N}(0, V_B)$. If assumptions 6.1, 7.1 and 7.2 hold, V_B can be consistently estimated by*

$$\hat{V}_B = \hat{\sigma}^2 \begin{bmatrix} \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0) \end{bmatrix}^T \big[\alpha^2 \hat{\mathbb{E}}(W_i W_i^T \mathbf{1}(X_i > x_0))^{-1} + \\ + (1-\alpha)^2 \hat{\mathbb{E}}(W_i W_i^T \mathbf{1}(X_i < x_0))^{-1}\big] \begin{bmatrix} \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0) \end{bmatrix},$$

where

$$\hat{\mathbb{E}}(W_i W_i^T \mathbf{1}(X_i > x_0)) = \frac{1}{n}\sum_{i=1}^{n} W_i W_i^T \mathbf{1}(X_i > x_0),$$

$$\hat{\mathbb{E}}(W_i W_i^T \mathbf{1}(X_i < x_0)) = \frac{1}{n}\sum_{i=1}^{n} W_i W_i^T \mathbf{1}(X_i < x_0),$$

$$\hat{\sigma}^2 = \frac{1}{1-\hat{p}_{x_0}}\left[\alpha \frac{1}{n}\sum_{i=1}^{n} (Y_i - W_i^T \hat{\gamma}^+)^2 \mathbf{1}(X_i > x_0) + (1-\alpha)\frac{1}{n}\sum_{i=1}^{n} (Y_i - W_i^T \hat{\gamma}^-)^2 \mathbf{1}(X_i < x_0)\right].$$

The convergence in probability of $\hat{\sigma}^2$ to σ^2 is established by noticing that $\hat{\sigma}^2$ is simply a weighted average of two standard estimators of the variance of ϵ_i using weighted least squares. The convergence of $\hat{V}_B$ follows from the LLN (p. 124 in Chow and Teicher (1997)) applied to $\hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)$, $\hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0)$, $\hat{\mathbb{E}}(W_i W_i^T \mathbf{1}(X_i > x_0))$ and $\hat{\mathbb{E}}(W_i W_i^T \mathbf{1}(X_i < x_0))$ (given assumptions 7.1 (1) and 7.2 (1) and (3)) and Slutsky's theorem (theorem 3.2.7 in Amemiya (1985)). The following theorem gives the discontinuity test properties in the linear case.

Theorem 7.3. *Let $0 \leqslant \lambda \leqslant 1$. Let Φ be the standard normal cumulative distribution function, and let $c_\lambda = \Phi^{-1}(\lambda)$. If the assumptions of theorems 6.1, 7.1 and 7.2 hold, then under*

$$H_0\text{: } X \text{ is exogenous,}$$

$$\mathbb{P}\left(\sqrt{n}\frac{\hat{\theta}}{\sqrt{\hat{V}_B}} \leqslant c_\lambda\right) \to \lambda \quad as \quad n \to \infty.$$

moreover, if result 1 is true, then under

$$H_1\text{: } X \text{ is endogenous,}$$

$$\mathbb{P}\left(\sqrt{n}\frac{\hat{\theta}}{\sqrt{\hat{V}_B}} > c_\lambda\right) \to 1 \quad as \quad n \to \infty,$$

and under the local alternatives $\frac{\theta}{\sqrt{n}}$,

$$\mathbb{P}\left(\sqrt{n}\frac{\hat{\theta}}{\sqrt{\hat{V}_B}} \leqslant c_\lambda\right) \to \Phi\left(\frac{c_\lambda \sqrt{V_B} - \theta}{\sqrt{V_A + V_B}}\right) \quad as \quad n \to \infty.$$

See proof in appendix B.2.2.

Remark 7.1. *Homoskedasticity can be relaxed. Let W be the matrix whose rows are the W_i^T, let W^+ be the matrix whose rows are the $\mathbf{1}(X_i > x_0)W_i^T$, and let W^- be the matrix whose rows are the $\mathbf{1}(X_i < x_0)W_i^T$. Suppose $\mathbb{V}ar(\epsilon \,|\, W) = \Sigma$, then*

$$V_B = \begin{bmatrix} \mathbb{E}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \mathbb{E}(g(Z_i)Z_i \,|\, X_i = x_0) \end{bmatrix}^T \left[\alpha^2 V_1 + (1-\alpha)^2 V_2\right] \begin{bmatrix} \mathbb{E}(g(Z_i) \,|\, X_i = x_0)\, x_0 \\ \mathbb{E}(g(Z_i)Z_i \,|\, X_i = x_0) \end{bmatrix},$$

where

$$V_1 = \mathbb{E}(W_i W_i^T \mathbf{1}(X_i > x_0))^{-1} \mathbb{E}\left(\underset{n\to\infty}{plim} \frac{W^{+T}\Sigma W^+}{n}\right) \mathbb{E}(W_i W_i^T \mathbf{1}(X_i > x_0))^{-1},$$
$$V_2 = \mathbb{E}(W_i W_i^T \mathbf{1}(X_i < x_0))^{-1} \mathbb{E}\left(\underset{n\to\infty}{plim} \frac{W^{-T}\Sigma W^-}{n}\right) \mathbb{E}(W_i W_i^T \mathbf{1}(X_i < x_0))^{-1},$$

and $\underset{n\to\infty}{plim}$ denotes the limit in probability, supposing the limits exist. V_1 can be estimated using the Eicker-White covariance matrix (see White (1980))of an OLS regression of the Y_i onto X_i and Z_i using only observations such that $X_i > x_0$, and V_2 can be estimated analogously, using only observations such that $X_i < x_0$.

7.2 The partially linear case

Suppose that for $X > x_0$, the conditional expectation satisfies

$$\mathbb{E}(Y_i \,|\, X_i, Z_i) = \tau^+(X_i) + Z_i^T \gamma^+, \tag{7.7}$$

and for $X < x_0$, the conditional expectation satisfies

$$\mathbb{E}(Y_i \,|\, X_i, Z_i) = \tau^-(X_i) + Z_i^T \gamma^-,$$

where $\tau^+(x_0)^\downarrow := \lim_{x\downarrow x_0} \tau^+(X)$ and $\tau^-(x_0)^\uparrow := \lim_{x\uparrow x_0} \tau^-(X)$ exist. If x_0 is the left boundary of $\mathcal{X}$, then $\tau^-(X_i) = 0$ for all X_i and $\gamma^- = 0$. If x_0 is the right boundary of $\mathcal{X}$, then $\tau^+(X_i) = 0$ for all X_i and $\gamma^+ = 0$.

Example 2. *(Censoring) Equation* (7.7) *can be derived inside the censoring model presented in section 6.1. Suppose*

$$f(X, Z, Q) = \psi_1(X) + z'\alpha_Z + \alpha_Q Q,$$
$$g(Z, Q) = \psi_2(Z'\pi_Z + Q),$$

where ψ_2 is invertible. Then, substituting into equation (6.4) *for $X > 0$,*

$$\mathbb{E}(Y \,|\, X, Z) = (\psi_1(X) + \alpha_Q \psi_2^{-1}(X))\, X + Z^T(\alpha_Z - \alpha_Q \pi_Z).$$

which translates into equation (7.7) *if $\tau^+(X) := \psi_1(X) + \alpha_Q \psi_2^{-1}(X)$, and $\gamma^+ := \alpha_Z - \alpha_Q \pi_Z$.*

In the partially linear case, $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\downarrow} = \tau^+(x_0)^{\downarrow} + Z_i^T\gamma^+$, and $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\uparrow} = \tau^-(x_0)^{\uparrow} + Z_i^T\gamma^-$. Hence, $\hat{\mathbb{E}}(Y_i \,|\, x_0, Z_i)^{\downarrow} = \hat{\tau}^+(x_0)^{\downarrow} + Z_i^T\hat{\gamma}^+$, and $\hat{\mathbb{E}}(Y_i \,|\, x_0, Z_i)^{\uparrow} = \hat{\tau}^-(x_0)^{\uparrow} + Z_i^T\hat{\gamma}^-$. Define

$$\hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0) = \frac{1}{\hat{p}_{x_0}}\frac{1}{n}\sum_{i=1}^{n} g(Z_i)\mathbf{1}(X_i = x_0),$$
$$\hat{\mathbb{E}}(g(Z_i)Z_i \,|\, X_i = x_0) = \frac{1}{\hat{p}_{x_0}}\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = x_0)g(Z_i)Z_i,$$

then

$$\begin{aligned} B_n = {} & \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)\left[\alpha(\hat{\tau}^+(x_0)^{\downarrow} - \tau^+(x_0)^{\downarrow}) + (1-\alpha)(\hat{\tau}^-(x_0)^{\uparrow} - \tau^-(x_0)^{\uparrow})\right] \\ & + \hat{\mathbb{E}}(g(Z_i)Z_i \,|\, X_i = x_0)^T\left[\alpha(\hat{\gamma}^+ - \gamma^+) + (1-\alpha)(\hat{\gamma}^- - \gamma^-)\right]. \end{aligned} \tag{7.8}$$

The following discussion refers to the estimation of $\tau^+(x_0)^{\downarrow}$ and γ^+. $\tau^-(x_0)^{\uparrow}$ and γ^- are estimated analogously. The estimation of the parametric component in the partially linear regression has been widely discussed in the literature. In the later papers (after Robinson (1988)), the generally adopted technique is that of subtracting the conditional expectation of Y_i given X_i so as to eliminate the nonparametric part. The resulting equation is

$$Y_i - \mathbb{E}(Y_i \,|\, X_i) = (Z_i - \mathbb{E}(Z_i \,|\, X_i))^T\gamma^+ + \epsilon_i, \quad \text{for} \quad X_i > x_0. \tag{7.9}$$

The coefficient of the constant term among the covariates is not identified and is eliminated in the subtraction, so Z_i in this equation does not include a constant term.

Robinson (1988) first suggested this approach. He estimated the conditional expectations using kernel regression, and peformed an OLS regression of $Y_i - \hat{\mathbb{E}}(Y_i \,|\, X_i)$ on $Z_i - \hat{\mathbb{E}}(Z_i \,|\, X_i)$, to obtain $\hat{\gamma}^+$. Robinson showed that the estimated $\hat{\gamma}^+$ converges to γ^+ at the rate $\sqrt{n}$, even though the regression includes nonparametric plugins. The following literature established the same $\sqrt{n}$ rate of convergence and the asymptotic distribution of γ^+ for an array of different nonparametric plugins. See for example Linton (1995) when the nonparametric component is estimated using local polynomial regression, and Li (2000) when the nonparametric component is estimated using series or spline orthogonal bases.

The basic technique for the estimation of the nonparametric component is rather intuitive. It consists of a nonparametric regression of $Y_i - Z_i^T\hat{\gamma}^+$ on X_i, and the variations depend on the nature of $\hat{\gamma}^+$ and the regression technique chosen. Since the rates of convergence of this component are slower than $\sqrt{n}$, the asymptotic behavior of the estimated nonparametric component is a simple extension of the results for regular nonparametric regression, because the estimated parametric component is estimated at the faster rate $\sqrt{n}$. The case of interest for this section is more delicate, because the value of interest is $\tau^+(x_0)^{\downarrow}$, which is the limit of the nonparametric component at

a boundary point. There are two difficulties, the first is that nonparametric estimation at boundary points requires especial attention in the choice of the estimator and in the asymptotic treatment. For this reason $\tau^+(x_0)^\downarrow$ is estimated using local polynomial regression, since this technique has been shown to possess excellent boundary properties (for example, p. 69 in Fan and Gijbels (1996)).

Though other techniques could also be used, such as for example a simple kernel regression using boundary kernels, the local polynomial regression is also desirable in that it requires no especial tailoring for the boundaries. Hence, the researcher needs to apply no extra discretion than for a regular nonparametric regression. Porter (2003) developed the asymptotic theory for the local polynomial estimator of the discontinuity in the regression discontinuity design. His method is to estimate the right and left limits of the discontinuous function at the point of discontinuity using local polynomial regression, and he derives results for arbitrary choice of the polynomial degree. This part provides the extension of his results to the partially linear case, in which the dependent variable in the local polynomial regression, $Y_i - Z_i^T\hat{\gamma}^+$, contains a plugin estimator of the parametric component. Though from an asymptotic point of view the extension is very simple, this section explicits the variance terms up to the $O(h)$ magnitude, which requires the careful consideration of the covariances between the parametric and nonparametric parts of the estimation. Moreover, the results are presented for a generic nonparametric plugin for $\hat{\mathbb{E}}(Y_i \,|\, X_i)$ and $\hat{\mathbb{E}}(Y_i \,|\, X_i)$, so that the plugins can be estimated with other, sometimes more practical, techniques such as series estimators.

The second difficulty is that x_0 is a point with positive probability in $\mathcal{X}$. The available theory on local polynomial estimators relies on the existence of a density function in a neighborhood of x_0. However, when using local polynomial estimators to estimate the limit of a function at a point, the observations at the point itself are not used. In fact, although Porter (2003) requires the existence of a density function, the proofs do not use the entire support of $dF(x)$ at once, but rather separate the observations to the right and to the left of x_0. This section adapts Porter's result using distribution functions conditional on $X_i \neq x_0$, which have a density function by assumption, though with possibly different right and left limits at x_0. As a consequence the same results as in Porter (2003) can be derived in terms of limits, therefore generalizing Porter's results to allow both for the positive probability of $X_i = x_0$, and also for the density function of $F(X_i \,|\, X_i = x)$ to have different right and left limits at x_0. It is important to notice that because the limits may be different, the variance estimator suggested by Porter in theorem 4 cannot be used in this case. Theorem 7.5 below proposes a different estimator which allows for the different right and left limits of the density at x_0.

If x_0 is an interior point, or is at the left boundary of the $\mathcal{X}$, the estimator $\tau^+(x_0)^\downarrow$ is defined in the following way. Given the kernel function k, the smoothing parameter

h, the polynomial degree p, and let $\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p$ be the solution the problem

$$\min_{a_0,\ldots,a_p} \frac{1}{n}\sum_{j=1}^{n} k\left(\frac{X_j - x_0}{h}\right) \mathbf{1}(X_j > x_0)\,[Y_j - Z_j^T\hat{\gamma}^+ - a_0 - a_1(X_j - x_0) - \ldots - a_p(X_j - x_0)^p]^2,$$

the local polynomial estimator of $\tau^+(x_0)^\downarrow$ is given by

$$\hat{\tau}^+(x_0)^\downarrow = \hat{a}_0 = e_1^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+(\mathrm{Y} - \mathrm{Z}\hat{\gamma}^+), \tag{7.10}$$

where $e_1 = (1, 0, \ldots, 0)^T$ has dimension $1 \times (p+1)$, X has rows equal to $(1, (X_j - x_0), \ldots, (X_j - x_0)^p)$, for $j = 1, ..., n$, D^+ is a $n \times n$ diagonal matrix with diagonal $\left\{\mathbf{1}(X_1 > x_0)\, k\left(\frac{X_1 - x_0}{h}\right), \ldots \ldots, \mathbf{1}(X_n > x_0)\, k\left(\frac{X_n - x_0}{h}\right)\right\}$, $\mathrm{Y} = (Y_1, \ldots, Y_n)^T$, and $\mathrm{Z} = [Z_1 \ \ldots \ Z_n]^T$. If x_0 is at the right boundary of $\mathcal{X}$, the estimator $\hat{\tau}^+(x_0)^\downarrow = 0$.

The next conditions make it possible to obtain the asymptotic distribution of B_n given in equation (7.8). The essence of the proof can be understood by observing that when $\hat{\tau}^+(x_0)^\downarrow$ is defined as in (7.10),

$$\begin{aligned}\hat{\tau}^+(x_0)^\downarrow &:= e_1^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+(\mathrm{Y} - \mathrm{Z}\hat{\gamma}^+) \\ &= e_1^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+(\mathrm{Y} - \mathrm{Z}\gamma^+) - e_1^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+ \mathrm{Z}(\hat{\gamma}^+ - \gamma^+),\end{aligned} \tag{7.11}$$

Define

$$\tilde{\tau}^+(x_0)^\downarrow = e_1^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+(\mathrm{Y} - \mathrm{Z}\gamma^+).$$

$\tilde{\tau}^+(x_0)^\downarrow$ is a simple local polynomial estimator of a boundary point seen, as discussed, in Porter (2003), but also examined in Fan and Gijbels (1996). Deriving its asymptotic distribution in this case needs only a small modification to account for the fact that X does not have a density function, since $\mathbb{P}(X_i = x_0) > 0$. It converges to a normally distributed random variable at the rate $\sqrt{nh}$. The second term can be considered jointly with the second term in equation (7.8), which converges at the rate $\sqrt{n}$. For testing in smaller samples, the results consider the effect of the estimation of γ^+ and γ^-. However both the bias and variance of $\hat{\tau}^+(x_0)^\downarrow$ and $\hat{\tau}^-(x_0)^\uparrow$ dominate the asymptotic behavior of $\hat{\theta}$.

Assumption 7.3.

1. $\mathbb{E}(|g(Z_i)|^{2+\xi_2} \,|\, X_i = x_0) < \infty$ *and* $\mathbb{E}(||g(Z_i)Z_i||^{2+\xi_2} \,|\, X_i = x_0) < \infty$, *for some* $\xi_2 > 0$.

2. *If* x_0 *is an interior point in* $\mathcal{X}$, *then the estimators* $\hat{\gamma}^+$ *and* $\hat{\gamma}^-$ *are defined as*

$$\begin{array}{ll}
\hat{\gamma}^+ = (\mathrm{Z}_+^T\mathrm{Z}_+)^{-1}\mathrm{Z}_+^T\mathrm{Y}_+, & \hat{\gamma}^- = (\mathrm{Z}_-^T\mathrm{Z}_-)^{-1}\mathrm{Z}_-^T\mathrm{Y}_-, \\
\mathrm{Y}_{i+} = (Y_i - \hat{\mathbb{E}}(Y_i \,|\, X_i)^+)\mathbf{1}(X_i > x_0), & \mathrm{Y}_{i-} = (Y_i - \hat{\mathbb{E}}(Y_i \,|\, X_i)^-)\mathbf{1}(X_i < x_0) \\
\mathrm{Z}_{i+} = (Z_i - \hat{\mathbb{E}}(Z_i \,|\, X_i)^+)\mathbf{1}(X_i > x_0), & \mathrm{Z}_{i-} = (Z_i - \hat{\mathbb{E}}(Z_i \,|\, X_i)^-)\mathbf{1}(X_i < x_0) \\
\hat{\mathbb{E}}(Y_i \,|\, X_i)^+ = \sum_{j=1}^n \mathbf{1}(X_j > x_0)\, T_{i,j}^+ Y_j, & \hat{\mathbb{E}}(Y_i \,|\, X_i)^- = \sum_{j=1}^n \mathbf{1}(X_j < x_0)\, T_{i,j}^- Y_j, \\
\hat{\mathbb{E}}(Z_i \,|\, X_i)^+ = \sum_{j=1}^n \mathbf{1}(X_j > x_0)\, T_{i,j}^+ Z_j, & \hat{\mathbb{E}}(Z_i \,|\, X_i)^- = \sum_{j=1}^n \mathbf{1}(X_j < x_0)\, T_{i,j}^- Z_j,
\end{array}$$

for some $T^+_{i,j}$ *and* $T^-_{i,j}$ *which are a function exclusively of the observations such that* $X_i > x_0$ *and* $X_i < x_0$ *respectively. Additionally,* $\sup_i \left\| \sum_{j=1}^n \mathbf{1}(X_j > x_0)\, T^+_{i,j} u_j - \mathbb{E}(u_i \,|\, X_i) \right\| = o_p(1)$ *for* $u_i = Z_i, \epsilon_i^2, \mathbb{E}(Z_i \,|\, X_i)\epsilon_i^2, Z_i \mathbb{E}(\epsilon_i^2 \,|\, X_i)$ *and* $\mathbb{E}(Z_i \,|\, X_i)\mathbb{E}(\epsilon_i^2 \,|\, X_i)$. $\hat{\gamma}^+$ *and* $\hat{\gamma}^-$ *satisfy*

$$\sqrt{n} \begin{bmatrix} \hat{\gamma}^+ - \gamma^+ \\ \hat{\gamma}^- - \gamma^- \\ A_n \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathcal{V}^+_\gamma & 0 & 0 \\ 0 & \mathcal{V}^-_\gamma & 0 \\ 0 & 0 & V_A \end{bmatrix} \right),$$

and there exist $\hat{\mathcal{V}}^+_{\gamma n}$ *and* $\hat{\mathcal{V}}^-_{\gamma n}$, *functions exclusively of data for which* $X_i > x_0$ *and* $X_i < x_0$ *respectively, and such that* $\hat{\mathcal{V}}^+_{\gamma n} \xrightarrow{p} \mathcal{V}^+_{\gamma n}$ *and* $\hat{\mathcal{V}}^-_{\gamma n} \xrightarrow{p} \mathcal{V}^-_{\gamma n}$. *Moreover,* $\mathbb{E}(\|\sqrt{n}(\hat{\gamma}^+ - \gamma^+)\|^{2+\xi_3})$ *and* $\mathbb{E}(\|\sqrt{n}(\hat{\gamma}^- - \gamma^-)\|^{2+\xi_3})$ *are uniformly bounded for all* n *and some* $\xi_3 > 0$. *If* x_0 *is the left boundary of* $\mathcal{X}$, *all is true except that* $\hat{\gamma}^- = 0$, $\mathcal{V}^-_\gamma = 0$, *and* $\hat{\mathcal{V}}^-_\gamma = 0$. *If* x_0 *is the right boundary of* $\mathcal{X}$, *all is true except that* $\hat{\gamma}^+ = 0$, $\mathcal{V}^+_\gamma = 0$, *and* $\hat{\mathcal{V}}^+_\gamma = 0$.

3. *There exist* $x^-, x^+ \in \mathbb{R}$, *with* $x^- < x_0 < x^+$ *such that* $F(x)$ *is twice continuously differentiable in* $[x_-\,,\, x_0) \cup (x_0\,,\, x^+]$ *with first derivative bounded away from zero and second derivative uniformly bounded in* $[x_-\,,\, x_0) \cup (x_0\,,\, x^+]$. *Define*

$$\begin{aligned} \phi(x_0)^\downarrow &:= \lim_{x \downarrow x_0} \tfrac{d}{dx} F(x), & \phi(x_0)^\uparrow &:= \lim_{x \uparrow x_0} \tfrac{d}{dx} F(x), \\ \phi'(x_0)^\downarrow &:= \lim_{x \downarrow x_0} \tfrac{d^2}{dx^2} F(x), & \phi'(x_0)^\uparrow &:= \lim_{x \uparrow x_0} \tfrac{d^2}{dx^2} F(x), \end{aligned}$$

then all of these quantities exist. Moreover, there exist $\hat{\phi}(x_0)^\downarrow$ *and* $\hat{\phi}(x_0)^\downarrow$, *consistent estimators of* $\phi(x_0)^\downarrow$ *and* $\phi(x_0)^\downarrow$ *respectively (see remark 7.2).*

4. *The function* $\tau^+(X)$ *is at least* $p+2$ *times continuously differentiable in* $(x_0, x^+]$, *and the function* $\tau^-(X)$ *is at least* $p+2$ *times continuously differentiable in* $[x^-, x_0)$. *Define*

$$\tau^{+(m)}(x_0)^\downarrow := \lim_{x \downarrow x_0} \tfrac{d^m}{dx^m} \tau^+(X), \qquad \tau^{-(m)}(x_0)^\uparrow := \lim_{x \downarrow x_0} \tfrac{d^m}{dx^m} \tau^-(X),$$

then these quantities exist for $m = 1, \ldots, p+2$.

5. *The variances* $\sigma^2(x) := \mathbb{E}(\epsilon_i^2 \,|\, X_i = x)$ *are at least* $p+2$ *continuously differentiable in* $[x^-, x_0) \cup (x_0, x^+]$. *The errors* $\epsilon_i^{\epsilon^2} = \epsilon_i^2 - \sigma^2(X_i)$ *have moments* $\mathbb{E}(|\epsilon_i^{\epsilon^2}|^{2+\xi_4} \,|\, X_i)$ *uniformly bounded for some* $\xi_5 > 0$. *Define*

$$\sigma^2(x_0)^\downarrow := \lim_{x \downarrow x_0} \sigma^2(x), \qquad \sigma^2(x_0)^\uparrow := \lim_{x \uparrow x_0} \sigma^2(x),$$

then these quantities exist.

