Nonseparable Models With Endogeneity and Sample Selection

By

Jen-Che Liao

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

University of Wisconsin-Madison

2013

Date of final oral examination: 4/30/13

The dissertation is approved by the following members of the Final Oral Committee:

Jack Porter, Professor of Economics, Chair
Bruce Hansen, Professor of Economics, Co-chair
Christopher Taber, Professor of Econmics
Xiaoxia Shi, Assistant Professor of Economics
Chunming Zhang, Professor of Statistics

Acknowledgments

My six-year life-changing journey during which I have grown and learned a tremendous amount is coming to an end. I have benefited substantially from the guidance of and interaction with many faculty and colleagues during my years at UW-Madison.

My biggest debt of gratitude comes to Jack Porter and Bruce Hansen, the best advisors I could have ever hoped for. They have always been generous with their time, support, and valuable and honest advice at all stages of my thesis research. I am sure that I will keep looking to my advisors as role models, aspiring me to bring the same level of enthusiasm, rigor, and insights to my own research. I also thank Chris Taber, Andrés Aradillas-López, and Xiaoxia Shi for helpful comments and feedback. Having been part of such a high quality group of econometric faculty motivates me to keep pushing forward. I have also benefited from discussions with many other professors. Among them were Hidehiko Ichimura, Rosa Matzkin, Arthur Lewbel, and Chunming Zhang.

It has been a great honor sharing these years from the very beginning in the program with my Taiwanese fellow students Ying-Ying Lee, Hsueh-Hsiang (Cher) Li, Chu-An Liu, and Cheng-Ying (Anita) Yang. I have been fortunate to have them along during my graduate study at UW-Madison, especially when achieving important milestones starting from "Ph.A." through "Ph.D." Many others provided a fruitful source of economic and econometric insights and questions. Among them were Andrew (Drew) Anderson, Laura Dague, Shengjie Hong, SeoJeong (Jay) Lee, Enrique Pinzon Garcia, Nelson Ramirez Rondan, Mai Seki, Naoya Sueishi, Jing Tao, and Jin Yan. I would also like to thank my badminton group partners, Atsuko Tanaka, Kittichai Saelee, and Hsuan-Chih (Luke) Lin. Playing badminton on a weekly basis helps make my life at UW-Madison more balanced and enjoyable.

Finally, I want to thank my family for standing behind me in all my endeavour. My wife, Mei-Shin, has always been supportive through my ups and downs and has dealt with my struggles while working and taking care of kids. Thanks for her sacrifice, patience, and love. Having two beloved kids, Heng-An and Yu-Fong, has been bringing tremendous happiness to my life. Although they often had to endure their dad's absence from their lives they seldom complained. But I know they have been hoping for my return for a long time. I delicate this dissertation to my dear family.

Contents

Page

Appendix

Page

List of Tables

List of Figures

Abstract

Nonseparable models are not additively separable in unobserved heterogeneity and therefore allow responses to policy interventions to vary across individuals with identical observed characteristics. This dissertation is a collection of three essays, each contributing to a different aspect of nonseparable econometrics models with endogeneity and sample selection.

Binary response is a very important special case of nonseparable models, as it has many applications. In the first chapter, we consider a triangular simultaneous equations model with a binary outcome that is identified under a weak quantile restriction which allows for general forms of heteroskedasticity. The proposed two-step estimation procedure combines Horowitz's (1992) smoothed maximum score estimator in semiparametric binary response models with a control function approach to the endogeneity problem. Rates of convergence and the asymptotic distribution are derived. In a simulation study, we present the finite-sample performance of the estimator and illustrate advantages of the proposed approach by comparing with other alternatives.

The second chapter provides an application of the methodology developed in Chapter 1 to an empirical context of female labor market participation with endogenous non-labor income. Using the data set extracted from the 2011 March Supplement to the US Current Population Survey, we find that, qualitatively similar to the probit estimates, accounting for endogeneity leads to a substantial increase in the magnitude of the non-labor income coefficient, being 62% ~ 77% larger than that in the smoothed maximum score estimation. The coefficient estimates for different quantiles are considerably different, implying that

strong full conditional independence may fail or heteroskedasticity may be present in the data set considered.

In the third chapter, we discuss what features of sample selection models without imposing additivity can be identified under various restrictions. We focus on a nonparametric nonseparable sample selection model with possibly endogenous regressors. Using a control function approach, we provide identification results and develop a three-step nonparametric estimator for the average structural function given selection. Convergence rates are derived. A simulation study compares the numerical performance of the proposed estimator with Das, Newey, and Vella (2003) estimator and Heckman's two-step estimator under correct specifications and misspecifications.

Chapter 1

Triangular Models of Binary Response under Quantile Restrictions

1.1 Introduction

The triangular simultaneous equations model has long been an important special class of structural models, with a recent interest in generalizing identification in linear triangular models to nonparametric setups under weak assumptions. This chapter considers identification and estimation of a triangular simultaneous equations model with a binary outcome under a weak nonparametric quantile restriction, allowing for both endogeneity and heteroskedasticity of general forms while being less susceptible to misspecification. The binary response model considered here generalizes the types of quantile restrictions pioneered by Manski (1975, 1985) and Horowitz (1992) to allow for the presence of (continuous) endogenous regressors.[1,2] Using a triangular model that motivates the use of a control function to correct for endogeneity, the proposed estimation method can be viewed as a partially linear version of the maximum score type of estimators with a nonparametrically generated regressor.

[1]The possibility of using median restrictions in endogenous binary response models was first introduced to the literature by Blundell and Powell (2004) and later employed by Hoderlein (2009) and Krief (2011). More discussions on this point follow shortly.

[2]Another important but fundamentally different setting is nonlinear models with endogenous regressors that are discretely distributed. Several recent papers have considered identification and estimation of such models, including, for example, Abrevaya, Hausman, and Khan (2010), Shaikh and Vytlacil (2011), and Vytlacil and Yildiz (2007).