6. *The kernel k is symmetric and has bounded support. For all j odd integers, $\int k^x(u)u^j du = 0$. Define $\upsilon_j = \int_0^\infty k(u)u^j du$ and $\omega_j = \int_0^\infty k^2(u)u^j du$, then*

$$\Upsilon_j = \begin{bmatrix} \upsilon_j \\ \vdots \\ \upsilon_{j+p} \end{bmatrix}, \Lambda_j = \begin{bmatrix} \upsilon_j & \dots & \upsilon_{j+p} \\ \vdots & & \vdots \\ \upsilon_{j+p} & \dots & \upsilon_{j+2p+1} \end{bmatrix}, \Omega_j = \begin{bmatrix} \omega_j \\ \vdots \\ \omega_{j+p} \end{bmatrix}, \Omega = \begin{bmatrix} \omega_j & \dots & \omega_{j+p} \\ \vdots & & \vdots \\ \omega_{j+p} & \dots & \omega_{j+2p+1} \end{bmatrix}.$$

7. *$\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$, and $\lim_{n\to\infty} h^{p+1}\sqrt{n} < \infty$.*

8. *The functions $\mathbb{E}(Z_i^d \,|\, X_i = x)$ and $\mathbb{E}(\epsilon_i^2 Z_i^d \,|\, X_i = x)$ are at least $p+2$ times continuously differentiable in $[x^-, x_0) \cup (x_0, x^+]$. The $\epsilon_i^z := Z_i - \mathbb{E}(Z_i \,|\, X_i)$ have moments $\mathbb{E}(||\epsilon_i^z||^{2+\xi_5} \,|\, X_i)$ uniformly bounded for some $\xi_5 > 0$. Define*

$$\begin{aligned}
\mathbb{E}(Z_i \,|\, X_i = x_0)^\downarrow &:= \lim_{x \downarrow x_0} \mathbb{E}(Z_i \,|\, X_i = x), \\
\mathbb{E}(Z_i \,|\, X_i = x_0)^\uparrow &:= \lim_{x \uparrow x_0} \mathbb{E}(Z_i \,|\, X_i = x), \\
\mathbb{E}(Z_i Z_i^T \,|\, X_i = x_0)^\downarrow &:= \lim_{x \downarrow x_0} \mathbb{E}(Z_i Z_i^T \,|\, X_i = x), \\
\mathbb{E}(Z_i Z_i^T \,|\, X_i = x_0)^\uparrow &:= \lim_{x \uparrow x_0} \mathbb{E}(Z_i Z_i^T \,|\, X_i = x), \\
\mathbb{E}(Z_i \epsilon_i^2 \,|\, X_i = x_0)^\downarrow &:= \lim_{x \downarrow x_0} \mathbb{E}(Z_i \sigma^2(X_i, Z_i) \,|\, X_i = x), \\
\mathbb{E}(Z_i \epsilon_i^2 \,|\, X_i = x_0)^\uparrow &:= \lim_{x \uparrow x_0} \mathbb{E}(Z_i \sigma^2(X_i, Z_i) \,|\, X_i = x),
\end{aligned}$$

then all of these quantities exist. Finally, define the notation

$$\begin{aligned}
\Sigma_z(x_0)^\downarrow &:= \mathbb{E}(Z_i Z_i^T \,|\, X_i = x_0)^\downarrow - \mathbb{E}(Z_i \,|\, X_i = x_0)^\downarrow \mathbb{E}(Z_i \,|\, X_i = x_0)^{\downarrow T} \\
\Sigma_z(x_0)^\uparrow &:= \mathbb{E}(Z_i Z_i^T \,|\, X_i = x_0)^\uparrow - \mathbb{E}(Z_i \,|\, X_i = x_0)^\uparrow \mathbb{E}(Z_i \,|\, X_i = x_0)^{\uparrow T} \\
c_{z\epsilon^2}(x_0)^\downarrow &:= \mathbb{E}(Z_i \epsilon_i^2 \,|\, X_i = x_0)^\downarrow - \mathbb{E}(Z_i \,|\, X_i = x_0)^\downarrow \sigma^2(x_0)^\downarrow \\
c_{z\epsilon^2}(x_0)^\uparrow &:= \mathbb{E}(Z_i \epsilon_i^2 \,|\, X_i = x_0)^\uparrow - \mathbb{E}(Z_i \,|\, X_i = x_0)^\uparrow \sigma^2(x_0)^\uparrow
\end{aligned}$$

9. *If x_0 is the left boundary of $\mathcal{X}$, then $\alpha = 1$, and if x_0 is the right boundary of $\mathcal{X}$, then $\alpha = 0$.*

Theorem 7.4. *If assumptions 6.1, 7.1 and 7.3 hold, then*

$$\sqrt{nh}\mathcal{V}_n^{-1/2}(\hat{\theta} - \theta - \mathcal{B}_n) \xrightarrow{d} \mathcal{N}(0, 1),$$

where

$$\mathcal{B}_n = \mathbb{E}(g(Z_i) \,|\, X_i = x_0) \left[\alpha \mathcal{B}_n^+ + (1-\alpha)\mathcal{B}_n^-\right],$$

$$\mathcal{B}_n^+ = \begin{cases} h^{p+1} \frac{\tau^{+(p+1)}(x_0)^{\lim}}{(p+1)!} e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+1}), & \text{if } p \text{ is odd,} \\ h^{p+2} \left[\frac{\tau^{+(p+1)}(x_0)^{\lim}}{(p+1)!} \frac{\phi'(x_0)^{\downarrow}}{\phi(x_0)^{\downarrow}}\right] e_1^T \Lambda_0^{-1} (\Upsilon_{p+2} - \Lambda_1 \Lambda_0 \Upsilon_{p+1}) & \\ \quad + \left[\frac{\tau^{+(p+2)}(x_0)^{\lim}}{(p+2)!}\right] e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+2}) & \text{if } p \text{ is even,} \end{cases}$$

and analogously for $\mathcal{B}_n^-$, substituting the "+" by "−" in the notation.

$$\mathcal{V}_n = \alpha^2 \left[\mathcal{V}_\tau^+ + 2\sqrt{h} C_+^T \mathcal{C}_{\tau\gamma}^+ + h C_+^T \mathcal{V}_\gamma^+ C_+\right] + (1-\alpha)^2 \left[\mathcal{V}_\tau^- + 2\sqrt{h} C_-^T \mathcal{C}_{\tau\gamma}^- + h C_-^T \mathcal{V}_\gamma^- C_-\right] + h V_A + o(h),$$

where if x_0 is an interior point or is at the left boundary of $\mathcal{X}$,

$$\begin{aligned} \mathcal{V}_\tau^+ &= \mathbb{E}(g(Z_i) \,|\, X_i = x_0)^2 \mathcal{V}^+, \\ \mathcal{V}^+ &= \frac{\sigma^2(x_0)^{\downarrow}}{\phi(x_0)^{\downarrow}} e_1^T \Lambda_0^{-1} \Omega \, \Lambda_0^{-1} e_1, \\ C_+ &= \mathbb{E}(g(Z_i) Z_i \,|\, X_i = x_0) - \mathbb{E}(g(Z_i) \,|\, X_i = x_0) \, \mathbb{E}(Z_i \,|\, X_i = x_0)^{\downarrow}, \\ \mathcal{C}_{\tau\gamma}^+ &= \left(\Sigma_z(x_0)^{\downarrow}\right)^{-1} c_{z\epsilon^2}(x_0)^{\downarrow}, \end{aligned}$$

and if x_0 is an interior point or is at the right boundary of the support of the X_i, $\mathcal{V}_\tau^-$, $\mathcal{V}_-$, C_- and $\mathcal{C}_{\tau\gamma}^-$ are defined analogously, substituting the "+" by "−" and $\downarrow$ by $\uparrow$ in the notation.

The proof is in section B.3.1 in the appendix, though the nature of it was already discussed in the beginning of this section. The following definitions concern the estimation of the variance $\mathcal{V}_n$. Define the operator

$$P_t^+ = e_t^T (\mathrm{X}^T D^+ \mathrm{X})^{-1} \mathrm{X}^T D^+.$$

Then, observe that $\hat{\tau}(x_0)^{\downarrow} = P_1^+(Y - Z\hat{\gamma}^+)$. Whenever x_0 is an interior point or is the left boundary of $\mathcal{X}$, the quantities $\hat{C}_+$, $\hat{\Sigma}_z(x_0)^{\downarrow}$, $c_{z\epsilon^2}(x_0)^{\downarrow}$ and $\hat{\sigma}^2(x_0)^{\downarrow}$ are defined in equations (7.12)-(7.14) below:

$$\begin{aligned} &\hat{\mathbb{E}}(Z_i \,|\, X_i = x_0)^{\downarrow} = (P_1^+ Z)^T, \\ &\hat{C}_+ := \alpha \, \hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0) - \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0) \, \hat{\mathbb{E}}(Z_i \,|\, X_i = x_0)^{\downarrow}. \end{aligned} \tag{7.12}$$

$$\text{Let } U_{ls} = \begin{bmatrix} Z_1^l Z_1^s \\ \vdots \\ Z_n^l Z_n^s \end{bmatrix}, \quad \hat{E}(Z_i Z_i^T \,|\, X_i = x_0)^{\downarrow} = \begin{bmatrix} P_1^+ U_{11} & \dots & P_1^+ U_{1d} \\ \vdots & & \vdots \\ P_1^+ U_{d1} & \dots & P_1^+ U_{dd} \end{bmatrix}, \text{ then}$$

$$\hat{\Sigma}_z(x_0)^{\downarrow} = \hat{E}(Z_i Z_i^T \,|\, X_i = x_0)^{\downarrow} - \hat{\mathbb{E}}(Z_i \,|\, X_i = x_0)^{\downarrow} \hat{\mathbb{E}}(Z_i^T \,|\, X_i = x_0)^{\downarrow}. \tag{7.13}$$

$$\text{Let } R_z^+ = \begin{bmatrix} (Y_1 - Z_1^T \hat{\gamma}^+)^2 Z_1^T \\ \vdots \\ (Y_n - Z_n^T \hat{\gamma}^+)^2 Z_n^T \end{bmatrix}, \text{ then}$$

$$\begin{aligned} &\hat{\mathbb{E}}(Z_i (Y_i - Z_i \hat{\gamma}^+)^2 \,|\, X_i = x_0)^{\downarrow} = P_1^+ R_z^{1+}, \\ &\hat{c}_{z\epsilon^2}(x_0)^{\downarrow} = \hat{\mathbb{E}}(Z_i (Y_i - Z_i \hat{\gamma}^+)^2 \,|\, X_i = x_0)^{\downarrow} - \hat{\mathbb{E}}(Z_i \,|\, X_i = x_0)^{\downarrow} \hat{\mathbb{E}}((Y_i - Z_i \hat{\gamma}^+)^2 \,|\, X_i = x_0)^{\downarrow}. \end{aligned} \tag{7.14}$$

$$\text{Let } R^+ = \begin{bmatrix} (Y_1 - Z_1^T \hat{\gamma}^+)^2 \\ \vdots \\ (Y_n - Z_n^T \hat{\gamma}^+)^2 \end{bmatrix}, \text{ then}$$

$$\begin{aligned} &\hat{\mathbb{E}}((Y_i - Z_i \hat{\gamma}^+)^2 \,|\, X_i = x_0)^{\downarrow} = P_1^+ R^+ \\ &\hat{\sigma}^2(x_0)^{\downarrow} = \mathbb{E}((Y_i - Z_i^T \hat{\gamma}^+)^2 \,|\, X_i = x_0)^{\downarrow} - (\hat{\tau}(x_0)^{\lim +})^2. \end{aligned} \tag{7.15}$$

Finally, if x_0 is an interior point or is the right boundary of $\mathcal{X}$, then $\hat{C}_-$, $\hat{\Sigma}_z(x_0)^{\uparrow}$, $c_{z\epsilon^2}(x_0)^{\uparrow}$ and $\hat{\sigma}^2(x_0)^{\uparrow}$ are defined analogously, substituting "+" by "-" and "$\downarrow$" by "$\uparrow$" in the notation.

Theorem 7.5. *Under H_0: X_i is exogenous, $\theta = 0$ and $V_A = 0$. If assumptions 6.1, 7.1 and 7.3 hold, then if*

$$\hat{\mathcal{V}}_n = \alpha^2 \left[\hat{\mathcal{V}}_\tau^+ + 2\sqrt{h} \hat{C}_+^T \hat{\mathcal{C}}_{\tau\gamma}^+ + h \hat{C}_+^T \hat{\mathcal{V}}_\gamma^+ \hat{C}_+ \right] + (1-\alpha)^2 \left[\hat{\mathcal{V}}_\tau^- + 2\sqrt{h} \hat{C}_-^T \hat{\mathcal{C}}_{\tau\gamma}^- + h \hat{C}_-^T \hat{\mathcal{V}}_\gamma^- \hat{C}_- \right],$$

where

$$\begin{aligned} \hat{\mathcal{V}}_\tau^+ &= \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)^2 \hat{\mathcal{V}}^+, & \hat{\mathcal{V}}_\tau^- &= \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)^2 \hat{\mathcal{V}}^-, \\ \hat{\mathcal{V}}^+ &= \tfrac{\hat{\sigma}^2(x_0)^{\downarrow}}{\hat{\phi}(x_0)^{\downarrow}} e_1^T \Lambda_0^{-1} \Omega \Lambda_0^{-1} e_1, & \hat{\mathcal{V}}^- &= \tfrac{\hat{\sigma}^2(x_0)^{\uparrow}}{\hat{\phi}(x_0)^{\uparrow}} e_1^T \Lambda_0^{-1} \Omega \Lambda_0^{-1} e_1, \\ \hat{\mathcal{C}}_{\tau\gamma}^+ &= \left(\hat{\Sigma}_z(x_0)^{\downarrow} \right)^{-1} \hat{c}_{z\epsilon^2}(x_0)^{\downarrow} & \hat{\mathcal{C}}_{\tau\gamma}^- &= \left(\hat{\Sigma}_z(x_0)^{\uparrow} \right)^{-1} \hat{c}_{z\epsilon^2}(x_0)^{\uparrow}, \end{aligned}$$

then $\hat{\mathcal{V}}_n - \mathcal{V}_n = o_p(1)$.

The proof is in section B.3.2. The following theorem gives the discontinuity test properties in the partially linear case.

Theorem 7.6. *Let $0 \leqslant \lambda \leqslant 1$, Φ be the standard normal cumulative distribution function, and $c_\lambda = \Phi^{-1}(\lambda)$. If theorems 6.1, 7.4 and 7.5 hold and $\sqrt{nh}h^{p+1} \to 0$ as $n \to \infty$, then under H_0: X is exogenous,*

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} \leqslant c_\lambda\right) \to \lambda \quad as \quad n \to \infty.$$

moreover, if result 1 is true, under H_1: X is endogenous,

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} > c_\lambda\right) \to 1 \quad as \quad n \to \infty,$$

and under the local alternatives $\frac{\theta}{\sqrt{nh}}$,

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} \leqslant c_\lambda\right) \to \Phi\left(c_\lambda - \frac{\theta}{\sqrt{\alpha^2\mathcal{V}_\tau^+ + (1-\alpha)^2\mathcal{V}_\tau^-}}\right) \quad as \quad n \to \infty.$$

The proof is offered in the appendix section B.3.3. Observe that the variance of the estimation of the nonparametric terms τ^+ and τ^+ is the only variance that affects the local power of the test in large samples. This occurs because the other components in $\mathcal{V}_n$ are $o(nh)$, or more specifically, $O(nh^{3/2})$.

Remark 7.2. *The estimation of $\phi(x_0)^\downarrow$ and $\phi(x_0)^\uparrow$ is not a trivial application of the literature of density estimation. When estimating limits of densities at boundary points, the same concerns as with the estimation of conditional expectations at boundary points arise, so $\hat{\phi}(x_0)^\downarrow$ and $\phi(x_0)^\uparrow$ must be chosen mindful of their boundary properties. Although local polynomial estimators have excellent boundary properties, they cannot be naturally transformed for density estimation, as it can be done with kernels. One solution is to estimate $\phi(x_0)^\downarrow$ with boundary kernels, as in Jones (1993). The application section uses a different approach, based on the estimator proposed in Lejeune and Sarda (1992), which consists on the local polynomial regression of the empirical distribution function $\hat{F}(X_i)$ on X_i using only observations such that $X_i > 0$. The coefficient of the constant term is an estimator of $\lim_{x\downarrow x_0} F(x)$, but the coefficient of the linear term is actually an estimator of $\lim_{x\downarrow x_0} \frac{d}{dx}F(x)$, which is exactly $\phi(x_0)^\downarrow$. Hence, in this case*

$$\hat{\phi}(x_0)^\downarrow = e_2^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+ \hat{F} = P_2^+ \hat{F},$$

where $\hat{F} = (\hat{F}_1, \ldots, \hat{F}_n)^T$, $\hat{F}_j = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \leqslant X_j)$. Analogously for $\hat{\phi}(x_0)^\uparrow$.

7.3 The nonparametric case

Let the conditional expectation be represented by the function f, so that

$$f_Y(x, z) := \mathbb{E}(Y_i \,|\, X_i = x, Z_i = z), \tag{7.16}$$

and define $f_Y(x_0, Z_i)^{\downarrow} := \lim_{x \downarrow x_0} f_Y(x, Z_i)$, $f_Y(x_0, Z_i)^{\uparrow} := \lim_{x \uparrow x_0} f_Y(x, Z_i)$, and suppose that these limits exist for all Z_i.

Example 3. *(Censoring) Equation* (7.16) *can be parameterized inside the censoring model presented in section 6.1. From equation* (6.4)*, observe that for* $X > 0$*,*

$$f_Y(X, Z) = f_1(X, Z, f_2^{-1}(X; Z)).$$

Assumption 6.3 can be modified to serve as a primitive of assumption 7.4 (3).

In this case, $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\downarrow} = f_Y(x_0, Z_i)^{\downarrow}$, and $\mathbb{E}(Y_i \,|\, x_0, Z_i)^{\uparrow} = f_Y(x_0, Z_i)^{\uparrow}$. Hence, $\hat{\mathbb{E}}(Y_i \,|\, x_0, Z_i)^{\downarrow} = \hat{f}(x_0, Z_i)^{\downarrow}$, and $\hat{\mathbb{E}}(Y_i \,|\, x_0, Z_i)^{\uparrow} = \hat{f}(x_0, Z_i)^{\uparrow}$. Define $\hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0) = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^n g(Z_i) \mathbf{1}(X_i = x_0)$ and $\hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0) = \frac{1}{\hat{p}_{x_0}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i = x_0) g(Z_i) Z_i$, then equation (7.3) cannot be simplified as in the previous cases. The present case will assume that the Z_i are random variables which can take a finite number of values. Similar results could be derived when the Z_i can take a countable number of values, and also when the Z_i are continuous or mixed random variables. The decision to present results in the finite case has the advantage of the simplicity, but is also done for practical reasons, as is explained in remark 7.4 below. The following exposition refers to the estimation of $f_Y(x_0, Z_i)^{\downarrow}$, and $f_Y(x_0, Z_i)^{\uparrow}$ is estimated analogously.

Let the $Z_i \in \{z^1, \ldots, z^M\}$, and define the estimator of $f_Y(x_0, z^m)^{\downarrow}$ in the following way. Given the kernel function k, the smoothing parameter h, the polynomial degree p, and let $\hat{a}_0, \ldots, \hat{a}_p$ be the solution to the problem

$$\min_{a_0, \ldots, a_p} \frac{1}{n} \sum_{j=1}^n k\left(\frac{X_j - x_0}{h}\right) \mathbf{1}(X_j > x_0) \left[Y_j - a_0 - a_1(X_j - x_0) - \ldots - a_p(X_j - x_0)^p\right]^2.$$

If x_0 is an interior point or is the left boundary of $\mathcal{X}$, the local polynomial estimator of $f_Y(x_0, z^m)^{\downarrow}$ is given by

$$\hat{f}(x_0, z^m)^{\downarrow} = \hat{a}_0 = e_1^T (\mathrm{X}^T D_m^+ \mathrm{X})^{-1} \mathrm{X}^T D_m^+ \mathrm{Y}, \tag{7.17}$$

where $e_1 = (1, 0, \ldots, 0)^T$ has dimension $1 \times (p+1)$, X has rows equal to $(1, (X_j - x_0), \ldots, (X_j - x_0)^p)$, $j = 1, \ldots, n$, D_m^+ is a $n \times n$ diagonal matrix with diagonal elements $\left\{\mathbf{1}(Z_1 = z^m)\mathbf{1}(X_1 > x_0)\, k\left(\frac{X_1 - x_0}{h}\right), \ldots, \mathbf{1}(Z_n = z^m)\mathbf{1}(X_n > x_0)\, k\left(\frac{X_n - x_0}{h}\right)\right\}$, and $\mathrm{Y} = (Y_1, \ldots, Y_n)^T$. If x_0 is the right boundary of $\mathcal{X}$, then $\hat{f}(x_0, z^m)^{\downarrow} = 0$.

Let $\hat{p}_{x_0}^m := \left(\sum_{i=1}^n \mathbf{1}(X_i = x_0)\right)^{-1} \sum_{i=1}^n \mathbf{1}(Z_i = z^m)\mathbf{1}(X_i = x_0)$ be an estimator of $p_{x_0}^m := \mathbb{P}(Z_i = z^m \,|\, X_i = x_0)$, hence

$$B_n = \alpha \sum_{m=1}^M \hat{p}_{x_0}^m \hat{\Gamma}(z^m)^+ g(z^m) + (1 - \alpha) \sum_{m=1}^M \hat{p}_{x_0}^m \hat{\Gamma}(z^m)^- g(z^m).$$

The next assumption provides conditions that allow the derivation of the asymptotic distribution of B_n.

Assumption 7.4.

1. *$dF(x, z^m) > 0$, for all m and $x \in (x^-, x^+) \cap \mathcal{X}$ (see remark 7.4 below for when this condition fails).*

2. *There exist $x^-, x^+ \in \mathbb{R}$, with $x^- < x_0 < x^+$ such that $\mathbb{P}(X_i \leqslant x\,,\, Z_i = z^m)$ is twice continuously differentiable in X with first derivative bounded away from zero and second derivative uniformly bounded for X in $(x_-\,,\, x_0) \cup (x_0\,,\, x^+)$ and all m. Define $\phi(x_0, z^m)^{\downarrow} := \lim_{x \downarrow x_0} \frac{d}{dx}\mathbb{P}(X_i \leqslant x\,,\, Z_i = z^m)$, $\phi(x_0, z^m)^{\uparrow} := \lim_{x \uparrow x_0} \frac{d}{dx}\mathbb{P}(X_i \leqslant x\,,\, Z_i = z^m)$, $\phi'(x_0, z^m)^{\downarrow} := \lim_{x \downarrow x_0} \frac{d^2}{dx^2}\mathbb{P}(X_i \leqslant x\,,\, Z_i = z^m)$, and $\phi'(x_0, z^m)^{\uparrow} := \lim_{x \uparrow x_0} \frac{d^2}{dx^2}\mathbb{P}(X_i \leqslant x\,,\, Z_i = z^m)$, then all of these quantities exist. Moreover, there exist $\hat{\phi}(x_0, z^m)^{\downarrow}$ and $\hat{\phi}(x_0, z^m)^{\uparrow}$, consistent estimators of $\phi(x_0, z^m)^{\downarrow}$ and $\phi(x_0, z^m)^{\uparrow}$ respectively (see remark 7.3 below).*

3. *The function $f_Y(x, z^m)$ is at least $p+2$ times continuously differentiable in X in $(x^-, x_0) \cap (x_0, x^+)$ for all m. Define $f^{(l)}(x_0)^{\downarrow} := \lim_{x \downarrow x_0} \frac{d^l}{dx^l} f_Y(x, z^m)$, and $f^{(l)}(x_0, z^m)^{\uparrow} := \lim_{x \downarrow x_0} \frac{d^l}{dx^l} f_Y(x, z^m)$, then these quantities exist for $l = 1, \ldots, p+2$ and all m.*

4. *The variances $\sigma^2(x, z^m) := \mathbb{E}(\epsilon_i^2 \mid X_i = x, Z_i = z^m)$ are continuous in $(x_-\,,\, x_0) \cup (x_0\,,\, x^+)$, and the limits $\sigma^2(x_0, z^m)^{\downarrow} := \lim_{x \downarrow x_0} \sigma^2(x, z^m)$ and $\sigma^2(x_0, z^m)^{\uparrow} := \lim_{x \uparrow x_0} \sigma^2(x, z^m)$ exist for all m. Moreover, the moments $\mathbb{E}(|\epsilon_i^2|^{2+\xi_6} \mid X_i = x, Z_i = z^m)$ are uniformly bounded for some $\xi_6 > 0$.*

5. *The kernel k is continuous, symmetric and has bounded support. For all j odd integers, $\int k(u) u^j du = 0$.*

6. *$\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$, and $\lim_{n\to\infty} h^{p+1}\sqrt{nh} < \infty$.*

7. *If x_0 is the left boundary of $\mathcal{X}$, then $\alpha = 1$, and if x_0 is the right boundary of $\mathcal{X}$, then $\alpha = 0$.*

Theorem 7.7. *If assumptions 6.1, 7.1 and 7.4 hold, then*

$$\sqrt{nh}\mathcal{V}_n^{-1/2}(\hat{\theta} - \theta - \mathcal{B}_n) \xrightarrow{d} \mathcal{N}(0, 1)$$

where

$$\mathcal{B}_n = \sum_{m=1}^{M} \hat{p}_{x_0}^m \, g(z^m) \left[\alpha \mathcal{B}_{m,n}^+ + (1-\alpha)\mathcal{B}_{m,n}^-\right]$$

$$\mathcal{B}_{m,n}^+ = \begin{cases} h^{p+1} \frac{f_Y^{+(p+1)}(x_0, z^m)^{\lim}}{(p+1)!} e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+1}), & \text{if } p \text{ is odd,} \\ h^{p+2}\left[\frac{f_Y^{+(p+1)}(x_0, z^m)^{\lim}}{(p+1)!}\frac{\phi'(x_0, z^m)^{\downarrow}}{\phi(x_0, z^m)^{\downarrow}}\right] e_1^T \Lambda_0^{-1}(\Upsilon_{p+2} - \Lambda_1 \Lambda_0 \Upsilon_{p+1}) & \\ \quad + \left[\frac{f_Y^{+(p+2)}(x_0, z^m)^{\lim}}{(p+2)!}\right] e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+2}), & \text{if } p \text{ is even,} \end{cases}$$

and analogously for $\mathcal{B}^-_{m,n}$, just substitute the "+" by "−" in the notation. Finally, Λ_0, Λ_1, Υ_{p+1} and Υ_{p+2} are defined in assumption 7.3 (6).

$$\begin{aligned}
\mathcal{V}_n &= \mathcal{V} + hV_A + o(h), \\
\mathcal{V} &= \sum_{m=1}^{M} \left(p^m_{x_0}\right)^2 g(z^m)^2 \left[\alpha^2 \mathcal{V}^+_m + (1-\alpha)^2 \mathcal{V}^-_m\right], \\
\mathcal{V}^+_m &= \frac{\sigma^2(x_0, z^m)^\downarrow}{\phi(x_0, z^m)^\downarrow} e_1^T \Lambda_0^{-1} \Omega \, \Lambda_0^{-1} e_1,
\end{aligned}$$

and analogously for $\mathcal{V}^-_m$, just substitute the "+" by "−" in the notation. Finally, Ω is defined in assumption 7.3 (6).

The proof is in section B.4.1 in the appendix. Its essence can be understood by observing that, since M is finite, the asymptotic distribution of B_n as defined in equation (7.3) can be trivially derived if the convergence of the $\hat{\Gamma}(z^m)^+$ and $\hat{\Gamma}(z^m)^-$ is ascertained. Since each z^m has positive probability in a neighborhood of x_0, the results in Porter (2003) can be applied with the same modifications as in the partially linear case to account for the fact that x_0 is a mass point.

The estimation of the variance depends on the estimation of $\sigma^2(x_0, z^m)^\downarrow$ and $\sigma^2(x_0, z^m)^\downarrow$. This step requires the estimation of the residuals. Define the operator

$$P^+_{t,m,x} = e_t^T (\mathrm{X}_x^T D^+_{x,m} \mathrm{X}_x)^{-1} \mathrm{X}_x^T D^+_{x,m}.$$

where X_x has rows equal to $(1, (X_i - x), \ldots, (X_i - x)^p)$, and $D^+_{x,m}$ is a diagonal matrix with diagonal elements equal to $\left\{\mathbf{1}(X_1 > x_0\,,\, Z_1 = z^m)\, k\left(\frac{X_1 - x_0}{h}\right), \ldots, \mathbf{1}(X_n > x_0\,,\, Z_n = z^m)\, k\left(\frac{X_n - x_0}{h}\right)\right\}$. Whenever x_0 is an interior point or is the left boundary of $\mathcal{X}$, $\hat{f}(x_0, z^m)^\downarrow = P^s_{1,m,x_0} Y$, and $\hat{\sigma}^2(x_0, z^m)^\downarrow$ is defined in equation (7.18) below. Define

$$\begin{aligned}
\hat{f}^+(X_i, z^m) &= P^+_{1,m,x} \mathrm{Y} \\
\hat{\epsilon}^+_i &= Y_i - \hat{f}^+(X_i, Z_i) \\
R &= ((\hat{\epsilon}^+_1)^2, \ldots, (\hat{\epsilon}^+_n)^2)^T, \\
\hat{\sigma}^2(x_0, z^m)^\downarrow &= P^+_{1,m,x_0} R
\end{aligned} \tag{7.18}$$

and if x_0 is the right boundary of $\mathcal{X}$, $\hat{\sigma}^2(x_0, z^m)^\downarrow = 0$. Analogously for $\hat{\sigma}^2(x_0, z^m)^\uparrow$, substituting "+" by "-" and "↓" by "↑" in the notation.

Assumption 7.5.

1. *The variances $\sigma^2(x, z^m)$ are at least $p+2$ times continuously differentiable in $(x_-\,,\, x_0) \cup (x_0\,,\, x^+)$ for all m, and $\lim_{x \downarrow x_0} d^{\,(l)}\sigma^2(x, z^m)$ and $\lim_{x \downarrow x_0} d^{\,(l)}\sigma^2(x, z^m)$ exist for $l = 1, \ldots, p+2$ and all m.*

2. *The moments* $\mathbb{E}\left((\epsilon_i^2 - \sigma_\epsilon^2(X_i, Z_i))^2 \mid X_i = x, Z_i = z^m\right)$ *are continuous and uniformly bounded in* $(x_-\,,\, x_0) \cup (x_0\,,\, x^+)$*, and the right and left limits when* $x \to x_0$ *exist for all* m*.*

3. $hn^{1/3}(\log n)^{-1/3} \to \infty$

Theorem 7.8. *Suppose assumptions 6.1, 7.1, 7.4 and 7.5 hold. Under* H_0*:* X_i *is exogenous,* $\theta = 0$ *and* $V_A = 0$*. Then*

$$\sqrt{nh}\hat{\mathcal{V}}_n^{-1/2}(\hat{\theta} - \mathcal{B}_n) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\hat{\mathcal{V}}_n = \hat{\mathcal{V}} := \sum_{m=1}^{M} \left(\hat{p}_{x_0}^m\right)^2 g(z^m)^2 \left[\alpha^2 \hat{\mathcal{V}}_m^+ + (1-\alpha)^2 \hat{\mathcal{V}}_m^-\right]$$

with

$$\hat{\mathcal{V}}_m^+ = \frac{\hat{\sigma}^2(x_0, z^m)^\downarrow}{\hat{\phi}(x_0, z^m)^\downarrow} e_1^T \Lambda_0^{-1} \Omega\, \Lambda_0^{-1} e_1, \quad \textit{and} \quad \hat{\mathcal{V}}_m^- = \frac{\hat{\sigma}^2(x_0, z^m)^\uparrow}{\hat{\phi}(x_0, z^m)^\uparrow} e_1^T \Lambda_0^{-1} \Omega\, \Lambda_0^{-1} e_1$$

The proof is in section B.4.2. It relies on Masry (1996)'s result about the uniform convergence of the local polynomial estimator applied to the estimated $\hat{\epsilon}_i^2$. The following theorem gives the discontinuity test properties in the partially linear case.

Theorem 7.9. *Let* $0 \leqslant \lambda \leqslant 1$*,* Φ *be the standard normal cumulative distribution function, and* $c_\lambda = \Phi^{-1}(\lambda)$*. If theorems 6.1, 7.7 and 7.8 hold and* $\sqrt{nh}h^{p+1} \to 0$ *as* $n \to \infty$*, then under* H_0*:* X *is exogenous,*

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} \leqslant c_\lambda\right) \to \lambda \quad as \quad n \to \infty.$$

moreover, if result 1 is true, under H_1*:* X *is endogenous,*

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} > c_\lambda\right) \to 1 \quad as \quad n \to \infty,$$

and under the local alternatives $\frac{\theta}{\sqrt{nh}}$*,*

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} \leqslant c_\lambda\right) \to \Phi\left(c_\lambda - \frac{\theta}{\sqrt{\mathcal{V}}}\right) \quad as \quad n \to \infty.$$

The proof is offered in section B.4.3.

Remark 7.3. *The estimation of* $\phi(x_0, z^m)^{\downarrow}$ *can be done as in the partially linear case (see remark 7.2), following the approach proposed in Lejeune and Sarda (1992). The values* $\hat{F}_m(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leqslant x)\mathbf{1}(Z_i = z^m)$ *are consistent estimators of* $\mathbb{P}(X_i \leqslant x, Z_i = z^m)$*. The approach consists on the local polynomial regression of the function* $\hat{F}_m(X_i)$ *on* X_i *at* x_0*, using only observations such that* $Z_i = z^m$ *and* $X_i > x_0$*. The coefficient of the constant term is an estimator of* $\lim_{x \downarrow x_0} \mathbb{P}(X_i \leqslant x, Z_i = z^m)$*, but the coefficient of the linear term is actually an estimator of* $\lim_{x \downarrow x_0} \frac{d}{dx}\mathbb{P}(X_i \leqslant x, Z_i = z^m)$*, which is exactly* $\phi(x_0, z^m)^{\downarrow}$*. Hence, in this case*

$$\hat{\phi}(x_0, z^m)^{\downarrow} = P^{+}_{2,m,x_0}\hat{F}_m,$$

where $\hat{F}_m = (\hat{F}_m(X_1), \ldots, \hat{F}_m(X_n))^T$*. Analogously for* $\hat{\phi}(x_0)^{\uparrow}$*.*

Remark 7.4. *The measure* ν *in* $\theta = \int G(\Delta(Z), Z)d\nu(z)$ *is chosen by the researcher. The measure chosen for the derivation of the estimators is* $F(z \,|\, X_i = x_0)$*, and from that derives the requirement that if* $\mathbb{P}(Z_i = z^m\,, X_i = x_0) > 0$*, then for estimation purposes it is necessary that* $dF(x, z^m) > 0$*, for all* X *in a neighborhood of* x_0*. All the results can be derived in exactly the same way if the measure chosen is* $F(z \,|\, X_i = x_0, z \in \bar{\mathcal{A}})$*, where* $\bar{\mathcal{A}}$ *is a finite subset of* $\mathcal{A} := \{z\,;\, dF(x, z) > 0,\ \forall x \in (x^-, x^+) \cap \mathcal{X}\}$*, as long as* $\bar{\mathcal{A}}$ *is not empty. Hence,* A_n *and* B_n *in equations* (7.2) *and* (7.3) *are substituted by*

$$A_n = \frac{1}{\hat{p}_{x_0,\bar{\mathcal{A}}}}\frac{1}{n}\sum_{i=1}^{n} \Delta(Z_i)g(Z_i)\mathbf{1}(X_i = x_0, Z_i \in \bar{\mathcal{A}}) - \mathbb{E}(\Delta(Z_i)g(Z_i) \,|\, X_i = x_0, Z_i \in \bar{\mathcal{A}})$$

$$B_n = \frac{1}{\hat{p}_{x_0,\bar{\mathcal{A}}}}\frac{1}{n}\sum_{i=1}^{n} [\alpha\hat{\Gamma}(Z_i)^{+} + (1-\alpha)\hat{\Gamma}(Z_i)^{-}]g(Z_i)\mathbf{1}(X_i = x_0, Z_i \in \bar{\mathcal{A}}).$$

where $\hat{p}_{x_0,\bar{\mathcal{A}}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = x_0, Z_i \in \bar{\mathcal{A}})$*. Assumption 6.1 remains the same, assumption 7.1 remains the same, except for the new definition of* $V_A := \mathbb{V}ar(\Delta(Z_i)g(Z_i \,|\, X_i = x_0, Z_i \in \bar{\mathcal{A}})$*, and assumptions 7.4 and 7.5 remain the same as long as* $\{z^1, \ldots, z^M\} = \bar{\mathcal{A}}$*. The results in theorems 7.7, 7.8 and 7.9 remain unchanged.*

Even when the $dF(Z_i)$ *does not have a finite support, the discontinuity test using a measure* ν *that integrates over a finite subset of the support of the* $dF(Z_i)$ *may be valid. For example, the same approach as above can be used when the support of* $dF(Z_i)$ *is countable, on when it is continuous but there exists a subset of values* Z *such that* $\mathbb{P}(Z_i = z) > 0$*. Define* $\mathcal{A} = \{z\,;\, dF(x, z) > 0,\ \forall x \in (x^-, x^+) \cap \mathcal{X}\,,\ \mathbb{P}(X_i = x_0\,,\ Z_i = z) > 0\,,$ *and* $\mathbb{P}(X_i \in (x^-, x^+)\backslash\{x_0\}\,,\ Z_i = z) > 0\}$*. Define* $\bar{\mathcal{A}} = \{z^1, \ldots, z^m\} \subset \mathcal{A}$*. As long as* $\bar{\mathcal{A}} \neq \emptyset$*, the procedure and the results hold exactly as in the case above.*

Chapter 8

When X is a discrete r.v.

When assumption 6.1 (1) fails, it is no longer possible to test the endogeneity of x from the discontinuity of $\mathbb{E}(Y \mid X = x, Z)$ in x at x_0, because $\lim_{x \downarrow x_0} \mathbb{E}(Y \mid X = x, Z)$ and $\lim_{x \uparrow x_0} \mathbb{E}(Y \mid X = x, Z)$ are not defined. However, it is still possible to test endogeneity if one is willing to make assumptions on the variation of $\mathbb{E}(Y \mid X = x, Z, Q)$ for given changes in x. This test will be undersized.

Assumption 8.1. *There exists a bounded neighborhood $\mathcal{N}$ of x_0 such that $\mathbb{P}(X \in \mathcal{N} \backslash \{x_0\}) > 0$, and for all $x, x' \in \mathcal{N}$, $x \neq x'$, and all Z and Q,*

$$b_L \leqslant \frac{E(Y \mid X = x, Z, Q) - E(Y \mid X = x', Z, Q)}{x - x'} \leqslant b_U.$$

Moreover, the sets $\mathcal{Z}_x$ are identical, $\forall x \in \mathcal{N}$.

Theorem 8.1. *Let $\mathcal{N}^+ := \mathcal{N} \cap (x_0, \infty)$, and suppose that $\mathbb{P}(X \in \mathcal{N}^+) > 0$. Let μ be a known measure in $\mathcal{N}^+$. Define*

$$\theta = \frac{\int \left[\int \mathbb{E}(Y \mid X, Z)\, d\mu(X) - \mathbb{E}(Y \mid X = x_0, Z)\right] d\nu(Z)}{\int X\, d\mu(X) - x_0}.$$

If assumption 8.1 holds, then, if X is exogenous, $b_L \leqslant \theta \leqslant b_U$.

This result is a simple consequence of the definition of the exogeneity of X, as seen in definition 1 in section 6. This condition is valid even when X is a discrete variable.

In order to build a test of endogeneity, suppose that $\mu(X) = F(X \mid X \in \mathcal{N}^+)$, and as in the previous cases, $\nu(Z) = F(Z \mid X = x_0)$. Then

$$\theta = \frac{\mathbb{E}(\mathbb{E}(Y \mid X \in \mathcal{N}^+, Z) \mid X = x_0) - \mathbb{E}(Y \mid X = x_0)}{\mathbb{E}(X \mid X \in \mathcal{N}^+) - x_0},$$

which is a desirable parameter from which to build a test statistic, because it eliminates the need to estimate $\mathbb{E}(Y \mid X = x_0, Z)$. Moreover, θ can be identified even when the support of the distribution of X is not continuous.

There are many approaches that could be taken in order to build a test statistic using θ as the reference parameter. In particular, it is direct to build the test so that the test statistic converges under H_0 at the $\sqrt{n}$ rate, even when $\mathbb{E}(Y \mid X, Z)$ is only nonparametrically indentifiable. This happens because θ aggregates over a positive probability subset of the distribution of X. This section develops a test which assumes that $\mathbb{E}(Y \mid X, Z)$ is linear in X and Z.

Assumption 8.2. *For $X \in \mathcal{N}^+ \backslash \{x_0\}$, $\mathbb{E}(Y \mid X, Z) = \beta X + Z^T \gamma$.*

In this case,

$$\begin{aligned}\theta &= \frac{\mathbb{E}(\beta\, \mathbb{E}(X \mid X \in \mathcal{N}^+) + Z^T\gamma \mid X = x_0) - \beta x_0 - \mathbb{E}(Z \mid X = x_0)^T\gamma}{\mathbb{E}(X \mid X \in \mathcal{N}^+) - x_0} \\ &= \frac{\beta\, \mathbb{E}(X \mid X \in \mathcal{N}^+) + \mathbb{E}(Z \mid X = x_0)^T\gamma - \beta x_0 - \mathbb{E}(Z \mid X = x_0)^T\gamma}{\mathbb{E}(X \mid X \in \mathcal{N}^+) - x_0} \\ &= \beta\end{aligned}$$

The test statistic is hence the method of moments estimator of β, which is the same as the estimator of β in a partitioned OLS regression of Y on X and Z using only observations for which $X \in \mathcal{N}^+$:

$$\hat{\theta} = (\mathrm{X}^T D(I - P_{\mathrm{Z}_+})D\mathrm{X})^{-1}\mathrm{X}^T D(I - P_{\mathrm{Z}_+})D\mathrm{Y}$$

where I is the $n \times n$ identity matrix, $P_{\mathrm{Z}_+} = D\mathrm{Z}(\mathrm{Z}^T D\mathrm{Z})^{-1}\mathrm{Z}^T D$, $\mathrm{X} = (X_1, \ldots, X_n)^T$, Z is a matrix with rows equal to Z_i^T, $i = 1, \ldots, n$, $D = Diag\{\mathbf{1}(X_1 \in \mathcal{N}^+ \backslash \{x_0\}), \ldots$ $\ldots, \mathbf{1}(X_n \in \mathcal{N}^+ \backslash \{x_0\})\}$, and provided both $\mathrm{X}^T D(I - P_{\mathrm{Z}_+})D\mathrm{X}$ and $\mathrm{Z}^T D\mathrm{Z}$ are invertible.

Assumption 8.3. *Suppose*

1. *The (Y_i, X_i, Z_i), $i = 1, \ldots, n$ are i.i.d*
2. *Let $\epsilon_i := Y_i - \mathbb{E}(Y_i \mid X_i, Z_i)$. Then, $\mathbb{V}ar(\epsilon_i \mid X_i, Z_i) = \sigma_\epsilon^2$ for all $X_i \in \mathcal{N}^+ \backslash \{x_0\}$, $i = 1, \ldots, n$*
3. *Let $W_i := (X_i, Z_i^T)^T$, then $\mathbb{E}(W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\}))$ exists and is invertible. Moreover, $\mathbb{E}(Z_i Z_i^T \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\}))$ is also invertible.*

Theorem 8.2. *If assumptions 8.2 and 8.3 hold, then*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V),$$

as $n \to \infty$, where

$$\begin{aligned}V = \sigma_\epsilon^2 \Big[&\mathbb{E}(X_i^2 \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\})) - \\ &- \mathbb{E}(X_i Z_i^T \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\}))\mathbb{E}(Z_i Z_i^T \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\}))^{-1} \cdot \\ &\cdot \mathbb{E}(Z_i X_i \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\})) \Big]^{-1}.\end{aligned}$$

Moreover, let

$$\hat{V} = \hat{\sigma}_\epsilon^2(\mathrm{X}^T D(I - P_{\mathrm{Z}_+})D\mathrm{X})^{-1}$$

with

$$\hat{\sigma}_\epsilon^2 = \left(\sum_{i=1}^{n} \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\})\right)^{-1} \hat{\epsilon}^T \hat{\epsilon}$$

where $\hat{\epsilon} = (I - D\mathrm{W}(\mathrm{W}^T D\mathrm{W})^{-1}\mathrm{W}^T D)Y$ *and* $\mathrm{W} = [\mathrm{X}\ \ \mathrm{Z}]$. *Then,*

$$\hat{V} \xrightarrow{p} V$$

as $n \to \infty$

The proof of the theorem is similar to the proofs of the convergence of the classical partitioned OLS estimator, and is given in appendix B.5.1.

Let

$$H_0\text{: } X \text{ is exogenous,}$$

and

$$H_1\text{: } X \text{ is endogenous,}$$

the discontinuity test of endogeneity will reject H_0 if $\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}$ or $\hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}$, for a positive scalar $c_{\lambda/2}$ which will be discussed soon. In order to establish the size and power against H_1 of this test, consider the following theorem

Theorem 8.3. *Suppose that assumptions 8.1, 8.2 and 8.3 hold. Then, under* H_0, *if* $\lambda_L < \lambda_U$,

$$\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}} \text{ or } \hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) \approx$$

$$\approx \Phi\left(\sqrt{n}\left(\frac{b_L - \theta}{\sqrt{V}}\right) - c_{\lambda/2}\right) + 1 - \Phi\left(\sqrt{n}\left(\frac{b_U - \theta}{\sqrt{V}}\right) + c_{\lambda/2}\right) \to 0$$

as $n \to \infty$. *Under* H_1, *if* $\lambda_L < \beta < \lambda_U$, *then*

$$\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}} \text{ or } \hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) \to 0$$

as $n \to \infty$. *If* $\beta = \lambda_L$ *or* $\beta = \lambda_U$, *then*

$$\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}} \text{ or } \hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) \to \lambda/2$$

as $n \to \infty$. *Finally, if* $\beta < \lambda_L$ *or* $\beta > \lambda_U$, *then*

$$\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}} \text{ or } \hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) \to 1$$

as $n \to \infty$

If $c_{\lambda/2} = \Phi^{-1}(\lambda/2)$ is the critical value associated with the confidence level $1-\lambda/2$ in the normal distribution, theorem 8.3 shows that the test is undersized, in the sense that under H_0, it rejects H_0 with probability smaller than λ. The test has non-trivial power agains H_1 if $\beta < \lambda_L$ or $\beta > \lambda_U$. The proof of this theorem is offered in the appendix section B.5.2.

Chapter 9

Applicability of the discontinuity test

The main assumption that guarantees that the discontinuity test is well defined is that $\mathbb{E}(Y \mid X, Z, Q)$ is continuous in X. Continuity is a phenomenon often argued in the literature, so this paper will not be concerned with when this assumption is valid. The non trivial power of this test depends on further assumptions, and the sufficient condition presented in chapter 6 requires that $dF(Q \mid X, Z)$ be discontinuous in X at $X = x_0$. Since this is not a condition commonly seen in the literature, this section will discuss when it can be encountered, and how it can be argued.

For the examples that the author has considered, the common characteristic is when X is a variable which presents a non-random concentration of observations at a known value $X = x_0$, or to one side of x_0. Depending on the reason for the concentration, it may be possible to infer that the observations at x_0 are distributed discontinuously different from the observations nearby.

A natural cause of concentration is censoring, i.e. when X cannot assume values above or below a certain value x_0. The example examined in part III is one such case. Suppose that X denotes the amount smoked per day. Since one can smoke any fraction of a cigarette, X is a continuous r.v. which cannot assume negative values. In the dataset discussed in chapter 13, more than 80% of the observations are concentrated at $x = 0$ (see appendix section C), so it is natural to suppose that the distribution of the characteristics of these observations is discontinuously different at $x = 0$ compared to those for which $x > 0$.

Though it is not possible to show that $dF(Q \mid X, Z)$ is discontinuous in X at $X = x_0$, it is possible to show that $dF(Z_d \mid X, Z_{-d})$ is discontinuous in X at $X = x_0$, where Z^d is one of the covariates, and Z_{-d} is the vector of the remaining covariates. A formal test of this condition could be, for example, to test for the discontinuity of the mean of the distribution of Z_d. This can be done by performing the discontinuity test of endogeneity, but substituting $\mathbb{E}(Z_d \mid X, Z_{-d})$ instead of $\mathbb{E}(Y \mid X, Z)$. Chapter 12 in part III does not perform such a formal test, but shows heuristic evidence that

$\mathbb{E}(Z_d \mid X)$ is discontinuous at $X = 0$.