There are numerous possible applications in economics that fit into the framework of heteroskedastic binary response models with endogenous regressors. One example that motivates this research and provides an application of this chapter concerns estimating the income effect on labor market participation, where non-labor income is likely to be endogenous (Blundell and Powell (2004)). In addition, accommodating (possibly arbitrary) heteroskedasticity in the choice model may be important in some applications because it allows for, for example, the random coefficients model, which is an important tool to model unobserved heterogeneity. Moreover, heteroskedasticity in the first-stage equation for endogenous income may also arise since the conditional variance of the income variable may vary with spouse's education, which serves as an instrumental variable. Another example of the framework is the impact of endogenous health assessments on the labor supply of older men, as studied by Maurer, Klein, and Vella (2011). A third example is consumer choice models with endogenous prices, in which the source of endogeneity comes from the fact that prices are typically correlated with the unobserved product characteristics.

The problems of endogeneity, heteroskedasticity, and nonnormality/asymmetry frequently arise in analysing microeconometric data. It appears essential to address these issues simultaneously. To do so we present a solution to the endogeneity problem in estimating underlying regression coefficients in semiparametric binary response models under weak nonparametric quantile restrictions. Our two-step semiparametric estimation procedure generalizes Horowitz's (1992) smoothed maximum score estimator to incorporate endogenous regressors. The first step consists of nonparametric estimation of the control variable, which is constructed as the conditional cumulative distribution function (CDF) of the endogenous variable given the instruments, as suggested by Imbens and Newey (2009) for a nonseparable first-stage model. In the second step, a series method is used to simultaneously estimate the finite-dimensional parameters and the nonparametric conditional quantile function using a generated control variable (as the argument) that has been estimated nonparametrically in the first stage.

One important feature of our new method is imposition of a weak quantile restriction. Using a median restriction as a special case, as first pointed out by Blundell and Powell (2004), one can identify and estimate the regression index coefficients of primary interest in an endogenous binary response model. This median restriction has recently been employed by Hoderlein (2009) and Krief (2011). However, this chapter is different from the aforementioned existing approaches in the following substantial ways. First, the method proposed here exploits general quantile restrictions, and as a result, this article's estimator could be used to develop a test procedure for the conditional independence assumption by comparing quantile coefficients on different parts of the conditional distribution. Second, instead of specifying the separable nonparametric (Hoderlein (2009)) or linear (Krief (2011)) first-stage equation, the first-stage model considered here is nonparametric and nonseparable in the unobservable, which motivates the use of the conditional CDF of the endogenous regressor given instruments as a control variable. Third, to the best of our knowledge, the proposed procedure is the first series estimator in the context of binary response models with endogeneity. The series method is particularly useful in our setting due to its convenience in imposing additive separability that arises naturally here through the control function approach.[3] Series also allow a straightforward extension to nonparametric models, as will be discussed in Section 1.6. Finally, unlike Blundell and Powell (2004) and Krief (2011) who propose kernel-weighted methods to eliminate the control function, we employ series methods to estimate the control function that may be useful for constructing tests for model specification such as endogeneity. On the other hand, our estimation procedure is also related to those using a series approximation to the infinite-dimensional parameter with generated regressors (Newey, Powell, and Vella (1999), Lee (2007), Newey (2009), and Imbens and Newey (2009)[4]).

[3]For the estimation of additive models, series methods have some advantages compared to the alternative kernel estimation methods, see Li (2000) for more discussions on this comparison.

[4]The framework of Imbens and Newey (2009) and Newey, Powell, and Vella (1999) is nonparametric mean regression models with nonparametrically generated regressors. Newey (2009) and Lee (2007) consider semiparametric mean and quantile regression models, respectively, with parametrically generated regressors.

As will be discussed further in Section 1.4, under quantile restrictions, the model considered has a *semiparametric partially linear* structure with the nonparametrically generated regressor in the binary response framework. From a technical perspective, the main complication for deriving statistical properties of our two-step estimator originates from the fact that the proposed estimation method involves two types of infinite-dimensional parameters: an unknown conditional quantile function that has to be nonparametrically estimated using a nonparametrically generated regressor. Estimation with generated regressors has received considerable attention recently in the econometrics literature (e.g., Blundell and Powell (2004); Blundell and Horowitz (2007), Hahn and Ridder (2010), Lee (2007), Li and Wooldridge (2002), Mammen, Rothe, and Schienle (2011, 2012), Newey, Powell, and Vella (1999), Rothe (2009), Song (2008), Sperlich (2009), Su and Ullah (2008), among others). The central issue of the generated regressor problem is to account for the contribution of the preliminary estimation uncertainty on the asymptotic behavior of the final estimator. To do this, we establish a stochastic expansion accounting for the nonparametric estimation error stemming from estimating the unknown conditional quantile function and the generated regressor. This result facilitates characterizing the roles played by the nonparametrically generated regressor. To be precise, using the stochastic expansion derived in the chapter and smoothing out the (discontinuous) objective function, we expand the first-order condition solved by the estimate of the parameter of interest and investigate the asymptotic behavior of each term in the expansion.

The literature on estimation of models with endogeneity consists of two main approaches: instrumental variables (IV) and control function (CF) methods.[5] For a general overview, see Blundell and Powell (2003) and Matzkin (2007). The hard problem with the nonparametric IV approach is what is known as the ill-posed inverse problem, which leads to difficulty in estimation and inference.[6] Furthermore, for the discrete response model, which can be viewed as

[5]In linear models, the IV method is the most common approach to deal with endogeneity. In fact, there are equivalences between the approaches between IV, CF, and two-stage least square in linear models.