If the data plots in chapter 12 are accepted as evidence of the discontinuity of many of the observable covariates in the structural function, it may be acceptable to infer that the same phenomenon happens with the unobservable Q. This line of argument is similar to the one used, for example, in the regression discontinuity literature to show that the observations cannot influence in which side of the threshold they are located. It is common practice in those cases to show that the observations seem to have the same frequencies at both sides of the threshold (See McCrary (2008) for a formal frequency test in the regression discontinuity context), and are similar with regard to the observable covariates (see for example Lee (2008)). A comparable approach is also adopted in the experimental literature to argue that the unobservables are equally distributed among the treatment and control groups: by showing that the two groups have the same distribution with regards to the observable characteristics, it may be acceptable to assume that the same happens with the unobservable characteristics as well (see for example LaLonde (1986)).

Censoring examples are commonly observed when the potentially endogenous variable X is a consumption good, which cannot be chosen in negative amounts, as is the case when X is the daily number of cigarettes smoked. In that case, the argument is that the observations at zero are discontinuously different from the observations at positive amounts. The discontinuity may exist because among everyone who chose zero there are not only those who would have optimally chosen zero in an unconstrained problem (who could indeed be similar to those who chose immediately positive amounts), but also those whose would have chosen negative amounts if they could (which can presumably be very different from those who chose immediately positive amounts). Other examples where censoring is naturally generated are $X =$ commute time, or $X =$ hours of work. Censoring can be artificially generated, for example by law imposed restrictions, such as minimum age required to drop out of school when $X =$ years of education, or minimum wage when $X =$ hourly wage.

Censored variables are just one example where selected concentration happens. Another instance of this phenomenon is when the potentially endogenous variable is a choice variable for which default values are specified. Standard contracts is one such case. For example, if $X =$ contribution rate on the 401K, the observations at the standard level may be discontinuously different from the observations at the tailored levels near the standard level (see Madrian and Shea (2001)). Another case is when there are social norms that stipulate that one choice is the default. For example, if $X =$ division of bequest among progeny, the social norm clearly dictates an equal division of the be quest among the progeny (see Wilhelm (1996)). Finally, there are instances where concentration happens because of some discontinuity in the incentives for people to make different choices. For instance, Saez (1999) shows evidence that taxpayers seem to concentrate in the first kink point of the US income tax schedule (i.e. where marginal rates jump from 0% to 15%).

Chapter 10

Conclusion of Part II

This part developed a nonparametric test of endogeneity which does not require instrumental variables. The two crucial assumptions for the applicability of the test are that the distribution of the unobservable conditional on the observable variables be discontinuous in the potentially endogenous variable at a known point, and that the structural equation relating the dependent variable and the observables be continuous in the running variable. The test consists in estimating the size of such discontinuities, averaging them over a given distribution of the covariates, and then testing for whether this average is equal to zero.

Part II provides test statistics and asymptotic distributions for the average of the discontinuities multiplied by arbitrary functions of the covariates, averaged over the distribution of the covariates at the threshold. This type of test eliminates the need of nonparametric estimation of one of the components of the parameter which is the basis of the test statistic. The estimation of the discontinuities is done for three different specifications of the conditional expectation of the dependent variable given the covariates when the running variable is different than the threshold. The first assumes that it is linear, the second that it is partially linear (nonparametric in the running variable and additively linear in the covariates), and the third that it is fully nonparametric, although with certain smoothness conditions. The test statistic is shown to converge at $\sqrt{n}$ rates, where n is the sample size, in the linear case, and at the rate $\sqrt{nh}$ in the partially linear and nonparametric cases. This rate is the same as that of a nonparametric regression with a single right-hand side variable, and this is achieved in spite of the presence of the covariates due to the aggregation over the measure of the covariates.

The estimation has to be sensible to the boundary nature of the threshold, even when it is not in fact a boundary point in the domain of the running variable. This is the case because this dissertation allows for the functional forms, as well as conditional distribution functions, variances etc., to be different at the right and left sides of the threshold. Hence, the threshold is treated as a boundary point in all cases, and estimation has to be mindful of boundary biases. The nonparametric estimators

use the local polynomial method, known for its automatic boundary carpentry and low bias. In that regard this dissertation is in accordance with the regression discontinuity literature, which uses the same kind of estimator. However that literature also assumes that the probability densities are continuous across the threshold, while this part allows for differences at the different sides of the threshold, which requires different estimators for the variance of the estimator.

Part II also presents a test of endogeneity developed for cases where X is a discrete variable. If it is reasonable to assume that the function f has bounded variation, and an upper and lower bound of such variation are known, then it is possible to build an undersized test of endogeneity. The test presented assumed that the expectation of Y conditional on X and Z is linear, and provided the test statistic and asymptotic distribution of the test based on that assumption.

The final chapter of part II discussed the applicability of the discontinuity test, particularly in what concerns the power of the test. The chapter considered in which situations it is reasonable to argue that the distribution of the unobservable Q conditional on X and Z is discontinuous in X at a known point x_0. The driving characteristic of the many examples is when x_0 is a mass-point.

Part III

The effects of maternal smoking in birth weight

Chapter 11

Introduction

The effects of smoking during pregnancy, known as "maternal smoking," on the birth weight of the child is an important topic of research in the medical literature for two main reasons. First, birth weight is seen as the primary measure of the newborn's health and as an excellent predictor of infant's survival and development (see Almond et al. (2005) p. 1032). Second, early studies about the effects of smoking in birth weight claimed impressive effects of the order of 500 grams (see Sexton and Hebel (1984)).

Let the variable CIG represent the average cigarettes smoked per day by the mother during pregnancy, let BW be the weight of the child at birth, and let Z represent a set of d covariates. These include detailed information about the mother, the father and the pregnancy. The interest is to uncover the causal relation between CIG and BW, which is expressed in the model

$$BW = m(CIG, Z, Q) + \varepsilon. \tag{11.1}$$

This relation is identifiable in the sense of theorem 6.1 in chapter if

$$\mathbb{P}\big(m(CIG, Z, Q) = m(CIG, Z, 0)\big) = 1$$

Otherwise, further measures must be taken to account for the presence of Q, such as searching for more complete datasets where hopefully Q can be observed, searching for instrumental variables, proxy variables etc.

The effect of maternal smoking on birth weight is an example where experiments that randomly and directly change the quantities smoked by the mothers cannot be generated for ethical reasons. Randomized trials in the field try to influence the amounts smoked indirectly through some kind of "propaganda" directed to a randomly selected part of the sample. The word "propaganda" will be used here to denote the set of smoking-related interventions that were randomly provided in such studies, such as informational phone calls, house visits, etc. A case can be made in favor of the reduced form effect on birth weight, which consists of the true parameter

of interest not being the effect of smoking on birth weight, but rather the effect of the smoking-related interventions on birth weight. However, propaganda can be of many different kinds, and may have radically different effects in different parts of the population depending on its content, its way of transmission and its scope. In this case, the effect of one kind of propaganda on birth weight is not necessarily a good predictor of the effects of other kinds of propaganda, and therefore the effects of smoking may constitute a better source of information to use to extrapolate between different options of public policy. Additionally, smoking rates can be affected not only through public policy, but also through medical recommendations. These reinforce even further the importance of knowing the effect of smoking versus the effect of smoke-related interventions. Moreover, studies that use propaganda to influence smoking behavior may also affect birth weight through other means, for example by providing information or raising health concerns that can make *all* pregnant women (including those who did not quit smoking) change other behaviors. If the actual direct effect of smoking on birth weight is small, although the estimated reduced form effect is large, then policy and medical attention directed at smoking may have comparatively less effect than the same resources directed at changing other habits, for example promoting propaganda for pregnant women to stop drinking, to eat better or to have more frequent prenatal visits.

According to the Cochrane Review (see Lumley et al. (2009)), the smoking cessation interventions in randomized trials had on average a significant but imprecisely estimated effect on birth weight. On average 6 out of 100 mothers quit smoking because of the intervention, and the average effect of the intervention on birth weight is 55 grams, with a 95% confidence interval between 10 grams and 90 grams. This implies that the effect of smoking cessation on birth weight is around 915 grams, with a 95% confidence interval between 167 grams and 1500 grams. Sexton and Hebel (1984), one of the most well-known among such studies, shows a great effect in smoking cessation (20% people quit smoking because of the intervention) and a reduced-form effect of 93 grams, with a 95% confidence interval between 15 and 170 grams, implying an effect of smoking cessation between 77 and 845 grams. These imprecise estimates present an even more ambiguous picture with regard to the relative importance of smoking cessation and smoke-related interventions. They seem to be mostly due to small samples. The Cochrane Review (Lumley et al. (2009)), a systematic review of the field, analyzes 72 randomized trials. These amount to a total sample size of just over 25,000 observations, with an average of around 350 observations per study.

Due to the difficulties associated with the randomized trials mentioned above, the literature in the field has focused on non-experimental data sources where large samples and a wide array of control variables are observed. All these studies rely on an assumption of selection on observables. Therefore, a test of endogeneity without instruments is more than a convenience, it is a necessity not only because it can help in the detection of endogeneity, but it can also contribute to validate a certain choice

of covariates over another.

Almond et al. (2005) is, to the author's knowledge, the most exhaustive analysis of this question using non-experimental data. They provide a detailed analysis of the costs of LBW using two independent empirical approaches, each employing a different source of variation on birth weight. The first approach uses variation of birth weight across twins in order to control for determinants of birth weight that are constant within a family, such as maternal smoking and gestation period. The second approach, of interest to this application, uses only singleton births and explores variation on birth weight due to variation of maternal smoking across families. The authors use both OLS regressions for a number of different specifications and subclassification on the propensity score to control for potential endogeneity due to family unobservables (see p. 1064 in Almond et al. (2005) for details of their method). More specifically, they estimate the difference in the birth weight and the probability of LBW between women who smoked and women who did not smoke during pregnancy, using the population of births of singletons from Pennsylvania from 1989 to 1991. They control for a rich set of covariates. Almond et al. (2005) compare nonsmoking directly with smoking mothers, disregarding the actual quantities smoked. They find that the children born of smoking mothers weigh 200 grams less than those of nonsmoking mothers, with a 95% confidence interval between 199 and 207 grams. For the case of LBW, they found that children of smoking mothers are 3.5% more likely to be of LBW than those of nonsmoking mothers, with a 95% confidence interval between 3.3% and 3.7%.

This part applies the discontinuity test to the full specification in Almond et al. (2005), using the same data set as in that paper. This part is divided in the following chapters. Chapter 12 develops the argument that the discontinuity test of endogeneity of the null hypothesis

$$H_0\text{: } CIG \text{ is exogenous in the structural function } m$$

against the alternative hypothesis

$$H_1\text{: } CIG \text{ is endogenous in the structural function } m$$

has power against H_1. For this it is necessary to argue that the conditional distribution of the unobservables Q that are dependent of CIG is discontinuous in CIG at $CIG = 0$.

Chapter 13 explains the methodology of the test. It includes a description of the dataset and all the techniques used to perform the discontinuity test in this problem. Chapter14 discusses the results of the test.

Chapter 12

Applicability of the discontinuity test to CIG in equation (11.1)

The argument for the applicability of the discontinuity test in the case of the effects of maternal smoking in birth weight depends on two crucial assumptions. The first is that the direct effect of smoking on birth weight may be viewed as continuous (assumption 6.1 (2)).

The second crucial assumption is that any unobservable variable correlated with CIG conditional on z has a distribution conditional on CIG and Z that may be viewed as discontinuous in CIG at a certain value of CIG (assumption 6.2). Since CIG cannot be negative, a candidate to be such a threshold is $CIG = 0$. In more empirical terms, the requirement is that the mothers that did not smoke during pregnancy have to be discontinuously different with regard to the unobservable variable from the mothers that smoked, even conditional on the covariates Z.

Though this cannot be confirmed for the unobservable variables Q, this phenomenon can be tested for the observable covariates Z. This heuristic evidence is analogous to the evidence provided in the applied Regression Discontinuity literature that covariates are continuous at the threshold, suggesting that unobservables are also continuous at the threshold. See Lee (2008) for an example. The test would be essentially the same as the discontinuity test, except that it would be performed with a variable z^s, $s = 1, \ldots, d$, instead of with the dependent variable Y, and by using the rest of the covariates as controls. No matter which continuous function of z^s is tested in the discontinuity test, if the distribution of z^s conditional on CIG and the rest of the covariates is continuous, so should the function be. If some of the z^s were found to be discontinuous at $CIG = 0$, this would be understood as evidence that an unobservable correlated with cigarettes is also discontinuous at $CIG = 0$. This is true unless z^s is a proxy for such a variable, in which case there is no identification issue in the first place.

The following figures provide heuristic evidence that the expectation of the observable covariates conditional on CIG is discontinuous at $CIG = 0$. It is not possible to

guarantee from these figures that any of the variables below would still be discontinuous conditional on the rest of the covariates. However, if the figures below are taken as evidence of discontinuities in the expectation of some of the z^s conditional on CIG, then at least one of the z^s has to be discontinuous in CIG conditional on the rest of the covariates. The figures were trimmed at $CIG = 40$, although the observed CIG goes up to 98. However, $CIG = 41$ to 98 account for only 0.05% of the full sample, and 0.2% of the mothers that smoked positive amounts. Table C.1 in appendix C shows the number of observations for each level of CIG . The dots correspond to the averages per CIG level and the lines show the 95% confidence interval of the mean per CIG level for low levels of CIG only. The confidence intervals are shown only

Figure 12.1: Mother's education (years)

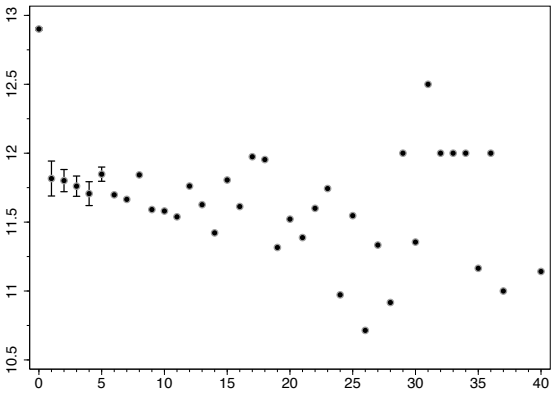


Figure 12.2: Mother's age

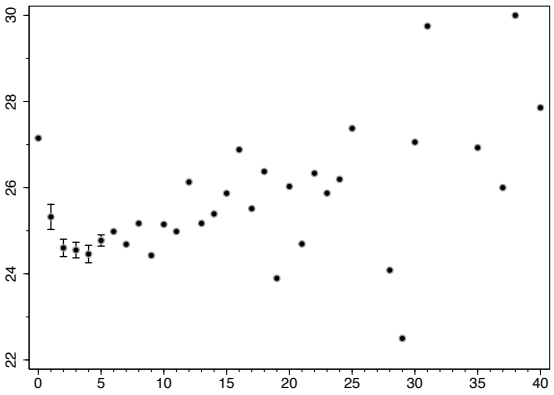


Figure 12.3: Mother is not married

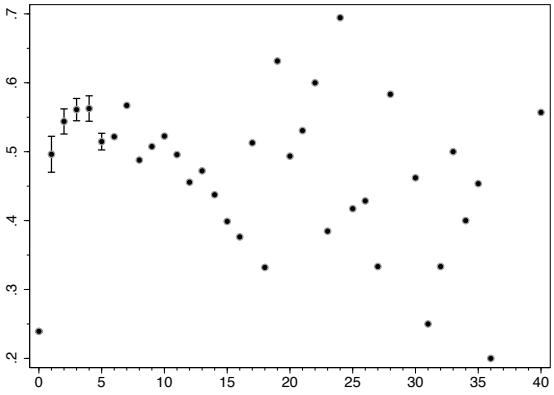
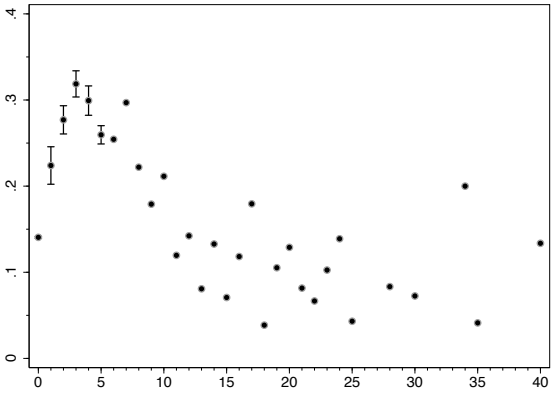


Figure 12.4: Mother is black



Figures 12.1 to 12.4: Horizontal axes represent average number of cigarettes smoked daily during pregnancy. Vertical axes represent: figure 12.1: mother's years of schooling, figure 12.2: mother's age at child's birth, figure 12.3: likelihood of mother not being married and figure 12.4: likelihood of mother being black. Dots represent average values referring to the pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean.

for small levels of cigarettes for exposition reasons. The confidence intervals become larger for $CIG > 20$, which correspond to only 1% of the full sample (6% of the sample of smokers). As suggested by table C.1 of appendix C, the confidence intervals increase as a reflection of the smaller sample sizes per CIG value. For values of CIG between 26 and 29, 31 and 34 and 36 and 39, sample size is extremely small per level of CIG , rarely above 10 and never above 15 observations.

Figures 12.1, 12.2, 12.3 and 12.4 are examples of covariates referring to the mother's demographic characteristics where there is a clear difference in the averages per level of CIG for zero versus just above zero cigarettes. Figure 12.1 shows the mother's education in years, with fairly constant averages of a little below 12 years for $0 < CIG < 8$, and increasing one full year of education for $CIG = 0$. Figure 12.2 shows the mother's age, which averages around 25 years old for low-level smoking mothers, and increases to 27 among the nonsmoking mothers. The marital status shifts from 50% of unmarried low-level smoking mothers to only 24% of unmarried nonsmoking mothers. The proportion of black women among the women surveyed has higher variation, but is constantly above 20%, and often closer to 30% for low-level smokers, and is only 14% for the nonsmokers.

The father's demographic characteristics present even higher differences for low-level smoking mothers relative to nonsmoking mothers. The education level, shown in figure 12.5, changes from below 11 years among the fathers of children of low-

Figure 12.5: Father's education (years)

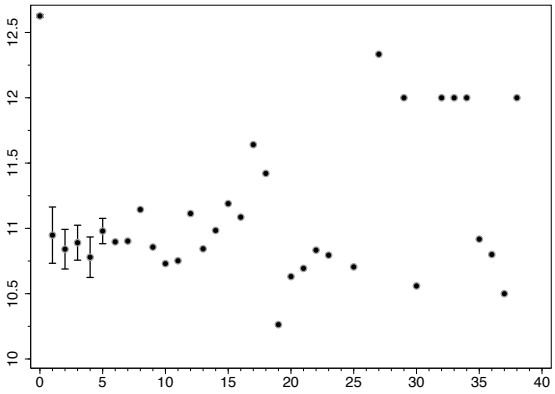
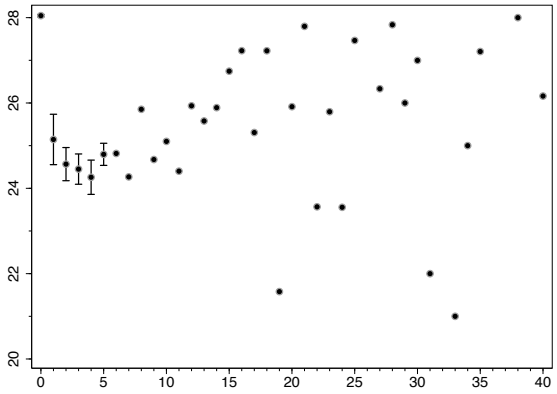


Figure 12.6: Father's age



Figures 12.5 to 12.6: Horizontal axes represent average number of cigarettes smoked daily by the mother during pregnancy. Vertical axes represent: figure 12.5: father's years of schooling, figure 12.6: father's age at child's birth. Dots represent average values referring to the pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean.

level smoking mothers to 12.6 years for fathers of children of nonsmoking mothers, increasing more than 1.5 years of education. Figure 12.6 shows that the average age

of the father is at most 25 years old for $0 < CIG < 10$, but increases to an average of 28 years of age among the fathers of children of nonsmoking mothers.

The behavioral characteristics of the mother also seem to change significantly when comparing low-level smoking mothers to nonsmoking mothers. Around 10% of the mothers consumed alcohol during pregnancy for all smoking levels until CIG=20, while only 2% of the nonsmoking mothers did the same. Low-level smoking mothers on average visited doctors for prenatal visits around 10 times, which is one less time than in the case of nonsmoking mothers.

Although the behavior of mothers seem to be discontinuously different at zero cigarettes, some contingencies that may have an influence on mother's behavior during pregnancy were not found to be discontinuous at zero cigarettes, such as the gender

Figure 12.7: Mother consumed alcohol

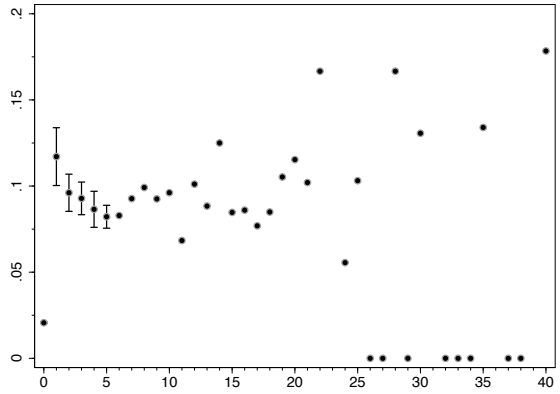


Figure 12.8: Number of prenatal visits

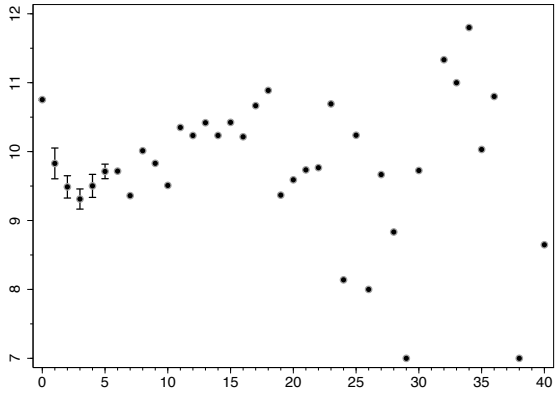


Figure 12.9: Gender of Newborn

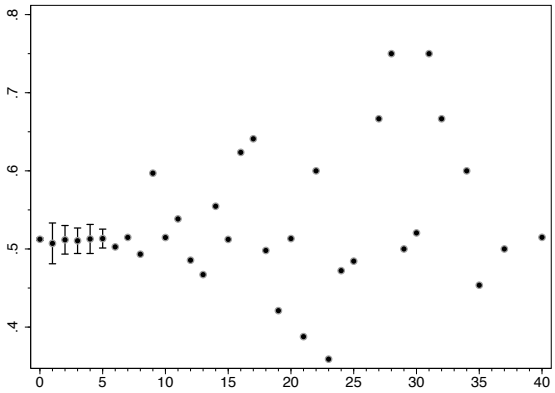
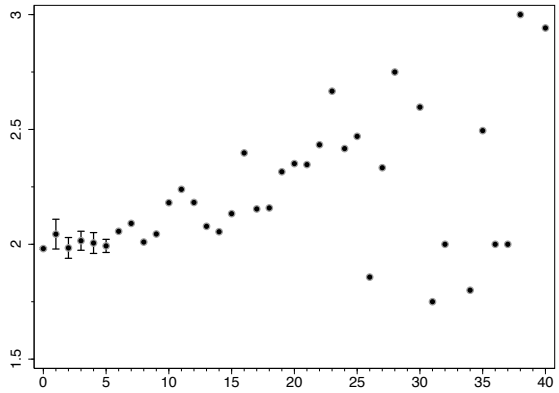


Figure 12.10: Order of Newborn



Figures 12.7 to 12.10: Horizontal axes represent average number of cigarettes smoked daily by the mother during pregnancy. Vertical axes represent: figure 12.7: likelihood of mother consuming alcohol during pregnancy, figure 12.8: number of prenatal visits, figure 12.9: likelihood of child being male and figure 12.10: order of the newborn among live births. Dots represent average values among pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean.

of the newborn (constant around 50% of males) and the birth order of the newborn (constant around the average of second birth).

Chapter 13

Methodology

This chapter describes the methodology used to test endogeneity in the problem of the estimation of the effects on maternal smoking in birth weight. The objective is to produce results that are comparable with Almond et al. (2005). However, their approach assumes that smoking is a binary variable. This part generalizes their specification, and assumes that smoking is a continuous variable.

The data set used is the same as in Almond et al. (2005). It is the annual, linked birth and infant death micro data produced by the National Center for Health Statistics (NCHS). This rich data set contains information for every newborn in Pennsylvania between 1989 and 1991 (488,144 observations, 94,205 smokers) such as mother's and father's demographic characteristics, mother's behaviors during pregnancy, mother's health history and risk factors, sex of the newborn, birth order of the newborn and whether the newborn was part of a multiple birth (i.e., whether the newborn is a singleton). The data also contains relevant information such as mother's and father's age, level of education and race, mother's marital status, foreign born status, number of previous live births and number of previous births where the newborn died. Other information includes maternal risk factors that are believed not to be affected by pregnancy smoking such as chronic hypertension, cardiac disease, lung disease and diabetes. Finally, the data has information related to maternal behavior such as number and timing of prenatal visits, whether the mother drinks and with what frequency, and number of cigarettes smoked per day (for a full list of the variables used see note 36 of Almond et al. (2005) in p.1064).

For the implementation of the test, it is assumed that for $CIG > 0$,

$$\mathbb{E}(BW \,|\, CIG, Z) = \tau(CIG) + Z^T\gamma.$$

This is equivalent to equation (7.7) with $Y = BW$ and $X = CIG$. The superscripts "+" omitted, since in this application $x_0 = 0$ is the left boundary point. Though this specification is not as flexible as a fully nonparametric approach, it allows the use of a high number of control variables, and therefore the immediate comparison with the most complete specification of Almond et al. (2005).

The discontinuity test developed in section 7.2 was performed for the variable CIG in the structural function of two outcome variables (denoted in the previous chapters as Y): birth weight (BW) and probability of low birth weight (LBW), defined as weight below 2500 grams. However, for simplicity the notation of the outcome variable in the rest of this section will remain BW. The test was performed for the dependent variables without the covariates, denoted specification I, and for the most complete specification provided in Almond et al. (2005) (footnote 36, p. 1064), denoted specification II.

The test statistic is calculated as in section 7.2. The only step in the construction of the test statistics which is not specified in chapter 7.2 is the estimator of $\mathbb{E}(BW_i \,|\, CIG_i)$ and $\mathbb{E}(Z_i \,|\, CIG_i)$ (used in the estimation of γ), though some requirements about this estimator are made in assumption 7.3 (2). Most kernel based estimators such as the Nadaraya-Watson or the local polynomial, as well as series estimators satisfy these requirements under roughly the same conditions, but the kernel-based techniques require one regression per different value of CIG_i in the sample, while the series estimators requires only one regression for the estimation of all the values required in the estimation of γ. A series estimator was therefore preferred over the kernel-based for practical reasons. The basis chosen is that of cubic B-splines, which have better local properties than classic global bases such as Fourier or power series (see p. 446 in Li and Racine (2007)). The knots of the spline basis were chosen to be $0.5, 3.5, 6.5, 9.5, 12.5, 17.5, 27.5, 37.5, 47.5, \ldots, 107.5$. Many other combinations of knots were attempted with virtually identical results.

Let the $\rho_j(CIG_i)$ represent the j-th element in the basis evaluated at CIG_i , and let ρ be the matrix whose rows are $(\rho_1(CIG_i)\mathbf{1}(CIG_i > 0), \ldots, \rho_N(CIG_i)\mathbf{1}(CIG_i > 0))$, where N is the number of elements of the basis used in the regression. Let $P_\rho^+ = \rho(\rho^T\rho)^{-1}\rho^T$ and I^+ be the $n \times n$ diagonal matrix with $\{\mathbf{1}(CIG_1 > 0), \ldots, \mathbf{1}(CIG_n > 0)\}$ in the diagonal. The estimator of the variance matrix of $\hat{\gamma}$ is given by

$$\hat{\mathcal{V}}_\gamma = n^+ (Z^T(I^+ - P_\rho^+)Z)^{-1} Z^T(I^+ - P_\rho^+)\hat{\Sigma}(I^+ - P_\rho^+)Z(Z^T(I^+ - P_\rho^+)Z)^{-1},$$

where $n^+ = \sum_{i=1}^n \mathbf{1}(CIG_i > 0)$. $\hat{\mathcal{V}}_\gamma$ is the Eicker-White covariance matrix (see White (1980)) of an OLS regression of $(I^+ - P_\rho^+)Y$ on $(I^+ - P_\rho^+)Z$. This can be useful if the researcher intends to estimate the standard errors using theorem 7.5 (see Li (2000) for the asymptotic behavior of the estimator of the parametric term in the partially linear model using series plugins). This work reports standard errors acquired instead by a bootstrap approach, which will be described later. An important restriction on the covariate list is that Z does not contain a constant term, because as can be seen in equation (7.9), in the partially linear models the constant term cannot be identified separately from $\tau(CIG)$.

For the local polynomial step, as described in equation (7.11), the choice parameters are the kernel, the degree of the polynomial, and the bandwidth size. The kernel used is Epanechnikov (rectangular and triangular kernels were also tested with virtu-

ally identical results) and the polynomial degree is 3, although degrees 2 and 1 were also tested with very similar results.

The bandwidth was chosen by a cross-validation technique (p. 15 in Li and Racine (2007)), which consisted in the estimation of $\tau(CIG)$ for $CIG = 1, \ldots, 20$ by a local polynomial regression of $BW_i - Z_i^T\hat{\gamma}$ using only observations for which $CIG_i > 0$, and $CIG_i \neq CIG$, for each bandwidth $h = 2, 3, ..., 20$, which yielded the values $\hat{\tau}_h(CIG)$, $h = 1, \ldots, 20$, $CIG = 1, \ldots, 20$. The chosen h^* is the one that satisfies

$$h^* = arg \min_{h=1,\ldots,20} \sum_{i=1}^{n} \left(BW_i - Z_i^T\hat{\gamma} - \hat{\tau}_h(CIG_i)\right)^2 \mathbf{1}(0 < CIG_i \leqslant 20).$$

The bandwidth that performed the best was $h = 2$, corresponding to roughly 1.5% of the observations such that $CIC > 0$, followed by $h = 3$, $h = 10$, $h = 11$ and $h = 6$, corresponding to 5%, 26%, 60% and 19% of the observations such that $CIG > 0$ respectively.

As stated previously, the standard errors were estimated by a bootstrap approach, which consisted in drawing 500 bootstrap samples of the data, and calculating $\hat{\theta}$ for each of those independently, exactly in the same way described above. The resulting standard deviations of the 500 values of $\hat{\theta}$ are the standard errors reported in chapter 14.

Chapter 14

Results

Tables 14.1 and 14.2 show the discontinuity test results for the birth weight and the probability of LBW equations.

Discontinuity Test Results

Table 14.1: **Birth Weight** Table 14.2: $\mathbb{P}$**(Birth Weight<2500g)**

C.V.			**I**	**II**	**C.V.**			**I**	**II**
h=2 (1.5%)	1	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	196** (14)	121** (14)	**h=2 (1.5%)**	1	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	-0.043** (0.007)	-0.016* (0.007)
h=3 (5%)	2	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	194** (17)	121** (17)	**h=3 (5%)**	2	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	-0.043** (0.008)	-0.016* (0.008)
h=6 (19%)	5	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	199** (52)	145* (62)	**h=6 (19%)**	5	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	-0.041* (0.017)	-0.019 (0.027)
h=10 (26%)	3	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	178** (30)	140** (32)	**h=10 (26%)**	3	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	-0.037** (0.014)	-0.023 (0.015)
h=11 (60%)	4	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	176** (25)	122** (25)	**h=11 (60%)**	4	$\hat{\boldsymbol{\theta}}$ ($\mathbb{SE}(\hat{\boldsymbol{\theta}})$)	-0.040** (0.012)	-0.022 (0.013)

Tables 14.1 and 14.2: In the first column, h is the bandwidth, and the percentage in parenthesis is the proportion of the sample of smokers used in the local polynomial regression for each value of the bandwidth. C.V. shows the position of the bandwidth in the cross-validation results. $\hat{\theta}$ is the discontinuity test statistic. The standard errors are the result of a bootstrap of the original sample with 200 repetitions. Specification I has no covariates and II is the same specification used in Almond et al. (2005) and described in the text. "**" means that the discontinuity test rejects at the 99% confidence level, "*" means that the test rejects at the 95% confidence level, but not at the 99% confidence level.

The results in table 14.1 present strong evidence of endogeneity for all specifications of birth weight and for all bandwidths. Table 14.2 indicates only weak evidence of endogeneity for the main specification of the probability of LBW and the preferred bandwidths ($h = 2$ and $h = 3$), and no evidence of endogeneity for larger bandwidths. For $h = 2$ and $h = 3$, specification II is rejected with 95% confidence (p-value 1.96) but not rejected with 99% (p-value 2.56) of confidence, and for all other bandwidths specification II is not rejected even with 90% confidence.

Figures 14.1 and 14.2 depict the main results from tables 14.1 and 14.2, respectively. Figure 14.1 shows the average birth weight for each level of CIG (black dots) and two other marks at zero cigarettes. The hollow dot and the "$\times$" point represent the predicted birth weight at zero cigarettes using specification I and II respectively. It can be seen in figure 14.1 that the covariates of specification II appear to help reduce the discontinuity of actual birth weight and predicted birth weight, but not enough for it to vanish.

Figure 14.2, which depicts the results for the probability of LBW analogously to 14.1, shows that the covariates of specification II help reduce the discontinuity of actual LBW and predicted LBW to a third of its original value. The results for the probability of LBW show that the discontinuity is small, so that if there is endogeneity in specification II, it is of low importance for LBW.

One of the two crucial assumptions of the discontinuity test of endogeneity may not be valid in the case of maternal smoking: namely that the effect of smoking on birth weight or on the probability of LBW is continuous at zero cigarettes. If that is the case, then one cannot disentangle the part of the discontinuity found in the test that is due to the discontinuous treatment effect and the part that is due to the endogeneity. In the results shown in this thesis, the discontinuity effects become smaller when more covariates are added, which may be an indication that at least part of the discontinuities are due to endogeneity. This suggests the necessity of better data sets or of the search for quasi-experimental variations of smoking.

Another possibility that should be considered is that the data on smoking is too roughly distributed. The unit of the variable CIG is "cigarettes per day," and one cigarette a day may be a significant amount of smoking. In this case, what seems to be a large discontinuity in $\mathbb{E}(BW \,|\, CIG, Z)$ may be a disproportionately higher treatment effect when comparing $CIG = 0$ to $CIG = 1$ for the same value of Z. In this case, an approach such as described in chapter 8 may be useful. Table 14.1 shows a discontinuity in birth weight of 121 grams. Interpreting the same result, not as a discontinuity, but as a discrete treatment effect inside of a selection on observables approach, the effect of smoking one cigarette is a decrease of 121 grams in birth weight. However, from Almond et al. (2005), the effect of smoking estimated in that paper is of 200 grams, and smoking pregnant women smoke in average 9 cigarettes

Figure 14.1: Results for birth weight, Specification II

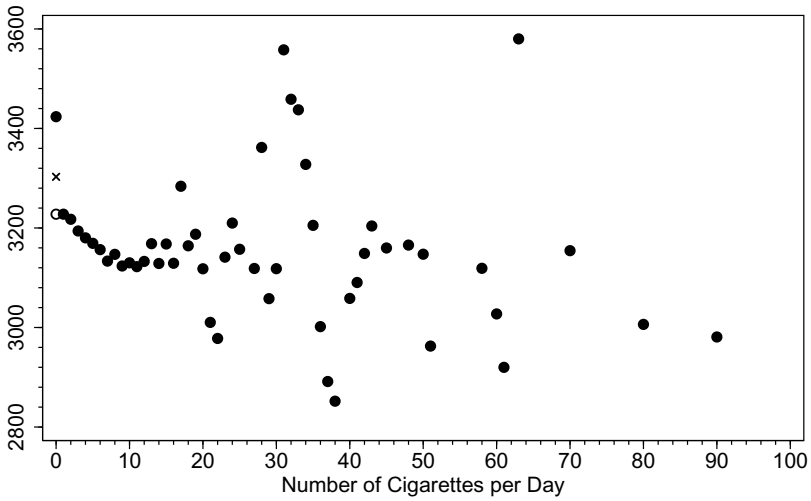
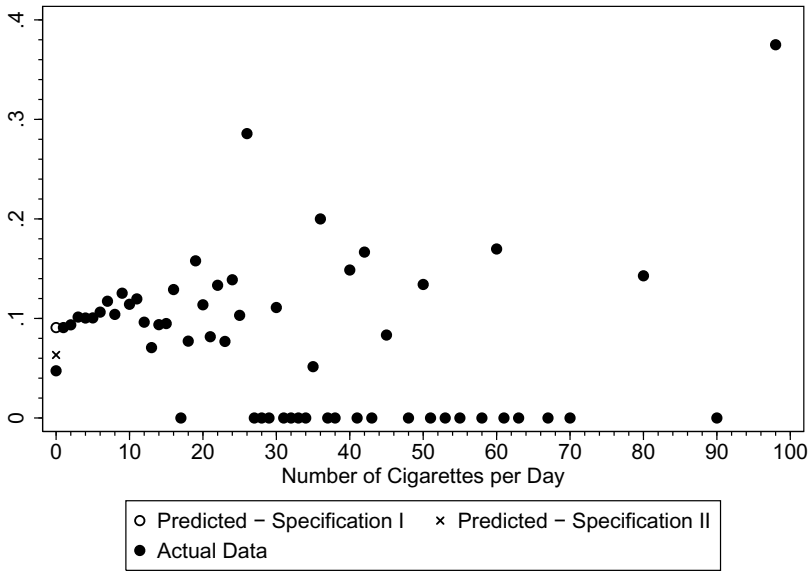


Figure 14.2: Results for the probability of LBW, Specification II



Figures 14.1 and 14.2: vertical lines represent birth weight in grams and likelihood of LBW (birth weight <2500 g) respectively. The solid dots represent averages among the pregnant mothers for each level of daily cigarette consumption. The hollow dot is the local polynomial predictor at zero cigarettes. The point "×" is the predictor after the effect of the covariates is removed. $\hat{\theta}$ is the difference between the × and the solid points at $CIG = 0$.

per day. Identical results were

$$BW = m(CIG, Z) + \varepsilon$$

under the selection on observables assumption that ε is independent of CIG and Z. In fact, the effect of smoking 10 cigarettes daily is a decrease of 200 grams in the birth weight. Hence, if selection on observables is to be believed, then the first cigarette

has an effect 50% higher than the following 9 cigarettes.

Bibliography

Douglas Almond, Kenneth Y. Chay, and David S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Takeshi Amemiya. The estimation of a simultaneous tobit model. *International Economic Review*, (20):169–181, 1979.

Takeshi Amemiya. *Advanced Econometrics*. Harvard Univ Press, Cambridge, Massachussetts, 1985.

Richard W. Blundell and Joel L. Horowitz. A non-parametric test of exogeneity. *Review of Economic Studies*, 74(4):1035–1058, 2007.

Richard W. Blundell and James L. Powell. Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, volume 2, pages 655–679, 2003.

Richard W. Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669, 2007.

David Card, David S. Lee, and Zhuan Pei. Quasi-experimental identification and estimation in the regression kink design. Working Papers 1206, Princeton University, Department of Economics, Industrial Relations Section, November 2009.

Xiaohong Chen and Yanqin Fan. Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series. *Journal of Econometrics*, 91(2):373–401, 1999.

Yuan S. Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Verlag, New York, 3rd edition, 1997.

Serge Darolles, Jean-Pierre Florens, and Eric Renault. Non parametric instrumental regression. IDEI Working Papers 228, Institut d'Économie Industrielle (IDEI), Toulouse, 2003.

Jianqin Fan and Irene Gijbels. *Local Polynomial Modelling and its Applications*. Chapman & Hall, London, United Kingdon, 1996.

Yanqin Fan and Qi Li. Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, 64(4):865–890, 1996.

A. Ronald Gallant and Douglas W. Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica*, (55):363–390, 1987.

Christian Gourieroux, Alain Monfort, Eric Renault, and Alain Trognon. Generalized residuals. *Journal of Econometrics*, (34):5–32, 1987.

Peter Hall and Joel L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

James Heckman. Dummy endogenous variables in a simultaneous equation system. *Econometrica*, (46):931–959, 1978.

Joel L. Horowitz. Applied nonparametric instrumental variables estimation. *Unpublished Manuscript, Department of Economics, Northwestern University*, 2009. URL `http://faculty.wcas.northwestern.edu/ jlh951/papers/FSPaper.pdf`.

Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: a guide to practice. *Journal of Econometrics*, 142(2):615–635, February 2008.

M. Chris Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

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

Lung-Fei Lee. Identification and estimation in binary choice models with limited (censored) dependent variables. *Econometrica*, (47):977–996, 1979.

Lung-Fei Lee. Tests for the bivariate normal distribution in econometric models with selectivity. *Econometrica*, (52):843–863, 1984.

Michael Lejeune and Pascal Sarda. Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis*, 14(4):457–471, 1992.

Qi Li. Efficient estimation of additive partially linear models. *International Economic Review*, 41(4):1073–1092, 2000.

Qi Li and Jeffrey S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.

Oliver Linton. Second order approximation in the partially linear regression model. *Econometrica*, 63(5):1079–1112, 1995.

Oliver Linton and Jens P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100, 1995.

Judith Lumley, Catherine Chamberlain, Therese Dowswell, Sandy Oliver, Laura Oakley, and Lyndsey Watson. Interventions for promoting smoking cessation during pregnancy (Cochrane Review). *The Cochrane Library*, 8(3), July 2009.

Brigitte C. Madrian and Dennis F. Shea. The Power of Suggestion: Inertia in 401 (k) Participation and Savings Behavior*. *Quarterly Journal of Economics*, 116(4): 1149–1187, 2001.

Elias Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–600, 1996.

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

Whitney K. Newey. Efficient estimation os limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics*, (36):231–250, 1987.

Whitney K. Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2):233–253, 1994.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 09 2003.

Adrian Pagan and Francis Vella. Diagnostic tests for models based on individual data: A survey. *Journal of Applied Econometrics*, (4):S29–S60, 1989.

Jack Porter. Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*, 2003. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.540`.

Peter M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56 (4):931–954, 1988.

David Ruppert and Matt P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3):1346–1370, 1994.

Emmanuel Saez. Do taxpayers bunch at kink points? Working Paper 7366, National Bureau of Economic Research, September 1999.

Mary Sexton and Richard Hebel. A clinical trial of change in maternal smoking and its effect on birth weight. *JAMA*, 251(7), Feb 1984.

Richard Smith and Richard W. Blundell. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica*, (54), 1987.

Francis Vella. A simple estimator for simultabeous models with censored endogenous regressors. *International Economic Review*, 34(2):441–457, May 1993.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

Mark O. Wilhelm. Bequest behavior and the effect of heirs' earnings: Testing the altruistic model of bequests. *The American economic review*, 86(4):874–892, 1996.

Appendix A

Theorems from part I

A.1 Statement and proof of result 3.1

Assumption A.1. *Assume that*

1. *The matrix* $\tilde{\mathrm{W}}$ *(whose rows are the* $W_i^T \mathbf{1}(X_i > 0)$*) has full column rank* $K+1$.
2. $(Z_i^T, Q_i, \varepsilon_i)^T$ *are i.i.d.*
3. $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$, $\mathbb{V}ar(\varepsilon_i \,|\, X_i, Z_i) = \sigma^2 < \infty$
4. $\mathbb{E}(g(Z_i)^2) < \infty$, $\mathbb{V}ar(g(Z_i)W_i\mathbf{1}(X_i = 0)) < \infty$, *and* $\mathbb{E}(W_i W_i^T \mathbf{1}(X_i > 0))$ *is finite and invertible.*
5. $\mathbb{V}ar(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0)) < \infty$.

Theorem A.1. *If assumption A.1 holds, then*

$$\sqrt{n}(\hat{\theta} - \delta\mathbb{E}(g(Z_i)\mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)\mathbf{1}(X_i = 0))) \xrightarrow{d} N\left(0, \sigma^2(p+\Delta) + \delta^2\omega^2\right).$$

Proof. Introducing some notation:

- $h(Z_i) := \mathbb{E}(X_i^* \,|\, X_i^* \leqslant 0, Z_i)$
- $\mu := \mathbb{E}(h(Z_i)g(Z_i)\mathbf{1}(X_i = 0))$
- $\omega^2 := \mathbb{V}ar(h(Z_i)g(Z_i)\mathbf{1}(X_i = 0))$

From A.1(1) and equation (2.4), it is possible to write

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\varepsilon_i \mathbf{1}(X_i = 0) + \delta\frac{1}{n}\sum_{i=1}^{n} h(Z_i)g(Z_i)\mathbf{1}(X_i = 0) + \\ + \frac{1}{n}\sum_{i=1}^{n} g(Z_i)W_i^T\mathbf{1}(X_i = 0)^T \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^T \mathbf{1}(X_i > 0)\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} W_i\varepsilon_i\mathbf{1}(X_i > 0)$$

$$\implies \sqrt{n}(\hat{\theta}-\delta\mu) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g(Z_i)\varepsilon_i \mathbf{1}(X_i=0) + \delta\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}[h(Z_i)g(Z_i)\mathbf{1}(X_i=0)-\mu+$$

$$+\frac{1}{n}\sum_{i=1}^{n} g(Z_i)W_i^T\mathbf{1}(X_i=0)\left(\frac{1}{n}\sum_{i=1}^{n} W_iW_i^T\mathbf{1}(X_i>0)\right)^{-1}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} W_i\varepsilon_i\mathbf{1}(X_i>0)\right)$$

$$=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\Bigg[\varepsilon_i\Bigg(g(Z_i)\mathbf{1}(X_i=0)+\left(\frac{1}{n}\sum_{i=1}^{n} g(Z_i)W_i^T\mathbf{1}(X_i=0)\right)\left(\frac{1}{n}\sum_{i=1}^{n} W_iW_i^T\mathbf{1}(X_i>0)\right)^{-1}\cdot$$

$$\cdot W_i\mathbf{1}(X_i>0)\Bigg)+\delta(h(Z_i)g(Z_i)\mathbf{1}(X_i=0)-\mu)\Bigg] \tag{A.1}$$

Introducing further notation to make it easier to deal with the term inside the sum in (A.1),

- $A_n := \frac{1}{n}\sum_{i=1}^{n} g(Z_i)W_i^T\mathbf{1}(X_i=0)\left(\frac{1}{n}\sum_{i=1}^{n} W_iW_i^T\mathbf{1}(X_i>0)\right)^{-1}$
- $A := \mathbb{E}(g(Z_i)W_i^T\mathbf{1}(X_i=0))\mathbb{E}(W_iW_i^T\mathbf{1}(X_i>0))^{-1}$
- $g_i = \delta(h(Z_i)g(Z_i)\mathbf{1}(X_i=0)-\mu)$

$$\sqrt{n}(\hat{\theta}-\delta\mu) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}[\varepsilon_i(g(Z_i)\mathbf{1}(X_i=0)+A_nW_i\mathbf{1}(X_i>0))+g_i]$$

$$=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}[\varepsilon_i(g(Z_i)\mathbf{1}(X_i=0)+AW_i\mathbf{1}(X_i>0))+g_i]+(A_n-A)\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iW_i\mathbf{1}(X_i>0)$$

By assumption A.1 (2-5), the random variables $R_i = \varepsilon_i[g(Z_i)\mathbf{1}(X_i=0)+AW_i\mathbf{1}(X_i>0)]+g_i$ are i.i.d., $\mathbb{E}(R_i)=0$, and $\mathbb{V}ar(R_i)=\sigma^2(p+\Delta)+\delta^2\omega^2<\infty$. To see this, observe that:

$$\begin{aligned}\mathbb{V}ar(R_i) =&\mathbb{E}((\varepsilon_i[g(Z_i)\mathbf{1}(X_i=0)+AW_i\mathbf{1}(X_i>0)]+g_i)^2)\\ =&\mathbb{E}(\varepsilon_i^2[g(Z_i)^2\mathbf{1}(X_i=0)+AW_iW_i^T\mathbf{1}(X_i>0)A^T+2g(Z_i)\mathbf{1}(X_i=0)\mathbf{1}(X_i>0)AW_i]+\\ &+g_i^2+2g_i\varepsilon_i[g(Z_i)\mathbf{1}(X_i=0)+AW_i\mathbf{1}(X_i>0)])\\ =&\sigma^2(p+\Delta+0)+\delta^2\omega^2+0=\sigma^2(p+\Delta)+\delta^2\omega^2<\infty\end{aligned}$$

Therefore, $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} R_i \xrightarrow{d} N(0,\sigma^2(p+\Delta)+\delta^2\omega^2)$. The random variables $\varepsilon_iW_i\mathbf{1}(X_i>0)$ are i.i.d., with $\mathbb{E}(\varepsilon_iW_i\mathbf{1}(X_i>0))=0$, and $\mathbb{V}ar(\varepsilon_iW_i\mathbf{1}(X_i>0))=\sigma^2\mathbb{E}(W_iW_i^T\mathbf{1}(X_i>0))<\infty$. Therefore $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iW_i\mathbf{1}(X_i>0)\xrightarrow{d} N(0,\sigma^2\mathbb{E}(W_iW_i^T\mathbf{1}(X_i>0)))$. Since $A_n-A\xrightarrow{p}0$ (by assumption A.1(4), by Slutsky's theorem the second term converges in probability to zero. By Slutsky's theorem again,

$$\sqrt{n}(\hat{\theta} - \delta\mu) \xrightarrow{d} N(0, \sigma^2(p + \Delta) + \delta^2\omega^2)$$

which gives the asymptotic distribution of the test statistic.