[6]Several regularization procedures such as Tikhonov regularization and series truncation have been proposed to overcome this difficulty. See Horowitz (2011) for an excellent illustration of the ill-posed inverse problem and a summary of recent literature.

a class of nonseparable models without strict monotonicity, the IV approach (assuming that the instruments are independent of the structural error) is in general not sufficient to achieve point identification without imposing further assumptions, as set out in Chesher (2010). In a series of recent papers, Chesher and his coauthors (Chesher (2009, 2011); Chesher, Rosen, and Smolinski (2011)) have provided partial identification results using a single-equation IV approach in the binary response setting. Hong and Tamer (2003) and Shaikh and Vytlacil (2008) consider identification and estimation under IV median restrictions that the latent disturbance is median independent of the instruments. Alternatively, one can address endogeneity concerns in the context of binary response models by the use of the CF approach.[7] Within this literature, in addition to Hoderlein (2009) and Krief (2011) discussed above, Blundell and Powell (2004) propose a leading estimator by using a pairwise-differencing (or matching) estimation procedure. More recently, Rothe (2009) develops a semiparametric maximum likelihood estimator, which is an extension of Klein and Spady (1993) to allow for endogeneity. These two kernel-based estimators are $\sqrt{n}$-consistent and asymptotically normal, but impose the strong restriction that the structural error is fully independent of the endogenous regressors conditional on the control variable. None of these are robust to the presence of general forms of heteroskedasticity in the binary outcome equation. In the context of nonseparable triangular models with a continuous outcome, Chesher (2003) uses a local version of the quantile control function restriction. Another strand of literature concerning both endogeneity and heteroskedasticity in the binary response models is the special regressor method proposed by Lewbel (2000). See Lewbel (2012) for a recent overview of the method and its applications and Lewbel, Dong, and Yang (2012) for a comparison to other types of estimators such as CF, maximum likelihood, and linear probability models.

[7]There have been many papers using the CF approach to deal with endogeneity in a variety of econometric models, including nonparametric separable and nonseparable triangular simultaneous equations models (Newey, Powell, and Vella (1999); Imbens and Newey (2009)), nonparametric sample selection models (Das, Newey, and Vella (2003)), semiparametric censored/uncensored quantile regression models (Blundell and Powell (2007); Lee (2007); Chernozhukov, Fernandez-Val, and Kowalski (2011)), among others. Kim and Petrin (2010) propose a hybrid of CF and IV in the nonparametric additive model.

The remainder of the chapter is organized as follows. Section 1.2 specifies a semiparametric triangular model of binary response and discusses the essential stochastic restrictions employed in this chapter as the identifying assumption. Section 1.3 considers identification of the model. Section 1.4 describes the proposed two-stage series control function estimator. Asymptotic properties of the proposed estimator are derived in Section 3.7, including consistency, rates of convergence, asymptotic normality, and more interestingly, how the first-stage estimation error of the generated regressor affects the asymptotic distribution of the final estimators. An extension to a fully nonparametric binary response model with endogeneity is discussed in Section 1.6. Section 3.8 reports the results of a simulation study. Section 3.9 presents concluding comments with suggestions for future research. The appendix contains the proofs of theorems and intermediate lemmas.

1.2 The Model

We begin with the structural relationship between a binary response variable Y and a vector of (possibly endogenous) explanatory variables X of the form

$$Y^* = m_0(X, U) = X'\gamma_0 + U, \tag{1.1}$$

$$Y = \mathbf{1}\{Y^* > 0\}, \tag{1.2}$$

where $\mathbf{1}\{\cdot\}$ denotes the usual indicator function that equals one if its argument is true and zero otherwise, Y is a scalar binary variable indicating the sign of the latent variable Y^*. The latent variable Y^* is determined by an unknown nonseparable structural function m_0 of a vector of observable covariates $X \in \mathbb{R}^{d_x}$ and an unobserved disturbance U. The structural function m_0 is assumed to be strictly increasing in its second argument U.

In this chapter, we consider a semiparametric single-index specification that is an important special case of the general latent outcome equation: $Y^* = m_0(X, U) = X'\gamma_0 + U$, where γ_0 is a $d_x \times 1$ dimensional vector of unknown parameters.[8]

[8] In Section 1.6, we further discuss a nonparametric separable specification: $m_0(X, U) = g_0(X) + U$ as a possible extension of our estimation procedure. Moreover, there are two other special cases that have

Let $X = (X_1, Z_1')'$, where X_1 is a continuously distributed, scalar endogenous explanatory variable[9] in the sense that X_1 is possibly correlated with U. The endogenous regressor X_1 is assumed to be determined by the following first-stage model

$$X_1 = h_0(Z, \eta), \qquad \eta \perp\!\!\!\perp Z, \tag{1.3}$$

where h_0 is a real-valued unknown function and is assumed to be strictly monotonic in a scalar disturbance η and $Z = (Z_1', Z_2')' \in \mathbb{R}^{d_z}$ are exogenous instruments.

To complete the model, we need to impose a crucial stochastic restriction that is the key to identification of the model. Throughout the chapter, we denote $Q_\tau(\cdot)$ as the τth quantile of the conditional distribution in question, e.g., $Q_\tau(Y \mid X = x) = F^{-1}_{Y|X=x}(\tau) \equiv \inf\{y : F_{Y|X=x}(y) \geq \tau\}$ is the τth quantile of a conditional CDF $F_{Y|X}(\cdot)$. Following the idea employed by Imbens and Newey (2009) to construct the control variable in nonseparable triangular simultaneous equations models with a nonseparable first-stage equation (1.3), it is easy to show that, under the strict monotonicity of h_0 in η,

$$V \equiv F_{X_1|Z}(X_1 \mid Z) = F_\eta(\eta).$$

Since V is a one-to-one function of η, conditioning on η is equivalent to conditioning on V. Then the stochastic restriction considered here is that the conditional distribution function of U given X and Z, $F_{U|X,Z}$, must satisfy the following quantile exclusion restrictions (QER):

$$Q_\tau(U \mid X, Z) = Q_\tau(U \mid X, V) = Q_\tau(U \mid V) \equiv \lambda_\tau(V) \qquad \text{a.s. for some known } \tau \in (0,1), \tag{1.4}$$

where $\lambda_\tau(\cdot)$ is a real-valued unknown function referring to the τth conditional quantile function of U given V. The first equality in (1.4) follows from the usual control function approach

attracted much attention among econometricians and statisticians but we do not pursue in this chapter. The first one is nonparametric fully additive models of the form: $m_0(X, U) = m_1(X_1) + \cdots + m_{d_x}(X_{d_x}) + U$, where m_j's are unknown functions. The second is the additive partially linear model: $m_0(X, U) = X_l'\gamma_0 + \sum_{j=l+1}^{d_x} m_{j0}(X_j)$, where $X = (X_l', X_{l+1}, \ldots, X_{d_x})'$. The latter is particularly useful when X contains discrete-valued covariates. See, for example, Li (2000) for a further discussion.