□

Appendix B

Theorems from part II

B.1 Identification theorems

B.1.1 Theorem 6.1:

The proof requires the following assumption:

Assumption B.1.

$$\lim_{x \downarrow x_0} \int \mathbb{E}(y \,|\, x, z, q)\, dF(Q \,|\, X = x, Z) = \int \mathbb{E}(y \,|\, x_0, z, q) \lim_{x \downarrow x_0} dF(Q \,|\, X = x, Z),$$
$$\lim_{x \uparrow x_0} \int \mathbb{E}(y \,|\, x, z, q)\, dF(Q \,|\, X = x, Z) = \int \mathbb{E}(y \,|\, x_0, z, q) \lim_{x \uparrow x_0} dF(Q \,|\, X = x, Z).$$

Proof. First, observe that assumption 6.1 (1) assures that all $\mathbb{E}(Y \,|\, X = x_0, Z)$, $\lim_{x \downarrow x_0} \mathbb{E}(Y \,|\, X = x, Z)$ and $\lim_{x \uparrow x_0} \mathbb{E}(Y \,|\, X = x, Z)$ are identified for all $z \in \mathcal{Z}_{x_0}$, unless x_0 is a boundary point, in which case either the right or left limit will not be identified. However, item (2) assures $\Delta(Z)$ will be identified, because when one of its parts is not identified, α is such that the part is null. Identification of θ follows because G is known and ν is identified.

From equation (6.1),

$$\begin{aligned}\theta = \int G\Big(\Big[\int \mathbb{E}(Y \,|\, X = x_0, Z, Q) dF(Q \,|\, X = x_0, Z) - \\ - \alpha \lim_{x \downarrow x_0} \int \mathbb{E}(Y \,|\, X, Z, Q) dF(Q \,|\, X = x, Z) - \\ - (1 - \alpha) \lim_{x \uparrow x_0} \int \mathbb{E}(y \,|\, x, z, q) dF(Q \,|\, X = x, Z) \Big], z \Big) d\nu(z)\end{aligned}$$

If X is exogenous, $\int \mathbb{E}(y \,|\, x, z, q) dF(Q \,|\, X = x, Z) = \mathbb{E}(Y \,|\, X = x, Z) \int dF(Q \,|\, X = x, Z) = \mathbb{E}(Y \,|\, X = x, Z)$, hence $\theta = \int G\big(\big[\mathbb{E}(y \,|\, x = x_0, z) - \alpha \lim_{x \downarrow x_0} \mathbb{E}(Y \,|\, X = x, Z) -$

$(1-\alpha)\lim_{x\uparrow x_0}\mathbb{E}(Y\,|\,X=x,Z)\big],z\big)d\nu(z)$. From assumption 6.1 (2), $\mathbb{E}(Y\,|\,X=x,Z)$ is continuous, and therefore $\theta=\int G\,(0,z)\,d\nu(z)=0$. □

B.1.2 Proof of Remark 6.4:

f continuous in X at x_0 implies that for a given value of z, and q, $\forall\epsilon>0$, $\exists\,\delta_{z,q}>0$ such that $|x-x_0|<\delta_{z,q}\implies|f_Y(x,z,q,\varepsilon)-f_Y(x_0,z,q,\varepsilon)|<\epsilon$. Hence,

$$\begin{aligned}|\mathbb{E}(y\,|\,x,z,q)-\mathbb{E}(y\,|\,x=x_0,z,q)|=&\Big|\int f_Y(x,z,q,\varepsilon)\,dF(\varepsilon\,|\,x,z,q)\\&\qquad-\int f_Y(x_0,z,q,\varepsilon)\,dF(\varepsilon\,|\,x=x_0,z,q)\Big|\\=&\left|\int\left(f_Y(x,z,q,\varepsilon)-f_Y(x_0,z,q,\varepsilon)\right)dF(\varepsilon)\right|\\\leqslant&\int|f_Y(x,z,q,\varepsilon)-f_Y(x_0,z,q,\varepsilon)|\;dF(\varepsilon)<\epsilon.\end{aligned}$$

B.2 Estimation in the linear case

B.2.1 Theorem 7.1:

First, observe that

$$\sqrt{n}\begin{bmatrix}\hat\delta^+-\delta^+\\ \hat\delta^--\delta^-\end{bmatrix}=\begin{bmatrix}\left[\frac{1}{n}\sum_{i=1}^n X_iX_i^T\mathbf{1}(X_i>x_0)\right]^{-1}&0\\0&\left[\frac{1}{n}\sum_{i=1}^n X_iX_i^T\mathbf{1}(X_i<x_0)\right]^{-1}\end{bmatrix}\cdot\\\cdot\;\sqrt{n}\frac{1}{n}\sum_{i=1}^n\begin{bmatrix}X_i\,\epsilon_i\,\mathbf{1}(X_i>x_0)\\X_i\,\epsilon_i\,\mathbf{1}(X_i<x_0)\end{bmatrix}.$$

Assumptions 7.1 (1), assumption 7.2 (3), the LLN and the continuous mapping theorem guarantee that $\left[\frac{1}{n}\sum_{i=1}^n X_iX_i^T\mathbf{1}(X_i>x_0)\right]^{-1}\xrightarrow{p}\mathbb{E}(X_iX_i^T\mathbf{1}(X_i>x_0))^{-1}$ and $\left[\frac{1}{n}\sum_{i=1}^n X_iX_i^T\mathbf{1}(X_i<x_0)\right]^{-1}\xrightarrow{p}\mathbb{E}(X_iX_i^T\mathbf{1}(X_i<x_0))^{-1}$. Since the ϵ_i are functions of Y_i, X_i and Z_i, they are i.i.d. Moreover,

$$\begin{aligned}\mathbb{C}ov(X_i\,\epsilon_i\,\mathbf{1}(X_i>x_0),X_i\,\epsilon_i\,\mathbf{1}(X_i<x_0))&=\mathbb{E}(X_i\,\epsilon_i\,\mathbf{1}(X_i>x_0))\mathbb{E}(X_i\,\epsilon_i\,\mathbf{1}(X_i<x_0))\\&=\mathbb{E}(X_i\,\mathbb{E}(\epsilon_i\,|\,X_i)\,\mathbf{1}(X_i>x_0))\mathbb{E}(X_i\,\mathbb{E}(\epsilon_i\,|\,X_i)\,\mathbf{1}(X_i<x_0))=0.\end{aligned}$$

Therefore, assumptions 7.1 (1) and 7.2 (2) and the vector CLT imply that

$$\begin{aligned}&\sqrt{n}\frac{1}{n}\sum_{i=1}^n\begin{bmatrix}X_i\,\epsilon_i\,\mathbf{1}(X_i>x_0)\\X_i\,\epsilon_i\,\mathbf{1}(X_i<x_0)\end{bmatrix}\xrightarrow{d}\\&\qquad\xrightarrow{d}\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\sigma^2\begin{bmatrix}\mathbb{E}(X_iX_i^T\mathbf{1}(X_i>x_0))&0\\0&\mathbb{E}(X_iX_i^T\mathbf{1}(X_i<x_0))\end{bmatrix}\right).\end{aligned}$$

Finally, Slutsky's theorem implies that

$$\sqrt{n}\begin{bmatrix}\hat{\delta}^+-\delta^+\\ \hat{\delta}^--\delta^-\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \sigma^2\begin{bmatrix}\mathbb{E}(X_iX_i^T\mathbf{1}(X_i>x_0)) & 0\\ 0 & \mathbb{E}(X_iX_i^T\mathbf{1}(X_i<x_0))\end{bmatrix}^{-1}\right).$$

By the continuous mapping theorem,

$$\alpha\sqrt{n}(\hat{\delta}^+-\delta^+)+(1-\alpha)\sqrt{n}(\hat{\delta}^--\delta^-)\xrightarrow{d}\mathcal{N}(0,v),$$

where

$$v=\sigma^2\left(\alpha^2\mathbb{E}(X_iX_i^T\mathbf{1}(X_i>x_0))^{-1}+(1-\alpha)^2\sigma^2\mathbb{E}(X_iX_i^T\mathbf{1}(X_i<x_0))^{-1}\right).$$

By assumption 7.1 item (1), assumption 7.2 item (1) and the strong LLN, $\hat{\mathbb{E}}(g(Z_i)\,|\,X_i=x_0)\xrightarrow{p}\mathbb{E}(g(Z_i)\,|\,X_i=x_0)$ and $\hat{\mathbb{E}}(g(Z_i)Z_i\,|\,X_i=x_0)\xrightarrow{p}\mathbb{E}(g(Z_i)Z_i\,|\,X_i=x_0)$, and since the limits are scalar, the convergence holds for the vector. By Slutsky's theorem,

$$\sqrt{n}B_n\xrightarrow{d}\mathcal{N}(0,V_B).$$

It is easy to establish the joint convergence of A_n and B_n with the same arguments as above. Observe that since $(\hat{\delta}^+-\delta^+)$ and $(\hat{\delta}^--\delta^-)$ use only data for which $X_i\neq x_0$,

$$\begin{aligned}\mathbb{C}ov(A_n,B_n)=\mathbb{E}\left(\frac{1}{\hat{p}_{x_0}^2}(\Delta(Z_i)g(Z_i)\mathbf{1}(X_i=x_0)-\theta)[x_0\;\; g(Z_i)Z_i\mathbf{1}(X_i=x_0)]\right)\cdot\\ \cdot\,\mathbb{E}\left(\alpha(\hat{\delta}^+-\delta^+)+(1-\alpha)(\hat{\delta}^--\delta^-)\right)=0\end{aligned}\tag{B.1}$$

because the weighted least squares estimators δ^+ and δ^- are unbiased. Equation (B.1) and the continuous mapping theorem imply that

$$\sqrt{n}(\hat{\theta}-\theta)=\sqrt{n}A_n-\sqrt{n}B_n\xrightarrow{d}\mathcal{N}(0,V_A+V_B).$$

B.2.2 Theorem 7.3

The convergence of $\mathbb{P}\left(\sqrt{n}\frac{\hat{\theta}}{\hat{V}_B}\leqslant c_\lambda\right)$ to λ as $n\to\infty$ under H_0 is a trivial consequence of theorem 7.2. Under H_1,

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}\frac{\hat{\theta}}{\sqrt{\hat{V}_B}}>c_\lambda\right)=\mathbb{P}\Bigg(\sqrt{n}\left(\frac{\hat{\theta}-\theta}{\sqrt{V_A+V_B}}\right)-\frac{c_\lambda\left(\sqrt{\hat{V}_B}-\sqrt{V_B}\right)}{\sqrt{V_A+V_B}}>\\ >\frac{c_\lambda\sqrt{V_B}}{\sqrt{V_A+V_B}}-\sqrt{n}\frac{\theta}{\sqrt{V_A+V_B}}\Bigg).\end{aligned}$$

From theorem 7.2 and the continuous mapping theorem, $\sqrt{\hat{V}_B} - \sqrt{V_B} \xrightarrow{p} 0$, and therefore, by the same theorem and Slutsky's theorem, $\sqrt{n}\left(\frac{\hat{\theta}-\theta}{\sqrt{V_A+V_B}}\right) - \frac{c_\lambda\left(\sqrt{\hat{V}_B}-\sqrt{V_B}\right)}{\sqrt{V_A+V_B}} \xrightarrow{d} \mathcal{N}(0,1)$. Since $-\sqrt{n}\frac{\theta}{\sqrt{V_A+V_B}} \to -\infty$ as $n \to \infty$, it is easy to prove that $\mathbb{P}\left(\sqrt{n}\left(\frac{\hat{\theta}-\theta}{\sqrt{V_A+V_B}}\right) - \frac{c_\lambda\left(\sqrt{\hat{V}_B}-\sqrt{V_B}\right)}{\sqrt{V_A+V_B}} > \frac{c_\lambda\sqrt{V_B}}{\sqrt{V_A+V_B}} - \sqrt{n}\frac{\theta}{\sqrt{V_A+V_B}}\right) \to 1$ as $n \to \infty$.

Under the alternatives $\theta/\sqrt{n}$, the same manipulations as above yield

$$\mathbb{P}\left(\sqrt{n}\frac{\hat{\theta}}{\sqrt{\hat{V}_B}} \leqslant c_\lambda\right) = \mathbb{P}\left(\sqrt{n}\left(\frac{\hat{\theta}-\theta/\sqrt{n}}{\sqrt{V_A+V_B}}\right) - \frac{c_\lambda\left(\sqrt{\hat{V}_B}-\sqrt{V_B}\right)}{\sqrt{V_A+V_B}} \leqslant \frac{c_\lambda\sqrt{V_B}-\theta}{\sqrt{V_A+V_B}}\right)$$

and since $\sqrt{n}\left(\frac{\hat{\theta}-\theta/\sqrt{n}}{\sqrt{V_A+V_B}}\right) - \frac{c_\lambda\left(\sqrt{\hat{V}_B}-\sqrt{V_B}\right)}{\sqrt{V_A+V_B}} \xrightarrow{d} \mathcal{N}(0,1)$, the result of the theorem follows immediately.

B.3 Estimation in the partially linear case

B.3.1 Theorem 7.4:

From equation (7.11), equation (7.8) can be rewritten as

$$\begin{aligned}
B_n &= B_n^1 + B_n^2 \\
B_n^1 &:= \hat{\mathbb{E}}(g(Z_i)\,|\,X_i = x_0)\left(\alpha[\tilde{\tau}^+(x_0)^{\lim} - \tau^+(x_0)^{\lim}] + (1-\alpha)[\tilde{\tau}^-(x_0)^{\lim} - \tau^-(x_0)^{\lim}]\right) \\
B_n^2 &:= \alpha\Big[\big(\hat{\mathbb{E}}(g(Z_i)Z_i\,|\,X_i = x_0)^T - \hat{\mathbb{E}}(g(Z_i)\,|\,X_i = x_0)\, e_1^T(\mathrm{X}^T D^+ \mathrm{X})^{-1}\mathrm{X}^T D^+ Z\big)\cdot \\
&\quad \cdot (\hat{\gamma}^+ - \gamma^+)\Big] + (1-\alpha)\Big[\big(\hat{\mathbb{E}}(g(Z_i)Z_i\,|\,X_i = x_0)^T - \\
&\quad - \hat{\mathbb{E}}(g(Z_i)\,|\,X_i = x_0)\, e_1^T(\mathrm{X}^T D^- \mathrm{X})^{-1}\mathrm{X}^T D^- Z\big)(\hat{\gamma}^- - \gamma^-)\Big].
\end{aligned}$$

It will be shown that B_n^1 converges at the rate $\sqrt{nh}$ and $(A_n + B_n^2)$ converges at the rate $\sqrt{n}$, and therefore $\mathbb{V}ar(\sqrt{nh}B_n^2) = O(h)$. The consequence of the disparity between the rates is that only the variance of B_n^1 will affect the asymptotic variance of $\hat{\theta}$. However, in order to study the influence of B_n^2 in small samples, the results will consider variance terms that are at least $O(h)$, which is the same to say that any term which is $o(h)$ will be considered irrelevant in the variance calculations and not taken into account.

For the distribution of B_n^1, observe that $\tilde{\tau}^+(x_0)^{\lim}$ and $\tilde{\tau}^-(x_0)^{\lim}$ are local polynomial regressions of $Y_i - Z_i^T\gamma^+$ and $Y_i - Z_i^T\gamma^-$ on X_i using only observations for which $X_i > x_0$ and $X_i < x_0$ respectively. These are standard local polynomial regressions

of the kind used in Porter (2003) for the estimation of the sides of the discontinuity in the regression discontinuity design. There is one crucial difference: Porter assumes that the running variable X_i has a density function in a neighborhood of x_0. Since here $\mathbb{P}(X_i = x_0) > 0$, this is no longer possible. However, by assumption 7.3 (3) the distribution function $F(x \,|\, x \neq x_0)$ has a density function in $[x^-, x_0) \cap (x_0, x^+]$, and it is equal to

$$\varphi(x) := \frac{\frac{d}{dx}F(x)}{\mathbb{P}(X_i \neq x_0)}.$$

Let the random variables X_i^+ be defined in $[x_0, \infty) \cap \mathcal{X}$ with density function $\tilde{\varphi}(x)^+ = \varphi(x)$ in $(x_0, x^+]$ and $\tilde{\varphi}(x_0)^+ = \lim_{x \downarrow x_0} \varphi(x)$. Then $\mathbb{P}(\mathrm{X}_i^+ = X_i \mathbf{1}(X_i > x_0)) = 1$. Define X_i^- analogously. Though in theorem 3 Porter assumes that the X_i have a density function in an open set $\mathcal{N} \ni x_0$, all the equations use either $X_i \mathbf{1}(X_i > x_0)$ or $X_i \mathbf{1}(X_i < x_0)$, and the results only require that $X_i \mathbf{1}(X_i \geqslant x_0)$ has a density in $[x_0, x^+)$ and that $X_i \mathbf{1}(X_i \leqslant x_0)$ has a density in $(x^-, x_0]$. Hence, theorem 3 in Porter (2003) can be applied to X_i^+ and X_i^-, and the results will be valid to $X_i \mathbf{1}(X_i > x_0)$ and $X_i \mathbf{1}(X_i < x_0)$ respectively with probability one. Assumption 7.3 (4)-(7) complete the requirements of the theorem. Let $\tilde{n} := \sum_{i=1}^n \mathbf{1}(X_i \neq x_0)$, Porter shows that

$$\sqrt{h\tilde{n}} \begin{pmatrix} \tilde{\tau}^+(x_0)^{\lim} - \tau^+(x_0)^{\downarrow} - \tilde{\mathcal{B}}_n^+ \\ \tilde{\tau}^-(x_0)^{\lim} - \tau^-(x_0)^{\uparrow} - \tilde{\mathcal{B}}_n^- \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tilde{\mathcal{V}}^+ & 0 \\ 0 & \tilde{\mathcal{V}}^- \end{bmatrix} \right) \tag{B.2}$$

where if p is odd,

$$\tilde{\mathcal{B}}_n^+ = h^{p+1} \frac{\tau^{+(p+1)}(x_0)^{\lim}}{(p+1)!} e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+1}) = \mathcal{B}_n^+$$

and if p is even,

$$\begin{aligned} \tilde{\mathcal{B}}_n^+ &= h^{p+2} \left[\frac{\tau^{+(p+1)}(x_0)^{\lim}}{(p+1)!} \frac{\tilde{\varphi}'(x_0)^+}{\tilde{\varphi}(x_0)^+} \right] e_1^T \Lambda_0^{-1} (\Upsilon_{p+2} - \Lambda_1 \Lambda_0 \Upsilon_{p+1}) \\ &\quad + \left[\frac{\tau^{+(p+2)}(x_0)^{\lim}}{(p+2)!} \right] e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+2}) \\ &= h^{p+2} \left[\frac{\tau^{+(p+1)}(x_0)^{\lim}}{(p+1)!} \frac{\phi'(x_0)^{\downarrow}}{\phi(x_0)^{\downarrow}} \right] e_1^T \Lambda_0^{-1} (\Upsilon_{p+2} - \Lambda_1 \Lambda_0 \Upsilon_{p+1}) \\ &\quad + \left[\frac{\tau^{+(p+2)}(x_0)^{\lim}}{(p+2)!} \right] e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+2}) = \mathcal{B}_n^+ \end{aligned}$$

and analogously for $\mathcal{B}_n^-$. Observe that $\mathbb{E}(\sigma_\epsilon^2(X_i, Z_i) \,|\, X_i = x \,,\, X_i \neq x_0) = \sigma^2(x)$ for all X in (x_0, x^+). Hence, if p is even or odd,

$$\begin{aligned} \tilde{\mathcal{V}}^+ &= \frac{\mathbb{V}ar(\epsilon_i \,|\, \mathrm{X}_i^+ = x_0)}{\tilde{\varphi}(x_0)^+} e_1^T \Lambda_0^{-1} \Omega \, \Lambda_0^{-1} e_1 = \mathbb{P}(X_i \neq x_0) \frac{\sigma^2(x_0)^{\downarrow}}{\phi(x_0)^{\downarrow}} e_1^T \Lambda_0^{-1} \Omega \, \Lambda_0^{-1} e_1 \\ &= \mathbb{P}(X_i \neq x_0) \, \mathcal{V}^+, \end{aligned}$$

and analogously for $\mathcal{V}^-$. By assumption 7.1 (1) and the LLN, $\tilde{n}/n \xrightarrow{p} \mathbb{P}(X_i \neq x_0) > 0$, and by the Continuous Mapping theorem and Slutsky's theorem,

$$\sqrt{hn}\begin{pmatrix} \tilde{\tau}^+(x_0)^{\lim} - \tau^+(x_0)^{\downarrow} - \mathcal{B}_n^+ \\ \tilde{\tau}^-(x_0)^{\lim} - \tau^-(x_0)^{\uparrow} - \mathcal{B}_n^- \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathcal{V}^+ & 0 \\ 0 & \mathcal{V}^- \end{bmatrix}\right). \tag{B.3}$$

Also, by assumption 7.1 item (1), assumption 7.2 item (1) and the LLN, $\hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0) \xrightarrow{p} \mathbb{E}(g(Z_i) \,|\, X_i = x_0)$. Hence, Slutsky's theorem and the continuous mapping theorem imply

$$\sqrt{nh}\big(B_n^1 - \mathcal{B}_n\big) \xrightarrow{d} \mathcal{N}\Big(0, \mathbb{E}(g(Z_i) \,|\, X_i = x_0)^2 \left[\alpha^2 \mathcal{V}^+ + (1-\alpha)^2 \mathcal{V}^-\right]\Big).$$

For determining the asymptotic distribution of $B_n^2 + A_n$, denote

$$\begin{aligned}
a_n^+ &:= e_1^T (\mathrm{X}^T D^+ \mathrm{X})^{-1} \mathrm{X}^T D^+ Z, \\
a_n^- &:= e_1^T (\mathrm{X}^T D^+ \mathrm{X})^{-1} \mathrm{X}^T D^- Z, \\
b_n^+ &:= \alpha \left[\hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0)^T - \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)\, a_n^+\right], \\
b_n^- &:= (1-\alpha) \left[\hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0)^T - \hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0)\, a_n^-\right],
\end{aligned}$$

then

$$B_n^2 + A_n = b_n^+(\hat{\gamma}^+ - \gamma^+) + b_n^-(\hat{\gamma}^- - \gamma^-) + A_n = \begin{bmatrix} b_n^+ & b_n^+ & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma}^+ - \gamma^+ \\ \hat{\gamma}^- - \gamma^- \\ A_n \end{bmatrix}.$$

First, observe that assumptions 7.3 (3) and (6)-(8) and Theorem 3 in Porter (2003) guarantee that $a_n^+ \xrightarrow{p} \mathbb{E}(Z_i \,|\, X_i = x_0)^{\downarrow}$ and $a_n^- \xrightarrow{p} \mathbb{E}(Z_i \,|\, X_i = x_0)^{\uparrow}$. By assumption 7.1 item (1), assumption 7.3 item (1) and the LLN, $\hat{\mathbb{E}}(g(Z_i) \,|\, X_i = x_0) \xrightarrow{p} \mathbb{E}(g(Z_i) \,|\, X_i = x_0)$, and $\hat{\mathbb{E}}(g(Z_i) Z_i \,|\, X_i = x_0) \xrightarrow{p} \mathbb{E}(g(Z_i) Z_i \,|\, X_i = x_0)$. Hence, by Slutsky's theorem, $[b_n^{+T} \quad b_n^{-T} \quad 1] \xrightarrow{p} [\alpha C_+^T \quad (1-\alpha) C_-^T \quad 1]$. From assumption 7.3 (2) and Slutsky's theorem,

$$\begin{aligned}
\sqrt{n}(B_n^2 + A_n) &\xrightarrow{d} \begin{bmatrix} \alpha C_+^T & (1-\alpha) C_-^T & 1 \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathcal{V}_\gamma^+ & 0 & 0 \\ 0 & \mathcal{V}_\gamma^- & 0 \\ 0 & 0 & V_A \end{bmatrix}\right) \\
&\sim \mathcal{N}\left(0, \alpha^2 C_+^T \mathcal{V}_\gamma^+ C_+ + (1-\alpha)^2 C_-^T \mathcal{V}_\gamma^- C_- + V_A\right).
\end{aligned} \tag{B.4}$$