[9]The estimator developed below can be extended to a multivariate endogenous variable case by specifying a reduce-form equation for each endogenous variable.

that conditional on Z, the endogenous variable X_1 varies only with V, which also implies $Q_\tau(U \mid Z, V) = Q_\tau(U \mid X, V)$.[10] As a result, equation (1.4) explicitly says that conditional on V, X and U are independent at τth quantile and hence justifies V as a control variable. The function $\lambda_\tau(\cdot)$ is left unspecified to reflect the fact that it is difficult to have a correct specification of the functional form of the stochastic relationship between the unobserved components U and V conditional on Z. If $\tau = .5$, equation (1.4) reduces to median exclusion restrictions. As will be discussed in the next section, the restrictions (1.4) imply a set of moment conditions that provide a basis for inference on β_0.

The triangular system (1.1)-(1.3)[11] can be viewed as a binary outcome variant of general triangular models without additivity studied by Imbens and Newey (2009). In comparison with the existing triangular semiparametric binary response models considered by Blundell and Powell (2004), Rothe (2009), Hoderlein (2009), and Krief (2011), equation (1.3) is a general nonparametric nonseparable model that allows the instruments Z to exert the influence on X_1 in flexible ways. For example, it allows for potential missspecification like non-linearity and heteroskedasticity that are commonly encountered in practice, as further discussed below.

CF and IV Stochastic Restrictions. We compare the QER (1.4) with other alternative stochastic restrictions that have been adopted to address endogeneity problems in the literature. First of all, the leading form of stochastic restrictions imposed in endogenous binary response models is the distributional exclusion restriction (DER), requiring that U be (fully) independent of X (or Z) conditional on η (Blundell and Powell (2004) and Rothe (2009)). Namely,

$$U \mid X, Z \overset{d}{\sim} U \mid X, \eta \overset{d}{\sim} U \mid Z, \eta \overset{d}{\sim} U \mid \eta, \tag{1.5}$$

where the symbol $\overset{d}{\sim}$ denotes equality of conditional distributions. The first two equalities follow if there exists a strictly monotonic function C such that $\eta = C(X, Z)$, implying there

[10] Equivalently, since V is one-dimensional, strictly increasing in both X_1 and Z, there is a one-to-one mapping between (X, Z), (Z, V), and (X, V).

[11] Equations (1.2) and (1.3) form a triangular structure in the sense that Y is absent in the structural equation in X_1.

is a one-to-one mapping between (X, Z), (X, η), and (Z, η). The last equality is a key conditional independence assumption. Restriction (1.5) is weaker than joint independence (JI) of all unobservable errors and the instruments

$$(U, \eta) \perp\!\!\!\perp Z, \tag{1.6}$$

because (1.5) allows for dependence of η on Z. On the other hand, compared to the IV (or the instrument independence) assumption:

$$U \perp\!\!\!\perp Z, \tag{1.7}$$

it is clear that the DER (1.5) is not more or less general than IV (1.7).[13] However, imposing (1.7) only is generally not sufficient to identify the endogenous binary response model, as pointed out in Section 3.1. In addition, it should be emphasized that the DER (1.5) seems somewhat restrictive since it implies a key requirement underlying the CF approach: the source of endogeneity resulting from the correlation between X_1 and U is only through their joint dependence on the first-stage error η. This restriction rules out, for example, direct dependence between X and U, and motivates us to consider the weaker assumption QER in the present chapter.

Compared to JI and DER, QER places weaker restrictions on the relation between (X, Z) and the distribution of U. The DER is equivalent to QER for *all* quantiles and not just one. As a result, it is clear that QER allows for not only endogeneity but also general forms of unknown heteroskedasticity. This generality may be important in some econometric applications. For example, QER permits a form of heteroskedasticity resulting from random coefficients, where the random coefficients characterize an individual's unobserved heterogeneity in preference or tastes regarding the attributes of the corresponding variables.

[12] This joint independence assumption is used in Florens, Heckman, Meghir, and Vytlacil (2008) and Imbens and Newey (2009) who deal with nonseparable triangular simultaneous equations models using the CF approach.

[13] This is the main reason why the assumptions underlying the CF are fundamentally different from those underlying IV, unless one is willing to impose the stronger joint independence between the instruments and unobservables, i.e. (1.6). Note that the rather strong JI encompasses the DER and IV as special cases. In that case, one may prefer using the IV approach because the CF method requires specifying the first-stage equation for the endogenous regressor that may suffer from misspecification.

Accounting for unobserved preference and taste variation is important in many empirical applications, particularly in labor supply and consumer demand analysis. See Pacifico (2012) and Horowitz (2009, Chapter 4) for related discussions. Nevertheless, it is worth mentioning that there is one important advantage of imposing the stronger DER: identifying coefficients γ_0 enables one to recover the probabilities and marginal effects of interest since DER implies a semiparametric single-index restriction of the form

$$\mathbb{E}(Y \mid X, \eta) = \mathbb{E}(\mathbf{1}\{U \leq X'\gamma_0\} \mid X, \eta) = \mathbb{E}(\mathbf{1}\{U \leq X'\gamma_0\} \mid \eta) = G_0(X'\gamma_0, \eta),$$

where G_0 is the conditional distribution function of U given η.[14] On the other hand, based on the same QER as (1.4), Blundell and Powell (2007) and Lee (2007) consider estimation of censored and uncensored the quantile regression models with endogeneity, respectively.

Example 1.1. (**Heterogeneous return to schooling in labor force participation with endogenous schooling**) For an economic example of our model, consider a random coefficients triangular binary response model: $Y = \mathbf{1}\{S\gamma(\varepsilon_1)+\varepsilon_2 \geq 0\}$ and $S = Z'\pi+\eta$, where Y is the indicator of employment status: $Y = 1$ if employed and $Y = 0$ otherwise, S is schooling, ε_2 is an unobserved random variable, Z is a set of instruments, and η is scalar individual ability. The random coefficient $\gamma(\varepsilon_1)$ is of the form: $\gamma(\varepsilon_1) = \bar{\gamma} + \varepsilon_1$ with the mean or median $\bar{\gamma}$ of the distribution of γ and an unobserved random variable ε_1. Then the model can be further written as $Y = \mathbf{1}\{S(\bar{\gamma} + \varepsilon_1) + \varepsilon_2 \geq 0\} = \mathbf{1}\{S\bar{\gamma} + (Z'\pi + \eta)\varepsilon_1 + \varepsilon_2 \geq 0\} \equiv \mathbf{1}\{S\bar{\gamma} + U \geq 0\}$, where $U \equiv (Z'\pi + \eta)\varepsilon_1 + \varepsilon_2$. In general, S is endogenous since S is correlated with U through η. Also, U is not independent of Z even when conditional on individual ability η and hence heteroskedastic. It also implies that assumptions JI, IV, and DER do not hold in this random coefficients context. In contrast, estimating $\bar{\gamma}$ under QER allows for the presence of this type of heteroskedasticity and endogeneity. Moreover, QER leads to modelling unobserved heterogeneity η nonparametrically, in the sense that it does not require one to assume parametric distributions for η and ε_2. Modelling unobserved heterogeneity in

[14] This restriction is the one on which the estimators of Blundell and Powell (2004) and Rothe (2009) base. Blundell and Powell (2004) also require that the function G_0 be monotonic in the first argument $X'\gamma_0$.

a flexible way may be important in some empirical applications, see, for example, Pacifico (2012) and Haan (2006) for more discussions on this point for discrete choice models of labor supply.

We use the following example adapted from Hoderlein (2009) to illustrate how the QER can be motivated.

Example 1.2. (**Motivating QER**) Consider a model of heterogeneous return to schooling in labor force participation with endogenous schooling: $Y = \mathbf{1}\{S \cdot \beta_1(P) + A \cdot \beta_2(P) \geq 0\}$ and $S = Z'\pi + A$, where Y is employment; S is schooling; A is ability; P is preferences for work; Z are instruments; β_1 and β_2 are random coefficients. This model implies $Y = \mathbf{1}\{S\bar{\beta}_1 + (Z'\pi + A)(\beta_1(P) - \bar{\beta}_1) + A\beta_2(P) \geq 0\}$ where $\bar{\beta}_1$ is the mean of $\beta_1(P)$. Suppose a strong but economically plausible assumption that we have instruments Z that satisfy $Z \perp\!\!\!\perp (A, P)$, implying $\mathbb{E}(A\beta_2(P) \mid S, A) \equiv \lambda(A)$. If we further assume that preference for work P is independent of all economic variables in the model, i.e., $P \perp\!\!\!\perp (S, Z, A)$, then we have $\mathbb{E}(\beta_1(P) \mid S, A) = \bar{\beta}_1$. Combining these two assumptions yields $\mathbb{E}[\underbrace{S \cdot (\beta_1(P) - \bar{\beta}_1) + A\beta_2(P)}_{\equiv U} \mid S, A] = \lambda(A)$. Furthermore, if the conditional distribution $F_{\beta_1(P), A\beta_2(P)|S,A}$ is assumed to be symmetric about $(\bar{\beta}_1, \lambda(A))$, we end up with median exclusion restrictions: $Q_{\tau=0.5}(U \mid S, A) = \lambda(A)$.

We end the discussion on stochastic restrictions by noting that, in contrast to the IV independence assumption, in practice it may be difficult to motivate the CF assumption from economic theory. In this view, QER considered in the chapter may be seen as a way to strike a balance between the lack of point identification from imposing the IV assumption (1.7) and the restrictive assumption DER in the CF approach. We note here that, similar to exogenous binary response models, it is natural to expect that the model implies no relationship between Y and X if, instead of (1.4), we impose the mean exclusion restrictions:

$$\mathbb{E}(U \mid X, Z) = \mathbb{E}(U \mid X, \eta) = \mathbb{E}(U \mid \eta).^{15} \tag{1.8}$$

[15]This type of restrictions is employed by, for example, Newey, Powell, and Vella (1999) in a nonparametric triangular simultaneous equations models. As pointed out by the authors, (1.8) allows for both endogeneity

Nonseparable first-stage Models. We note that there is another major difference between the CF and IV approaches: the CF method requires a first-stage equation for the endogenous variable added to the system of models, whereas IV is a single-equation approach. In this respect, it seems reasonable to specify a flexible first-stage equation to avoid misspecification error. Equation (1.3) is of a general form that encompasses several commonly used specifications as special cases.[16,17] One important class of the first-stage specification assumes h to be additively separable with a conditional mean restriction

$$X_1 = \mathbb{E}(X_1 \mid Z) + \eta, \tag{1.9}$$

where $\mathbb{E}(X_1 \mid Z)$ can be either specified to be linear (e.g., Krief (2011)) or left unspecified (Newey, Powell, and Vella (1999), Blundell and Powell (2004), Rothe (2009), and Hoderlein (2009)). Lee (2007) considers a semiparametric quantile regression version of Newey, Powell, and Vella (1999) with a parametric quantile first-stage equation $X_1 = Z'\pi(\tau) + \eta_\tau$ where the τ conditional quantile of η_τ given Z is assumed to be zero.

The presence of heteroskedasticity in the first-stage model can be a concern in practice. For example, as mentioned in the Introduction, in the empirical data used by Blundell and Powell (2004), the conditional variance of the logarithm of other income variables may vary with the education level of the spouse. In particular, additivity in the first-stage model (1.9)

and heteroskedasticity, and the latter often results from individual heterogeneity in demand functions (Brown and Walker (1989)).

[16]Strictly speaking, scalar heterogeneity and strict monotonicity imposed on equation (1.3) are restrictive from an economic perspective. Kasy (2011) shows that one-dimensionality of the first-stage heterogeneity η is both a necessary and sufficient condition for the existence of control functions that are valid to correct for endogeneity in the triangular simultaneous equations models. This dimensionality requirement is violated by, for example, the first-stage random coefficient model where both the intercept and slope coefficient are random. Interestingly, in his more recent paper, Kasy (2012) further shows that in continuous triangular models under monotonicity in the instrument while dropping restrictions on heterogeneity and functional form, the CF approach is still valid even though conditional independence: $U \perp\!\!\!\perp X \mid V$ fails to hold. Nonexistence of the control function or multi-dimensionality of the first-stage heterogeneity is beyond the scope of the chapter.