Since $\sqrt{nh}(B_n^2 + A_n) \xrightarrow{p} 0$, Slutsky's theorem guarantees the joint convergence of $\sqrt{nh}(B_n^1 - \mathcal{B}_n)$ and $\sqrt{nh}(B_n^2 + A_n)$. The only remaining task is to calculate the covariance $nh\, \mathbb{C}ov(B_n^1 - \mathcal{B}_n, B_n^2 + A_n)$ up to the $O(h)$ level. This result requires that

one return to the proof of theorem 3 in Porter (2003), p. 44., and refer to equations (17) to (22) in that paper. B_n^1 can be rewritten as

$$\sqrt{nh}(B_n^1 - \mathcal{B}_n) = e_1^T \left[\alpha D_{n+} \sum_{s=17}^{19} E_s + (1-\alpha) D_{n-} \sum_{s=20}^{22} E_s \right]$$

where E_s is the numerator in equation (s) in p.44 in Porter (2003), the notation translates here as

$$d_i = \mathbf{1}(X_i > x_0), \qquad Z_i = \frac{1}{h}\mathrm{X}_i, \qquad D_{n+} = (n^{-1}\mathrm{X}^T D^+ \mathrm{X})^{-1}, \qquad B_{n+} = \mathcal{B}_n^+,$$

$$Y_i^+ = \tau^+(X_i) - \tau^+(x_0)^{\downarrow} - \tau^{+(1)}(x_0)^{\lim}(X_i - x_0) - \cdots - \frac{1}{p!}\tau^{+(p)}(x_0)^{\lim}(X_i - x_0)^p + \varepsilon_i,$$

$$\mu_j^+(x) = \tau^+(x) - \left[\tau^+(x_0)^{\downarrow} + \tau^{+(1)}(x_0)^{\lim}(X_i - x_0) + \cdots + \frac{1}{j!}\tau^{+(j)}(x_0)^{\lim}(X_i - x_0)^j \right],$$

and the terms with the "-" sign are defined analogously. Hence,

$$\begin{aligned} \frac{1}{h} nh\, \mathbb{C}ov(B_n^1 - \mathcal{B}_n, B_n^2 + A_n) = \Bigg[& \alpha \sum_{s=17}^{19} \frac{1}{\sqrt{h}} \mathbb{C}ov(e_1^T D_{n+} E_s, \sqrt{n}(B_n^2 + A_n)) \\ & + (1-\alpha) \sum_{s=20}^{22} \frac{1}{\sqrt{h}} \mathbb{C}ov(e_1^T D_{n-} E_s, \sqrt{n}(B_n^2 + A_n)) \Bigg] \end{aligned}$$

We will deal with the first term, for $s = 17, 18$ and 19. The second term, for $s = 20, 21$ and 22 is analogous.

$$\begin{aligned} \frac{1}{\sqrt{h}} \mathbb{C}ov(e_1^T D_{n+} E_s, \sqrt{n}(B_n^2 + A_n)) = & \frac{1}{\sqrt{h}} \mathbb{C}ov(e_1^T D_{n+} E_s, \sqrt{n} B_n^2) + \\ & + \frac{1}{\sqrt{h}} \mathbb{C}ov(e_1^T D_{n+} E_s, \sqrt{n} A_n) \end{aligned}$$

A_n is composed exclusively by observations for which $X_i = x_0$, while $D_{n+}E_s$ is composed exclusively by observations for which $X_i > x_0$. They are therefore independent, and since $\mathbb{E}(A_n) = 0$, $\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+} E_s, \sqrt{n} A_n) = 0$. $B_n^2 = b_n^+(\hat{\gamma}^+ - \gamma^+) + b_n^-(\hat{\gamma}^- - \gamma^-)$. The term $b_n^-(\hat{\gamma}^- - \gamma^-)$ is composed exclusively of observations for which $X_i = x_0$ and $X_i > x_0$. It is therefore independent of $D_{n+}E_s$, and $\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+} E_s, b_n^- \sqrt{n}(\hat{\gamma}^- - \gamma^-)) = 0$. The term E_{19} is not random, and therefore, $\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+} E_{19}, b_n^+ \sqrt{n}(\hat{\gamma}^+ - \gamma^+)) = 0$. For the term E_{18}, we will use Hölder's inequality:

$$\begin{aligned} \frac{1}{\sqrt{h}} \mathbb{C}ov(e_1^T D_{n+} E_{18}, b_n^+ \sqrt{n}(\hat{\gamma}^+ - \gamma^+)) = & \mathbb{E}(e_1^T D_{n+} E_{18} b_n^+ \sqrt{n}(\hat{\gamma}^+ - \gamma^+)) \\ & - \mathbb{E}(e_1^T D_{n+} E_{18}) \mathbb{E}(b_n^+ \sqrt{n}(\hat{\gamma}^+ - \gamma^+)) \\ & \leqslant \mathbb{E}((e_1^T D_{n+} E_{18})^2)^{1/2} \mathbb{E}(n(b_n^+(\hat{\gamma}^+ - \gamma^+))^2)^{1/2} \\ & + \mathbb{E}(|e_1^T D_{n+} E_{18}|) \mathbb{E}(|b_n^+ \sqrt{n}(\hat{\gamma}^+ - \gamma^+)|) \end{aligned}$$

Porter shows that $\mathbb{V}ar(e_1^T D_{n+} E_{18}) = o(h^{-(p+1)})$, and from assumption 7.3 (2), $\mathbb{E}(n(b_n^+(\hat{\gamma}^+ - \gamma^+))^2)^{1/2}$ is uniformly bounded. Hence, $\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+} E_{18}, b_n^+\sqrt{n}(\hat{\gamma}^+ - \gamma^+)) = o(1)$.

The only remaining term is $\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+} E_{17}, b_n^+\sqrt{n}(\hat{\gamma}^+ - \gamma^+))$. It is necessary to have a better understanding of $\hat{\gamma}^+$. $b_n^+(\hat{\gamma}^+ - \gamma^+) = b_n^+(\mathrm{Z}_+^T\mathrm{Z}_+)^{-1}\mathrm{Z}_+^T(\mathrm{Y}_+ - \mathrm{Z}_+\gamma^+)$, and for observations such that $X_i > x_0$,

$$\begin{aligned}\mathrm{Y}_{i+} - \mathrm{Z}_{i+}\gamma^+ &= Y_i - Z_i^T\gamma^+ - \sum_{j=1}^{n} \mathbf{1}(X_j > x_0)\, T_{i,j}^+(Y_j - Z_j^T\gamma^+) \\ &= \tau^+(X_i) + \epsilon_i - \sum_{j=1}^{n} \mathbf{1}(X_j > x_0)\, T_{i,j}^+(\tau^+(X_j) + \epsilon_j)\end{aligned}$$

Let $T^+ = [T_{i,j}^+\mathbf{1}(X_j > x_0)]$ the $n \times n$ matrix with entry $T_{i,j}\mathbf{1}(X_i > x_0, X_j > x_0)$ in line i, column j, $\tau^+ = (\tau^+(X_1), \ldots, \tau^+(X_n))^T$, and $P_\gamma^+ := b_n^+(\mathrm{Z}_+^T\mathrm{Z}_+)^{-1}\mathrm{Z}_+^T$. Then $b_n^+(\hat{\gamma}^+ - \gamma^+) = P_\gamma^+(I^+ - T^+)(\tau^+ + \epsilon)$, where $I^+ = Diag\{\mathbf{1}(X_1 > x_0), \ldots, \mathbf{1}(X_n > x_0)\}$. Also, $e_1 D_{n+} E_{17} = \sqrt{nh} P_1^+\epsilon$. Hence, since it can be easily shown that $E(\frac{1}{\sqrt{h}} e_1^T D_{n+} E_{17}) = o(1)$ and since $b_n^+\sqrt{n}(\hat{\gamma}^+ - \gamma^+))$ is uniformly bounded,

$$\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+} E_{17}, b_n^+\sqrt{n}(\hat{\gamma}^+ - \gamma^+)) = n\mathbb{E}(P_1^+\epsilon(\hat{\gamma}^+ - \gamma^+)^T b_n^{+T})$$

$$\begin{aligned}&\left(P_1^+\epsilon\right) b_n^+(\hat{\gamma}^+ - \gamma^+) = \left(P_1^+\epsilon\right)(\hat{\gamma}^+ - \gamma^+)^T b_n^{+T} = P_1^+\epsilon(\tau^{+T} + \epsilon^T)(I^+ - T^+)^T P_\gamma^{+T} \\ &\implies \mathbb{E}(\left(P_1^+\epsilon\right) b_n^+(\hat{\gamma}^+ - \gamma^+)) = \mathbb{E}(P_1^+\epsilon^2(I^+ - T^+)^T P_\gamma^{+T})\end{aligned}$$

where $\epsilon^2 = Diag\{\epsilon_1^2, \ldots, \epsilon_n^2\}$. Define $\mathbb{E}(\epsilon^2 \,|\, X) = Diag\{\mathbb{E}(\epsilon_i^2 \,|\, X_i = X_1), \ldots, \mathbb{E}(\epsilon_i^2 \,|\, X_i = X_n)\}$ and $\hat{\mathbb{E}}(\epsilon^2 \,|\, X) = Diag\{\hat{\mathbb{E}}(\epsilon_i^2 \,|\, X_i = X_1\}, \ldots, \hat{\mathbb{E}}(\epsilon_i^2 \,|\, X_i = X_n))^T$. We can then rewrite $\epsilon^2(I^+ - T^+)^T P_\gamma^{+T}$ as

$$\begin{aligned}P_\gamma^T(I^+ - T^+)\epsilon^2 = \; & b_n^+(\mathrm{Z}_+^T\mathrm{Z}_+)^{-1} Z^T(I^+ - T^{+T})(I^+ - T^+)\epsilon^2 \\ = \; & b_n^+(\mathrm{Z}_+^T\mathrm{Z}_+)^{-1}\Big[Z^T\epsilon^2 - \mathbb{E}(Z \,|\, X)^T\epsilon^2 - Z^T\mathbb{E}(\epsilon^2 \,|\, X) + \\ & + \mathbb{E}(Z \,|\, X)^T\mathbb{E}(\epsilon^2 \,|\, X) - (\hat{\mathbb{E}}(Z \,|\, X) - \mathbb{E}(Z \,|\, X))^T\epsilon^2 - \\ & - Z^T(\hat{\mathbb{E}}(\epsilon^2 \,|\, X) - \mathbb{E}(Z \,|\, X)) + (\hat{\mathbb{E}}(Z \,|\, X) - \mathbb{E}(Z \,|\, X))^T\hat{\mathbb{E}}(\epsilon^2 \,|\, X) + \\ & + \mathbb{E}(Z \,|\, X)(\hat{\mathbb{E}}(\epsilon^2 \,|\, X) - \mathbb{E}(\epsilon^2 \,|\, X))\Big].\end{aligned}$$

Hence

$$\begin{aligned}nP_1^+\epsilon^2(I^+ - T^+)^T P_\gamma^{+T} = \Big[&P_1^+\epsilon^2 Z - P_1^+\epsilon^2\mathbb{E}(Z \,|\, X) - P_1^+\mathbb{E}(\epsilon^2 \,|\, X)Z \\ &+ P_1^+\mathbb{E}(\epsilon^2 \,|\, X)\mathbb{E}(Z \,|\, X)\Big]\left(\frac{\mathrm{Z}_+^T\mathrm{Z}_+}{n}\right)^{-1} b_n^{+T} + U_n,\end{aligned}$$

$$
\begin{aligned}
|U_n| \leqslant & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{X_i - x_0}{h}\right)\mathbf{1}(X_i > 0)\mathrm{X}_i\epsilon_i^2\right| \sum_{s=1}^{d}\sup_i\left|\hat{\mathbb{E}}(Z_i^s \mid X_i) - \mathbb{E}(Z_i^s \mid X_i)\right| |b_n^s| \\
& + \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{X_i - x_0}{h}\right)\mathbf{1}(X_i > 0)\mathrm{X}_i Z_i\right\| \sum_{s=1}^{d}\sup_i\left|\hat{\mathbb{E}}(\epsilon_i^2 \mid X_i) - \mathbb{E}(\epsilon_i^2 \mid X_i)\right| |b_n^s| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{X_i - x_0}{h}\right)\mathbf{1}(X_i > 0)\mathrm{X}_i\right| \sup_i\left|\hat{\mathbb{E}}(\epsilon_i^2 \mid X_i)\right| \cdot \\
& \qquad \cdot \sum_{s=1}^{d}\sup_i\left|\hat{\mathbb{E}}(Z_i \mid X_i) - \mathbb{E}(Z_i \mid X_i)\right| |b_n^s| \\
& + \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{X_i - x_0}{h}\right)\mathbf{1}(X_i > 0)\mathrm{X}_i\mathbb{E}(Z_i \mid X_i)\right\| \cdot \\
& \qquad \cdot \sum_{s=1}^{d}\sup_i\left|\hat{\mathbb{E}}(\epsilon_i^2 \mid X_i) - \mathbb{E}(\epsilon_i^2 \mid X_i)\right| |b_n^s| \,.
\end{aligned}
$$

Because the terms with the kernels can easily be shown to be bounded, $\sup_i \left\|\hat{\mathbb{E}}(\epsilon_i^2 \mid X_i)\right\|$ is asymptotically bounded, because $\sup_i \left\|\sum_{j=1}^{n}\mathbf{1}(X_j > x_0)\, T_{i,j}^{+}\epsilon_j^2 - \mathbb{E}(\epsilon_i^2 \mid X_i)\right\| = o_p(1)$, and the other terms are $o_p(1)$ by assumption 7.3 (2), $U_n = o_p(1)$.

Finally, by assumption 7.3 (2),

$$
\begin{aligned}
& \left(\frac{\mathrm{Z}_+^T\mathrm{Z}_+}{n}\right) \xrightarrow{p} \mathbb{V}ar(Z_i \mid X_i) \\
\Longrightarrow \; & nP_1^+\epsilon^2(I^+ - T^+)^T P_\gamma^{+T} \xrightarrow{p} \alpha\big(\mathbb{E}(\epsilon_i^2 Z_i \mid X_i = x_0)^{\downarrow} - \\
& \qquad - \mathbb{E}(\epsilon_i^2 \mid X_i = x_0)^{\downarrow}\mathbb{E}(Z_i Z_i \mid X_i = x_0)^{\downarrow}\big)\, C^+ \\
& \qquad = \alpha\, C^+(\Sigma_z(x_0)^{\downarrow})^{-1}c_{z\epsilon^2}(x_0)^{\downarrow}.
\end{aligned}
$$

Therefore,

$$
\alpha\sum_{s=17}^{19}\frac{1}{\sqrt{h}}\mathbb{C}ov(e_1^T D_{n+}E_s, \sqrt{n}(B_n^2 + A_n)) \to \alpha^2\, C^+(\Sigma_z(x_0)^{\downarrow})^{-1}c_{z\epsilon^2}(x_0)^{\downarrow}
$$

and the result for $s = 20, 21$ and 22 is analogous.

B.3.2 Theorem 7.5:

The convergence of $\hat{\mathcal{V}}$ follows from observing that $\alpha\hat{C}_+ = b_n^+$ and $(1-\alpha)\hat{C}_- = b_n^-$ and its convergence is established in the previous section. Theorem 4 in Porter (2003) guarantees the convergence of $\hat{\mathcal{V}}^+$ and $\hat{\mathcal{V}}^-$, as long as $\hat{\sigma}^2(x_0)^{\lim s} \to \sigma^2(x_0)^{\lim s}$ and

$\hat{\phi}(x_0)^{\lim s} \to \phi(x_0)^{\lim s}$ as $n \to \infty$. The latter is guaranteed by item (3) in assumtpion 7.3. We show the former for $s =$ "$+$", the result for $s =$ "$-$" is analogous.

$$\begin{aligned}\mathbb{E}((Y_i - Z_i^T\hat{\gamma}^+)^2 \,|\, X_i = x_0)^{\downarrow} &= e_1^T(\mathrm{X}^T W^{s+}\mathrm{X})^{-1}\mathrm{X}^T D^+ R^+ \qquad \text{(B.5)}\\ &= e_1^T(\mathrm{X}^T D^+\mathrm{X})^{-1}\mathrm{X}^T D^+ \tilde{R}^+ \\ &\quad - e_1^T(\mathrm{X}^T D^+\mathrm{X})^{-1}\mathrm{X}^T D^+(R^+ - \tilde{R}^+) \qquad \text{(B.6)}\end{aligned}$$

where $\tilde{R}^+ = (\tilde{R}_1^+, \ldots, \tilde{R}_n^+)^T$, $\tilde{R}_i^+ = (Y_i - Z_i^T\gamma^+)^2$. We begin by showing that the second term is $o_p(1)$. Notice that

$$\begin{aligned}R_i^+ - \tilde{R}_i^+ &= (Y_i - Z_i^T\hat{\gamma}^+)^2 - (Y_i - Z_i^T\gamma^+)^2 \\ &= \left[(Y_i - Z_i^T\hat{\gamma}^+) + (Y_i - Z_i^T\gamma^+)\right] Z_i^T(\hat{\gamma}^+ - \gamma^+)\end{aligned}$$

Let

$$\begin{aligned}\hat{U}^+ &= e_1^T(\mathrm{X}^T D^+\mathrm{X})^{-1}\sum_{i=1}^n \mathbf{1}(X_i > x_0)\, k\left(\frac{X_i - x_0}{h}\right)(Y_i - Z_i^T\hat{\gamma}^+)Z_i^T \\ \tilde{U}^+ &= e_1^T(\mathrm{X}^T D^+\mathrm{X})^{-1}\sum_{i=1}^n \mathbf{1}(X_i > x_0)\, k\left(\frac{X_i - x_0}{h}\right)(Y_i - Z_i^T\gamma^+)Z_i^T\end{aligned}$$

Then the second term in (B.5) is

$$e_1^T(\mathrm{X}^T D^+\mathrm{X})^{-1}\mathrm{X}^T D^+(R^+ - \tilde{R}^+) = (\hat{U}^+ + \tilde{U}^+)(\hat{\gamma}^+ - \gamma^+)$$

From equation (7.11) and the proof of theorem 7.4, it's easy to see that

$$\hat{U}^+ + \tilde{U}^+ \xrightarrow{p} 2\,(\tau^+(x_0)^{\downarrow})\,\mathbb{E}(Z_i \,|\, X_i = x_0)^{\downarrow}$$

and since $\hat{\gamma}^+ - \gamma^+ \xrightarrow{p} 0$ by assumption 7.3 item (2), the second term in (B.5) is $o_p(1)$.

The first term in (B.5) is a local polynomial regression of $(Y_i - Z_i^T\gamma^+)^2$ on X_i at x_0. Hence, from assumption 7.3 items (2)-(7) and theorem 4.1 in Ruppert and Wand (1994),

$$e_1^T(\mathrm{X}^T D^+\mathrm{X})^{-1}\mathrm{X}^T D^+\tilde{R}^+ \xrightarrow{p} \lim_{x\downarrow x_0}\mathbb{E}(Y_i - Z_i^T\gamma^+)^2 \,|\, X_i = x) = \lim_{x\downarrow x_0}\mathbb{E}(\epsilon_i^2 \,|\, X_i = x) = \sigma^2(x_0)^{\downarrow}.$$

The proof for the convergence of $\hat{c}_{z\epsilon^2}(x_0)^{\lim}$ is analogous. It is only necessary to observe that

$$\begin{aligned}\mathbb{E}(Z_i(Y_i - Z_i\gamma^+)^2 \,|\, X_i) - \mathbb{E}(Z_i \,|\, X_i)\mathbb{E}((Y_i - Z_i\gamma^+)^2 \,|\, X_i) = \\ = \mathbb{E}(Z_i\epsilon_i^2 \,|\, X_i) - \mathbb{E}(Z_i \,|\, X_i)\mathbb{E}(\epsilon_i^2 \,|\, X_i).\end{aligned}$$

B.3.3 Theorem 7.6:

Observe that

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} > c_\lambda\right) = \mathbb{P}\left(\sqrt{nh}\left(\frac{\hat{\theta}-\theta-\mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right) - \frac{c_\lambda\left(\sqrt{\hat{\mathcal{V}}_n}-\sqrt{\mathcal{V}_n}\right)}{\sqrt{\mathcal{V}_n}} + \frac{\sqrt{nh}\mathcal{B}_n}{\sqrt{\mathcal{V}_n}} > \right.$$
$$\left. > c_\lambda - \frac{\sqrt{nh}\theta}{\sqrt{\mathcal{V}_n}}\right).$$

From theorem 7.5 and the continuous mapping theorem, $\sqrt{\hat{\mathcal{V}}_n} - \sqrt{\mathcal{V}_n} \xrightarrow{p} 0$. Moreover, $\sqrt{nh}\mathcal{B}_n \to 0$, since $\sqrt{nh}h^{p+1} \to 0$. Hence, by theorems 7.4 and Slutsky's, $\sqrt{nh}\left(\frac{\hat{\theta}-\theta-\mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right) - \frac{c_\lambda\left(\sqrt{\hat{\mathcal{V}}_n}-\sqrt{\mathcal{V}_n}\right)}{\sqrt{\mathcal{V}_n}} + \frac{\sqrt{nh}\mathcal{B}_n}{\sqrt{\mathcal{V}_n}} \xrightarrow{d} \mathcal{N}(0,1)$. Under H_0, $\theta = 0$, and the first result follows immediately. Under H_1, since $h \to 0$, $\mathcal{V}_n \to \alpha^2\mathcal{V}_\tau^+ + (1-\alpha)^2\mathcal{V}_\tau^-$, and therefore $-\frac{\sqrt{nh}\theta}{\sqrt{\mathcal{V}_n}} \to -\infty$, from which the second result follows.

Under the alternatives $\theta/\sqrt{nh}$, observe that

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} > c_\lambda\right) = \mathbb{P}\left(\sqrt{nh}\left(\frac{\hat{\theta}-\theta/\sqrt{nh}-\mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right) - \frac{c_\lambda\left(\sqrt{\hat{\mathcal{V}}_n}-\sqrt{\mathcal{V}_n}\right)}{\sqrt{\mathcal{V}_n}} + \frac{\sqrt{nh}\mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right.$$
$$\left. + \theta\left(\frac{1}{\sqrt{\mathcal{V}_n}} - \frac{1}{\sqrt{\alpha^2\mathcal{V}_\tau^+ + (1-\alpha)^2\mathcal{V}_\tau^-}}\right) > c_\lambda - \frac{\theta}{\sqrt{\alpha^2\mathcal{V}_\tau^+ + (1-\alpha)^2\mathcal{V}_\tau^-}}\right).$$

and since $\theta\left(\frac{1}{\sqrt{\mathcal{V}_n}} - \frac{1}{\sqrt{\alpha^2\mathcal{V}_\tau^+ + (1-\alpha)^2\mathcal{V}_\tau^-}}\right) \xrightarrow{p} 0$, by Slutsky's theorem the third result of the theorem follows.