[17]Hoderlein and Sasaki (2011) consider a nonseparable triangular model allowing for multiple unobservables in the first stage. They show that the necessary and sufficient condition for identification is monotonicity in an index of potentially vector-valued unobservables, underscoring the importance of monotonicity in endogenous nonseparable models.

would not hold, for example, in the model that exhibits conditional heteroskedasticity of a multiplicative form, i.e.,

$$X_1 = \mathbb{E}(X_1 \mid Z) + \sigma(Z)\eta, \tag{1.10}$$

where $\sigma(Z)$ is a unknown function of the instruments and the source of endogeneity comes from potential correlation between U and the homoskedasticity error η. Then in general QER: $Q_\tau(U \mid X, \eta) = Q_\tau(U \mid \eta)$ does not necessarily imply $Q_\tau(U \mid X, \sigma(Z)\eta) = Q_\tau(U \mid \sigma(Z)\eta)$, meaning that recovering and conditioning on the first-stage heteroskedastic error $\sigma(Z)\eta$ may not be able to correct for endogeneity. In contrast, it is straightforward to observe that, under model (1.10),

$$F_{X_1|Z}(X_1 \mid Z) = F_\eta\left(\frac{X_1 - \mathbb{E}(X_1 \mid Z)}{\sigma(Z)}\right), \tag{1.11}$$

where $F_\eta(\cdot)$ denotes the CDF of η and is assumed to be strictly monotonic. This observation shares the same spirit with the well known propensity score as a control variable in the sample selection literature. Specifically, consider a heteroskedastic sample selection model (Chen and Khan (2003)): $Y = D \times Y^* = D \times (X'\beta_0 + \sigma_2(X)U)$ and $D = \mathbf{1}\{\mu(Z) - \sigma_1(Z)\eta \geq 0\}$, where $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ are unknown scale functions. Then the propensity score defined by

$$\mathrm{PS}(Z) = \mathbb{E}(D \mid Z) = \Pr(D = 1 \mid Z) = F_\eta\left(\frac{\mu(Z)}{\sigma_1(Z)}\right)$$

is robust to heteroskedasticity in the selection equation.

1.3 Identification

In this section, we present the conditions under which the finite-dimensional parameters β_0 in an endogenous binary response model (1.1)-(1.4) can be identified under assumption QER. To begin, note that for any real, monotone increasing function $m(\cdot)$, we have $Q_\tau(m(Y) \mid X) = m(Q_\tau(Y \mid X))$. For the purposes of identification and estimation, we focus in this chapter on the following conditional quantile of the observed dependent variable Y

given the regressors X and instruments Z implied by the model:

$$\begin{aligned} Q_\tau(Y \mid X, Z) &= Q_\tau(\mathbf{1}\{m_0(X,U) > 0\} \mid X, Z) = \mathbf{1}\{Q_\tau(m_0(X,U) \mid X, Z) > 0\} \\ &= \mathbf{1}\{m_0(X, Q_\tau(U \mid X, Z)) > 0\} = \mathbf{1}\{m_0(X, Q_\tau(U \mid V)) > 0\} \\ &= \mathbf{1}\{m_0(X, \lambda_\tau(V)) > 0\} \equiv \mathbf{1}\{m_\tau^*(X,V) > 0\}, \end{aligned} \tag{1.12}$$

where the second, third, and fourth equalities follow from the fact that the indicator is a monotone increasing function, the assumption of monotonicity of m_0 in U, and the QER (1.4), respectively. Note that by definition $m_\tau^*(X,V)$ is the τth quantile of the latent variable Y^* conditional on X and Z (or equivalently on X and V).

In the nonseparable latent variable formulation (1.1), we might want to know the effects of certain covariates on the latent variable Y^*. For example, Y^* can be interpreted as the underlying utility difference in the consumer choice modelling analysis, in which one may be interested in a consumer's willingness to pay for a marginal improvement in an attribute by calculating the ratio of the attribute's coefficient to the price coefficient. In such cases, one structural parameter of interest is the quantile structural function (QSF) defined by Imbens and Newey (2009)[18] as the τth quantile of $m_0(x,U)$. That is, $Q_\tau(m_0(x,U)) = m_0(x, Q_\tau(U)) \equiv m_0(x, q_\tau)$, where q_τ is the τth quantile of the marginal distribution of U.[19]

Since equation (1.12) plays an essential role in subsequent identification results, we briefly talk about identification of m_τ^*. The formal justification for identification of m_τ^* in equation (1.12) implied by models (1.1)-(1.4) is based on Manski (1988) and Matzkin (1992).

Definition Let the function $m_\tau^* : M \to \mathbb{R}$, $M \subset \mathbb{R}^{d_x + d_z}$. Suppose the vector (X, Z) possesses a probability density function inducing a probability measure G and denote the support of

[18] It is also called the quantile structural effect by Chernozhukov, Gagliardini, and Scaillet (2012). Imbens and Newey (2009) discuss identification and estimation of the quantile structural function in a triangular model without additivity via the control function. Chernozhukov, Imbens, and Newey (2007), Horowitz and Lee (2007), and Chernozhukov, Gagliardini, and Scaillet (2012) consider estimation of the QSF in nonseparable models by using the IV approach.

[19] With the QSF(x) evaluated at two different values $\bar{x}_1$ and $\bar{x}_2$ of X, we can interpret the difference $m_0(\bar{x}_1, q_\tau) - m_0(\bar{x}_2, q_\tau)$ as the quantile treatment effect. In the example of labor market participation for instance, $Y^* = m_0(X,U)$ denotes the (latent) willingness of the individual to participate in the labor force. $m_0(\bar{x}_1, q_\tau) - m_0(\bar{x}_2, g_\tau)$ measures the structural effect of, say, a change in income from $\bar{x}_1$ to $\bar{x}_2$ on the τth quantile of the willingness-to-participate Y^*.

G by S_G and $M_S \equiv M \cap S_G$. We say that m_τ^* is identified in a set $\mathcal{M}$ if $m_\tau^* \in \mathcal{M}$ and for any $\bar{m}_\tau \in \mathcal{M}$ such that $m_\tau^* \neq \bar{m}_\tau$[20] there exists a set $\mathcal{A}$ such that for all $(x,z) \in \mathcal{A} \subset M_S$ and $G(\mathcal{A}) > 0$

$$\Pr(Y = 1 \mid x, z; m_\tau^*) \neq \Pr(Y = 1 \mid x, z; \bar{m}_\tau).$$

Note that (1.12) is equivalent to saying that

$$\Pr(Y = 1 \mid X, Z) \begin{array}{l} > \\ = 1 - \tau \\ < \end{array} \iff m_\tau^*(X, V) \begin{array}{l} > \\ = 0. \\ < \end{array}$$

We also note that m_τ^* is identified if there exists a unique function $m_\tau^* \in \mathcal{M}$ such that the relationship[21]

$$Q_\tau(Y \mid X, V) = \mathbf{1}\{m_\tau^*(X, V) > 0\}$$

holds for $(x,z) \in \mathcal{A}$ with positive probability. The first assumption is about identification of m_τ^*.

Assumption ID. *Let $\bar{m}_\tau \neq m_\tau^*$ be any other function that belongs in $\mathcal{M}$ and satisfies* (1.12)*. Then m_τ^* is observationally distinguishable from $\bar{m}_\tau$ if and only if*

$$\Pr\Big((x,z) : \{m_\tau^*(x,v) < 0 \leq \bar{m}_\tau(x,v)\} \cup \{\bar{m}_\tau(x,v) < 0 \leq m_\tau^*(x,v)\}\Big) > 0. \tag{1.13}$$

Let $\mathcal{A}_1$ be the set of (x,z) such that equation (1.13) holds. Let $(m_\tau^*, F_{U|X,Z})$ and $(\bar{m}_\tau, \bar{F}_{U|X,Z})$ denote the truth and the specified alternative, respectively. The reason for the aforementioned identification argument is analogous to Proposition 2 of Manski (1988): for each pair $(x,z) \in \mathcal{A}_1$, there exists no $\bar{F}_{U|X,Z}$ satisfying both. We will give conditions on the primitives of the models under which the identification condition (1.13) is fulfilled.

We now turn to identification of semiparametric models (1.1)-(1.4). It is well known that the index coefficients $\gamma_0' \equiv (\gamma_{10}, \beta_0')$ can only be identified up to scale. As a result, we

[20] Two functions m_τ^* and $\bar{m}_\tau$ are said to be different if they attain different values on a subset of S_G that has positive probability.

[21] An equivalent representation to this relationship is the following nonlinear conditional quantile restriction: $\mathbb{E}\big[\mathbf{1}\{Y \leq \mathbf{1}\{m_\tau^*(X,V) > 0\}\} - \tau \mid X, Z\big] = 0$.

normalize the coefficient of the endogenous regressor X_1, γ_{10}, to be $|\gamma_{10}| = 1$ and represent the linear index by $X'\beta_0 = X'\begin{pmatrix}\gamma_{10}\\ \beta_0\end{pmatrix}$ for notation convenience. Under the linear structure in which $m^*_\tau(X,V) = X'\beta_0 + \lambda_\tau(V)$, we immediately have

$$Q_\tau(Y \mid X, Z) = Q_\tau(Y \mid X, V) = \mathbf{1}\{X'\beta_0 + \lambda_\tau(V) > 0\}. \tag{1.14}$$

In words, the conditional quantile of Y is a binary outcome version of a partially linear regression function with a generated regressor V. If the latent variable Y^* underlying the observed binary data were observed, the model becomes $Q_\tau(Y^* \mid X, Z) = X'\beta_0 + \lambda_\tau(V)$. Then this model is equivalent to endogenous quantile regression models studies by Lee (2007) except that the control variable V is parametrically estimated in the first stage there. We also note that in the absence of the unknown function $\lambda_\tau(\cdot)$, model (1.14) with $\tau = .5$ is Manski (1975, 1985) exogenous binary response model with median restrictions. The following conditions are sufficient for identification of the index coefficients β_0 and the unknown function λ_τ.

Assumption SPID. *Define $\varepsilon_\tau = Y^* - X'\gamma_0 - \lambda_\tau(V)$. Assume that*

(a) The conditional density of ε_τ given X and Z is continuous and bounded away from zero uniformly in a small neighborhood of 0 and uniformly over X and Z.

(b) The first component of γ_0 satisfies $|\gamma_{10}| = 1$ and denote $\gamma_0' = (\gamma_{10}, \beta_0')$.

(c) Conditional on the control variable V, X contains at least one continuously distributed component X_c with nonzero coefficient. Denote remaining components of X by X_{-c}. For almost every $x_{-c} = (x_1, \ldots, x_{c-1}, x_{c+1}, \ldots, x_{d_x})$ and z the distribution of X_c conditional on $X_{-c} = x_{-c}$ and $Z = z$ has an everywhere positive density with respect to the Lebesgue measure.

(d) The support of the distribution of X_{-c} is not contained in any proper linear subspace of $\mathbb{R}^{d_x - 1}$.

(e) The functions $\lambda_\tau(V)$ and $F_{X_1|Z}(X_1, Z) \equiv v_0(X_1, Z)$ are continuously differentiable with continuous distributions almost everywhere.

(f) There is at least one component of Z that is not included in X, say $Z_{21} \in Z_2$ and with probability one $\partial v_0(X_1, Z)/\partial Z_{21} \neq 0$.

(g) $0 < P(Y = 1 \mid X = x, Z = z) < 1$ *for almost every* x *and* z.