B.4 Estimation in the nonparametric case

B.4.1 Theorem 7.7:

The proof is similar to the proof of the convergence of the nonparametric term in the partially linear case. The essence of the argument is that since the support if $dF(Z_i)$ is finite, all arguments can be done separately for each possible value of Z_i. We begin by deriving the asymptotic distribution of $\hat{\Gamma}(z^m)^+$. This is a standard local polynomial regression of the kind used in Porter (2003) for the estimation of one side of the discontinuity in the regression discontinuity design. There are two differences. First, $\hat{\Gamma}(z^m)^+$ uses only data for which $Z_i = z^m$. Second, the results in Porter assume that the variable X_i has a density function in a neighborhood of x_0. Assumption 7.4

item (1) implies that $\mathbb{P}(X_i = x_0 \,|\, Z_i = z^m) > 0$, so this is no longer possible. However, from item (2), the conditional distribution function

$$\mathbb{P}(X_i \leqslant x \,|\, X_i > x_0 \,,\, Z_i = z^m) = \frac{\mathbb{P}(X_i \leqslant x \,,\, Z_i = z^m) - \mathbb{P}(X_i \leqslant x_0 \,,\, Z_i = z^m)}{\mathbb{P}(X_i > x_0 \,,\, Z_i = z^m)}$$

has a density function in (x_0, x^+), and it is equal to

$$\varphi_m(x) := \frac{\frac{d}{dx}\mathbb{P}(X_i \leqslant x \,,\, Z_i = z^m)}{\mathbb{P}(X_i > x_0 \,,\, Z_i = z^m)}.$$

Though theorem 3 in Porter dependens on the existence of a density function in (x_0, x^+), it is not dependent on the existence of a density function at the discontinuity point x_0, as long as the right limit of $\varphi_m(x)$ at x_0 exists. From assumption 7.4 item (2), this is true and

$$\varphi_m(x_0)^{\downarrow} := \lim_{x \downarrow x_0} \varphi_m(x) := \frac{\phi(x_0, z^m)^{\downarrow}}{\mathbb{P}(X_i > x_0 \,,\, Z_i = z^m)}.$$

Assumption 7.4 (3)-(6) complete the requirements of Theorem 3 in Porter (2003). Let $n_m^+ := \sum_{i=1}^n \mathbf{1}(X_i > x_0)\mathbf{1}(Z_i = z^m)$,

$$\sqrt{h n_m^+}\left(\hat{\Gamma}(z^m)^+ - \tilde{\mathcal{B}}_{m,n}^+\right) \xrightarrow{d} \mathcal{N}\left(0 \,,\, \tilde{\mathcal{V}}_m^+\right) \tag{B.7}$$

where if p is odd,

$$\tilde{\mathcal{B}}_n^+ = h^{p+1}\frac{f_Y^{+(p+1)}(x_0, z^m)^{\lim}}{(p+1)!} e_1^T \Lambda_0^{-1} \Upsilon_{p+1} + o(h^{p+1}) = \mathcal{B}_{m,n}^+$$

and if p is even,

$$\begin{aligned}
\tilde{\mathcal{B}}_n^+ &= h^{p+2}\left[\frac{f_Y^{+(p+1)}(x_0, z^m)^{\lim}}{(p+1)!}\frac{\varphi'(x_0, z^m)^{\downarrow}}{\varphi(x_0, z^m)^{\downarrow}}\right] e_1^T \Lambda_0^{-1}(\Upsilon_{p+2} - \Lambda_1\Lambda_0\Upsilon_{p+1}) \\
&\quad + \left[\frac{f_Y^{+(p+2)}(x_0, z^m)^{\lim}}{(p+2)!}\right] e_1^T \Lambda_0^{-1}\Upsilon_{p+1} + o(h^{p+2}) \\
&= h^{p+2}\left[\frac{f_Y^{+(p+1)}(x_0, z^m)^{\lim}}{(p+1)!}\frac{\phi'(x_0, z^m)^{\downarrow}}{\phi(x_0, z^m)^{\downarrow}}\right] e_1^T \Lambda_0^{-1}(\Upsilon_{p+2} - \Lambda_1\Lambda_0\Upsilon_{p+1}) \\
&\quad + \left[\frac{f_Y^{+(p+2)}(x_0, z^m)^{\lim}}{(p+2)!}\right] e_1^T \Lambda_0^{-1}\Upsilon_{p+1} + o(h^{p+2}) = \mathcal{B}_{m,n}^+
\end{aligned}$$

Also, observe that $\mathbb{E}(\sigma_\epsilon^2(X_i, Z_i) \,|\, X_i = x\,,\, X_i > x_0, Z_i = z^m) = \sigma_m^2(x)$ for all X in (x_0, x^+). Hence, if p is even or odd,

$$\begin{aligned}\tilde{\mathcal{V}}_m^+ &= \frac{\sigma^2(x_0)^\downarrow}{\varphi(x_0, z^m)^\downarrow} e_1^T \Lambda_0^{-1} \Omega \, \Lambda_0^{-1} e_1 = \mathbb{P}(X_i > x_0, Z_i = z^m) \frac{\sigma^2(x_0)^\downarrow}{\phi(x_0, z^m)^\downarrow} e_1^T \Lambda_0^{-1} \Omega \, \Lambda_0^{-1} e_1 \\ &= \mathbb{P}(X_i > x_0\,,\, Z_i = z^m)\, \mathcal{V}_m^+\end{aligned}$$

By assumption 7.1 (1) and the LLN, $n_m^+/n \xrightarrow{p} \mathbb{P}(X_i > x_0\,,\, Z_i = z^m)$, and by the continuous mapping theorem and Slutsky's theorem,

$$\sqrt{nh}\left(\hat{\Gamma}(z^m)^+ - \mathcal{B}_{m,n}^+\right) \xrightarrow{d} \mathcal{N}\left(0\,,\, \mathcal{V}_m^+\right).$$

The exact same reasoning applied to $\hat{\Gamma}(z^m)^-$ will yield the equivalent result for the left limit. Moreover, the result in Porter (2003) states the joint convergence of $\sqrt{nh}\left(\hat{\Gamma}(z^m)^+ - \mathcal{B}_{m,n}^+\right)$ and $\sqrt{nh}\left(\hat{\Gamma}(z^m)^- - \mathcal{B}_{m,n}^-\right)$ by the Cràmer Wold device. The $\hat{\Gamma}(z^m)^+$ are independent for all m, and also independent from the $\hat{\Gamma}(z^m)^-$, because they are built using different parts of the sample, hence by continuous mapping theorem,

$$\begin{aligned}\alpha\sqrt{nh}\,\left(\hat{\Gamma}(z^m)^+ - \mathcal{B}_{m,n}^+\right) + (1-\alpha)\alpha\sqrt{nh}\,\left(\hat{\Gamma}(z^m)^- - \mathcal{B}_{m,n}^-\right) \xrightarrow{d} \\ \xrightarrow{d} \mathcal{N}\left(0\,,\, \alpha^2\mathcal{V}_m^+ + (1-\alpha)^2\mathcal{V}_m^-\right).\end{aligned}$$

Assumption 7.1 (1) and (2), the LLN and Slutsky's theorem imply that $\hat{p}_{x_0}^m \xrightarrow{p} p_{x_0}^m$ jointly for all m. By Slutsky's theorem again, $\sqrt{nh}\,(B_n - \mathcal{B}_n) =$

$$\begin{aligned}&= \begin{bmatrix} \hat{p}_{x_0}^1 & \cdots & \hat{p}_{x_0}^M \end{bmatrix} \begin{bmatrix} \alpha\sqrt{nh}\,\left(\hat{\Gamma}(z^1)^+ - \mathcal{B}_{1,n}^+\right) + (1-\alpha)\alpha\sqrt{nh}\,\left(\hat{\Gamma}(z^1)^- - \mathcal{B}_{1,n}^-\right) \\ \vdots \\ \alpha\sqrt{nh}\,\left(\hat{\Gamma}(z^M)^+ - \mathcal{B}_{M,n}^+\right) + (1-\alpha)\alpha\sqrt{nh}\,\left(\hat{\Gamma}(z^M)^- - \mathcal{B}_{M,n}^-\right) \end{bmatrix} \xrightarrow{d} \\ &\xrightarrow{d} \mathcal{N}\left(0\,,\, \begin{bmatrix} p_{x_0}^1 & \cdots & p_{x_0}^M \end{bmatrix} Diag\{\alpha^2\mathcal{V}_m^+ + (1-\alpha)^2\mathcal{V}_m^-\} \begin{bmatrix} p_{x_0}^1 & \cdots & p_{x_0}^M \end{bmatrix}^T\right) \\ &\sim \mathcal{N}\left(0\,,\, \mathcal{V}\right)\end{aligned}$$

The joint convergence of $\sqrt{nh}B_n$ and $\sqrt{nh}A_n$ is guaranteed by Slutsky's theorem, because $\sqrt{nh}A_n \xrightarrow{p} 0$. In order to derive the small sample covariance, the same considerations as in the correlation between the $\hat{\Gamma}(z^m)^+$ and the $\hat{\Gamma}(z^m)^-$ for all m apply here, namely that they are independent from A_n because they are built using different observations. A_n may be correlated with $\mathcal{B}_n$. Equation (7.4) and lemma **??** imply that $nh\,\mathbb{E}(A_n\mathcal{B}_n) = \sqrt{h}\,O\left(h^{p+1}\right) = h\,O(h^{p+1/2})$, which is of order smaller than h, and therefore the correlation os negligible. Hence, the small sample variance is

$$\mathcal{V} + hV_A + o(h) = \mathcal{V}_n$$

which concludes the demonstration.

B.4.2 Theorem 7.8:

The previous section showed that $\hat{p}^m_{x_0} \xrightarrow{p} p^m_{x_0}$. It only remains to prove that $\hat{\sigma}^2(x_0, z^m)^\downarrow \xrightarrow{p} \sigma^2(x_0, z^m)^\downarrow$ and $\hat{\sigma}^2(x_0, z^m)^\uparrow \xrightarrow{p} \sigma^2(x_0, z^m)^\uparrow$ for all m. In the beginning of section 7.7 in the appendix, it is shown that the restriction to the observations such that $X_i \in (x_0, x^+)$ and $Z_i = z^m$ has a density function in (x_0, x^+) equal to $\varphi_m(x)$. The proof will use Masry (1996)'s result on the uniform convergence of the multivariate local polynomial. From assumption 7.4 and Theorem 6 in that article, if $Z_i = z^m$,

$$\begin{aligned}\sup_{x\in(x_0,x^+)} \left|\hat{f}^+(X_i, Z_i) - f(X_i, Z_i)\right| &= \sup_{x\in(x_0,x^+)} \left|\hat{f}^+(X_i, z^m) - f(X_i, z^m)\right| \\ &= O\left(\left(\frac{\log n}{nh}\right)^{1/2} + h^{p+1}\right)\end{aligned}$$

almost surely. Define $D^+_m = \{i\,;\, X_i > x_0 \text{ and } Z_i = z^m\}$, then by the continuous mapping theorem,

$$\sup_{i\in D^+_m} \left|(\hat{\epsilon}^s_i)^2 - \epsilon_i^2\right| = O\left(\left(\frac{\log n}{nh}\right)^{1/2} + h^{p+1}\right) \quad \text{a.s.}$$

Let $\tilde{R} = (\epsilon_1^2, \ldots, \epsilon_n^2)^T$,

$$\hat{\sigma}^2(x_0, z^m)^\downarrow = P^+_{1,m,x_0}\tilde{R} + P^+_{1,m,x_0}(R - \tilde{R})$$

The first term is a simple local polynomial regression of the ϵ_i^2 onto X_i at x_0, and by theorem 3 in Porter (2003), it is a consistent estimator of $\lim_{x\downarrow x_0} \mathbb{E}(\epsilon_i^2 \,|\, X_i = x, Z_i = z^m) = \sigma^2(x_0, z^m)^\downarrow$. For the second term, let $(v)_i$ denote the i-th element of vector v, and since $\left(P^+_{1,m,x_0}\right)_i$ is different from zero only if $i \in D^+_m$,

$$\begin{aligned}\left|P^+_{1,m,x_0}(R - \tilde{R})\right| &= \left|\left(\frac{\mathrm{X}_x^T W^s_{x,m} \mathrm{X}_x}{nh}\right)^{-1} \frac{\mathrm{X}_x^T W^s_{x,m}(R - \tilde{R})}{nh}\right| \\ &\leqslant \left\|\left(\frac{\mathrm{X}_x^T W^s_{x,m} \mathrm{X}_x}{nh}\right)^{-1}\right\| \left\|\frac{\mathrm{X}_x^T W^s_{x,m}(R - \tilde{R})}{h}\right\| \\ &\leqslant \left\|\left(\frac{\mathrm{X}_x^T W^s_{x,m} \mathrm{X}_x}{nh}\right)^{-1}\right\| \sup_{i\in D^+_m} \left|\left(\frac{\mathrm{X}_x^T W^s_{x,m}}{nh}\right)_i \left((\hat{\epsilon}^s_i)^2 - \epsilon_i^2\right)\right| \\ &\leqslant \left\|\left(\frac{\mathrm{X}_x^T W^s_{x,m} \mathrm{X}_x}{nh}\right)^{-1}\right\| \left|\left(\frac{\mathrm{X}_x^T W^s_{x,m}}{h}\right)_i\right| \sup_{i\in D^+_m} \left|(\hat{\epsilon}^s_i)^2 - \epsilon_i^2\right|\end{aligned}$$

Observe that $\left(\frac{\mathrm{X}_x^T W^s_{x,m}}{h}\right)_i = \mathbf{1}(i \in D^+_m)\frac{1}{h}k\left(\frac{X_i - x_0}{h}\right)(a_0 + a_1(X_i - x_0) + \cdots + a_p(X_i - x_0)^p)$. From assumption 7.4 (5), the kernel has bounded support, and since k is continuous,

there exists $\bar{k}$ such that $|k(u)| \leqslant \bar{k}$ for all u. Let $u^{\max} := \sup_u \{u\,;\, k(u) \neq 0\}$, define $x_h^{\max} := x_0 + u^{\max} h$. Hence,

$$\begin{aligned}\left|\left(\frac{\mathrm{X}_x^T W_{x,m}^s}{h}\right)_i\right| &\leqslant \frac{1}{h}\bar{k}\left[|a_0| + |a_1||x_h^{\max} - x_0| + \cdots + |a_p||x_h^{\max} - x_0|^p\right] \\ &\leqslant \frac{1}{h}\bar{k}\left[|a_0| + |a_1 u^{\max}|\, h + \cdots + |a_p (u^{\max})^p|\, h^p\right] \\ &\leqslant \frac{C}{h}, \quad \text{for } n \text{ large enough.}\end{aligned}$$

$$\implies \left|P_{1,m,x_0}^+(R - \tilde{R})\right| \leqslant \frac{C}{h}\left\|\left(\frac{\mathrm{X}_x^T W_{x,m}^s \mathrm{X}_x}{nh}\right)^{-1}\right\| \sup_{i \in D_m^+}\left|(\hat{\epsilon}_i^s)^2 - \epsilon_i^2\right|$$

By the convergence of $P_{1,m,x_0}^+ \tilde{R}$ and the continuous mapping theorem, there exists a $(p+1) \times (p+1)$ positive definite matrix M such that for all $\delta > 0$,

$$\mathbb{P}\left(\left\|\left(\frac{\mathrm{X}_x^T W_{x,m}^s \mathrm{X}_x}{nh}\right)^{-1} - M^{-1}\right\| > \delta\right) \to 0$$

Hence,

$$\begin{aligned}\implies \left|P_{1,m,x_0}^+(R - \tilde{R})\right| &\leqslant \frac{C}{h}\left(\left\|M^{-1}\right\| + o_p(1)\right) \sup_{i \in D_m^+}\left|(\hat{\epsilon}_i^s)^2 - \epsilon_i^2\right| \\ &= \left[\left(\frac{(\log n)^{1/3}}{n^{1/3} h}\right)^{3/2} + h^p\right] O_p(1) \quad \text{a.s.}\end{aligned}$$

From assumption 7.5 (3), $hn^{1/3}(\log n)^{-1/3} \to \infty$, and from assumption 7.4 (7), $h \to 0$. Hence, $\left|P_{1,m,x_0}^+(R - \tilde{R})\right| \xrightarrow{p} 0$. The proof of the convergence of $\hat{\sigma}^2(x_0, z^m)^{\uparrow}$ is analogous.

B.4.3 Theorem 7.9:

Analogously to the proof of theorem 7.6,

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} > c_\lambda\right) = \mathbb{P}\left(\sqrt{nh}\left(\frac{\hat{\theta} - \theta - \mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right) - \frac{c_\lambda\left(\sqrt{\hat{\mathcal{V}}_n} - \sqrt{\mathcal{V}_n}\right)}{\sqrt{\mathcal{V}_n}} + \frac{\sqrt{nh}\mathcal{B}_n}{\sqrt{\mathcal{V}_n}} > \right.$$
$$\left. > c_\lambda - \frac{\sqrt{nh}\theta}{\sqrt{\mathcal{V}_n}}\right).$$

From theorem 7.8 and the continuous mapping theorem, $\sqrt{\hat{\mathcal{V}}_n} - \sqrt{\mathcal{V}_n} \xrightarrow{p} 0$. Moreover, if $\sqrt{nh}h^{p+1} \to 0$, $\sqrt{nh}\mathcal{B}^+_{m,n} \to 0$ and $\sqrt{nh}\mathcal{B}^-_{m,n} \to 0$. Hence $\sqrt{nh}\mathcal{B}_n \to 0$. Hence, by theorems 7.7 and Slutsky's, $\sqrt{nh}\left(\frac{\hat{\theta}-\theta-\mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right) - \frac{c_\lambda\left(\sqrt{\hat{\mathcal{V}}_n}-\sqrt{\mathcal{V}_n}\right)}{\sqrt{\mathcal{V}_n}} + \frac{\sqrt{nh}\mathcal{B}_n}{\sqrt{\mathcal{V}_n}} \xrightarrow{d} \mathcal{N}(0,1)$. Under H_0, $\theta = 0$, and the first result follows immediately. Under H_1, since $h \to 0$, $\mathcal{V}_n \to \mathcal{V}$, and therefore $-\frac{\sqrt{nh}\theta}{\sqrt{\mathcal{V}_n}} \to -\infty$, from which the second result follows.

Under the alternatives $\theta/\sqrt{nh}$, observe that

$$\mathbb{P}\left(\sqrt{nh}\frac{\hat{\theta}}{\sqrt{\hat{\mathcal{V}}_n}} > c_\lambda\right) = \mathbb{P}\Bigg(\sqrt{nh}\left(\frac{\hat{\theta}-\theta/\sqrt{nh}-\mathcal{B}_n}{\sqrt{\mathcal{V}_n}}\right) - \frac{c_\lambda\left(\sqrt{\hat{\mathcal{V}}_n}-\sqrt{\mathcal{V}_n}\right)}{\sqrt{\mathcal{V}_n}} + \frac{\sqrt{nh}\mathcal{B}_n}{\sqrt{\mathcal{V}_n}} + \theta\left(\frac{1}{\sqrt{\mathcal{V}_n}} - \frac{1}{\sqrt{\mathcal{V}}}\right) > c_\lambda - \frac{\theta}{\sqrt{\mathcal{V}}}\Bigg).$$

and since $\theta\left(\frac{1}{\sqrt{\mathcal{V}_n}} - \frac{1}{\sqrt{\mathcal{V}}}\right) \xrightarrow{p} 0$, by Slutsky's theorem the third result of the theorem follows.

B.5 Theorems when X is discrete

B.5.1 Theorem 8.2:

Let $\delta := (\beta, \gamma^T)^T$, and $\hat{\delta} = (\mathrm{W}^T D\mathrm{W})^{-1}\mathrm{W}^T D\mathrm{Y} =: (\hat{\beta}, \hat{\gamma}^T)^T$, then

$$\sqrt{n}(\hat{\delta} - \delta) = \left(\frac{1}{n}\sum_{i=1}^n W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\})\right)^{-1} \sqrt{n}\frac{1}{n}\sum_{i=1}^n W_i \epsilon_i \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\})$$

Assumption 8.3 guarantees that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^n W_i \epsilon_i \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\}) \xrightarrow{d} \mathcal{N}(0, \sigma_\epsilon^2\, \mathbb{E}[W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\})]),$$

and that

$$\frac{1}{n}\sum_{i=1}^n W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\}) \xrightarrow{p} \mathbb{E}[W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\})].$$

The continuous mapping theorem and Slutsky's theorem then guarantee that

$$\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\epsilon^2 \mathbb{E}(W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+\backslash\{x_0\}))^{-1}\right).$$

By the partitioned inverse, let $[\cdot]_{i,j}$ denote the element in row i and column j in the matrix "$\cdot$". It is direct to derive:

1. $\hat{\beta} = (\mathrm{X}^T D\mathrm{X} - \mathrm{X}^T D\mathrm{Z}(\mathrm{Z}^T D\mathrm{Z})^{-1}\mathrm{Z}^T\mathrm{X})^{-1} = \hat{\theta}$
2. $\left[\mathbb{E}(W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+ \backslash \{x_0\}))^{-1}\right]_{1,1} = \Big(\mathbb{E}(W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+))^{-1} -$
$$- \mathbb{E}(X_i Z_i^T \mathbf{1}(X_i \in \mathcal{N}^+)) \cdot \mathbb{E}(Z_i Z_i^T \mathbf{1}(X_i \in \mathcal{N}^+))^{-1} \mathbb{E}(Z_i X_i^T \mathbf{1}(X_i \in \mathcal{N}^+))\Big)^{-1}$$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\epsilon^2 \mathbb{E}(W_i W_i^T \mathbf{1}(X_i \in \mathcal{N}^+))^{-1}\right).$$

which completes the proof of $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$. The convergence in probability of $\hat{\sigma}_\epsilon^2$ follows form the convergence of the OLS variance estimator, and the convergence of $\hat{V}$ follows from the partitioned inverse reasoning applied to the proof of the convergence of the OLS variance estimator.

B.5.2 Theorem 8.3:

$$\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}} \text{ or } \hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) =$$
$$= \mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) + \mathbb{P}\left(\hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right).$$

$$\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) = \mathbb{P}\left(\sqrt{n}(\hat{\theta} - \theta) + c_{\lambda/2}(\sqrt{\hat{V}} - \sqrt{V}) < \sqrt{n}(b_L - \theta) - c_{\lambda/2}\sqrt{V}\right) \tag{B.8}$$

By theorem 8.2, the continuous mapping theorem and Slutsky's theorem, $\sqrt{n}(\hat{\theta} - \theta) + c_{\lambda/2}(\sqrt{\hat{V}} - \sqrt{V}) \xrightarrow{d} \mathcal{N}(0, V)$ Under H_0, $\theta < b_L$. For all $\varepsilon > 0$, let $z_\varepsilon = \Phi^{-1}(\varepsilon)$, then there exist $n_0 \in \mathbb{N}$ such that $n > n_0$ implies $\sqrt{n}(b_L - \theta)/\sqrt{V} - c_{\lambda/2} < z_\varepsilon$, and therefore, (B.8) $< \varepsilon$. Hence, $\mathbb{P}\left(\hat{\theta} < b_L - c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) \xrightarrow[n\to\infty]{} 0$. The proof that $\mathbb{P}\left(\hat{\theta} > b_U + c_{\lambda/2}\sqrt{\frac{\hat{V}}{n}}\right) \xrightarrow[n\to\infty]{} 0$ is similar, and so are the proofs to the other parts of the theorem.

Appendix C

Empirical Appendix

The following table contains the frequencies of data per number of cigarettes smoked daily. 80% of the observations did not smoke, and 19% smoke from one to 20 cigarettes a day. Hence 94% of the smoking observations smoke up to 20 cigarettes, and 60% smoke up to 10. There is a high concentration of observations at $CIG = 5, 10, 15, 25, \ldots$, but it is not immediate to determine whether this is the result of rounding on reporting, which would induce measurement error in the variable CIG, or actual higher frequencies of smoking in multiples of 5. Given that a common pack of cigarettes contains 20 cigarettes, at least part of the higher frequencies at the 5's may be due to a preference for consuming cigarettes in quarter pack units.

Section 11 does not account for measurement error in the variable CIG. However, if the distribution of the measurement error conditional on CIG and Z is discontinuous at $CIG = 0$, the discontinuity test will detect its presence.

Table C.1: **Data Frequency**

CIG	**Frequency**	**Percent**	**Cumulative**
0	393,939	80.70	80.70
1	1,469	0.30	81.00
2	2,986	0.61	81.61
3	3,759	0.77	82.38
4	2,890	0.59	82.98
5	6,838	1.40	84.38
6	2,618	0.54	84.91
7	1,758	0.36	85.27
8	1,644	0.34	85.61
9	335	0.07	85.68
10	32,720	6.70	92.38

CIG	Frequency	Percent	Cumulative
11	117	0.02	92.41
12	801	0.16	92.57
13	396	0.08	92.65
14	128	0.03	92.68
15	4,568	0.94	93.61
16	93	0.02	93.63
17	39	0.01	93.64
18	259	0.05	93.69
19	19	0.00	93.70
20	25,333	5.19	98.89
21	49	0.01	98.90
22	30	0.01	98.90
23	39	0.01	98.91
24	36	0.01	98.92
25	417	0.09	99.00
26	7	0.00	99.01
27	3	0.00	99.01
28	12	0.00	99.01
29	2	0.00	99.01
30	2,993	0.61	99.62
31	4	0.00	99.62
32	3	0.00	99.62
33	2	0.00	99.62
34	5	0.00	99.62
35	97	0.02	99.64
36	5	0.00	99.65
37	2	0.00	99.65
38	1	0.00	99.65
39	0	0.00	99.65
40	1,474	0.30	99.95
> 40	254	0.05	100.00
Total	488,144		