Assumption SPID(a) guarantees uniqueness of the conditional τth quantile of ε_τ given X and Z; (b) is a standard scale normalization restriction in the binary response model literature; (c) requires that in addition to X_1 there is at least one exogenous regressors, say X_c, that is continuously distributed with unbounded support; (d) implies there is no exact linear relation among components of X_{-c}; (c) along with (d) ensures that for all $\bar{\beta} \neq \beta_0$ or $\bar{\lambda}_\tau(\cdot) \neq \lambda_\tau(\cdot)$ we have $R(\beta) = \int_{S_\beta} dF_{X,V} > 0$, where $S_\beta = \{x \in \mathbb{R}^{d_x}, z \in \mathbb{R}^{d_z} : \mathbf{1}\{x'\bar{\beta} + \bar{\lambda}_\tau(v) > 0\} \neq \mathbf{1}\{x'\beta_0 + \lambda_\tau(v) > 0\}\}$ and $F_{X,Z}$ is the joint cumulative distribution function of (X, Z); (f) is an exclusion restriction; (g) is standard for smoothed maximum score estimators.

Theorem 1.1. *Suppose Assumption SPID is satisfied in models* (1.1)-(1.4)*. Then the finite-dimensional parameter vector* β_0 *and the infinite-dimensional parameter* $\lambda_\tau(\cdot)$ *are identified.*

Corollary 1.1. *Suppose Assumption SPID is satisfied in models* (1.1)-(1.4)*. Then the finite-dimensional parameter vector* β_0 *and the infinite-dimensional parameter* $\lambda_\tau(\cdot)$ *are unique joint minimizers of the population objective function* $Q(\bar{\beta}, \bar{\lambda}_\tau) \equiv \mathbb{E}\big([(1-\tau) - Y]\mathbf{1}\{X'\bar{\beta} + \bar{\lambda}_\tau(V)\}\big)$*, where* $(\bar{\beta}, \bar{\lambda}_\tau)$ *is a pair of generic elements in the parameter space.*

A potential technical issue arises because the instrument Z is assumed to have compact support for the technical reason in the asymptotic analysis. This is typically required for establishing the uniform convergence of the first stage CDF estimator for $F_{X_1|Z}(X_1|Z)$. Horowitz (2009) (Corollary 4.1) gives conditions sufficient for identification of β_0 when X has bounded support in the exogenous case. It seems natural to expect to be able to relax the unbounded support condition in the endogenous case. Specifically, similar to Horowitz's (2009) argument, identification is still possible if $\text{Supp}(X'\beta_0 + \lambda_\tau(V) \mid \tilde{X}_{-c}, V)$ includes an interval containing $X'\beta_0 + \lambda_\tau(V) = 0$ for sufficiently many values of $\tilde{X}_{-c}$ and V. The set $S(\bar{\beta}, \bar{\lambda}_\tau) = \{(x, z) : -x'_{-c}\bar{\beta} - \bar{\lambda}_\tau(v) \leq x_c < -x'_{-c}\beta_0 - \lambda_\tau(v)\}$ can be equivalently rewritten as $S(\bar{\beta}, \bar{\lambda}_\tau) = \{(x, z) : -\tilde{x}'_{-c}(\bar{\beta} - \beta_0) - (\bar{\lambda}_\tau(v) - \lambda_\tau(v)) \leq x'\beta_0 + \lambda_\tau(v) < 0\}$. We impose conditions guaranteeing $\Pr[-\tilde{x}'_{-c}(\bar{\beta} - \beta_0) - (\bar{\lambda}_\tau(v) - \lambda_\tau(v)) \neq 0] > 0$ if $\bar{\beta} \neq \beta_0$ or $\bar{\lambda}_\tau \neq \lambda_\tau$. For almost

every $(\tilde{x}_{-c}, v) \in N_\delta \in \mathbb{R}^{d_x-1}$, the conditional distribution $F_{X'\beta_0+\lambda_\tau(V)|\tilde{X}_{-c}=\tilde{x}_{-c}, V=v}$ has a positive probability density everywhere on $I_\delta \equiv [-\delta, \delta]$.

Corollary 1.2 states a result for identification of β_0 when Z has bounded support.[22]?

Corollary 1.2. *Suppose Assumption SPID is satisfied except that Z has bounded support in models* (1.1)-(1.4)*. If for some $\delta > 0$, there are an interval $I_\delta = [-\delta, \delta]$ and a set $N_\delta \in \mathbb{R}^{d_x-1}$ such that (a) N_δ is not contained in any proper linear subspace of $\mathbb{R}^{d_x-1}$; (b) $P(\in N_\delta) > 0$; (c) for almost every $\tilde{x}_{-c} \in N_\delta$, the distribution of $X'\beta_0 + \lambda_\tau(V)$ conditional on $\tilde{X}_{-c} = \tilde{x}_{-c}$ has a probability density that is everywhere positive on I_δ, then the finite-dimensional parameter vector β_0 and the infinite-dimensional parameter $\lambda_\tau(\cdot)$ are identified.*

1.4 Estimation

We have presented the identification for the finite-dimensional parameter vector β_0 in a triangular binary response model under a quantile restriction. In this section we propose a series control function estimator of β_0. A straightforward extension to a nonparametric binary response model with endogeneity is discussed in Section 1.6. An interval estimator of choice probabilities by looking at a number of different quantiles is given in Appendix Section 1.9.

We propose to estimate β_0 based on equation (1.14). Recall that equation (1.14) looks like the partially linear form of the conditional quantile function of Y given X and Z. Thus it can be viewed as a generalization of quantile restrictions in the exogenous case (Kordas (2006)): $\text{Med}(Y \mid X) = \mathbf{1}\{X'\beta_0 > 0\}$ to partially linear models. Using the well-known result that for any random variable Y, $\mathbb{E}(\tau|Y-b|\mathbf{1}\{Y \geq b\} + (1-\tau)|Y-b|\mathbf{1}\{Y < b\})$ is minimized at

[22]Another interesting question is whether or not instruments are allowed to be discrete to achieve identification. As pointed out by D'Haultfoeuille and Fvrier (2012), imposing monotonicity restrictions either on the outcome or on the first stage equation can achieve point identification only when the instrument is continuous. Examples include Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007), Imbens and Newey (2009), and our model. They also show that point identification using a discrete instrument can be achieved with monotonicity on both equations.

$b = Q_\tau(Y)$, we have

$$\begin{aligned}
&\mathbb{E}\big[\rho_\tau(Y_i - \mathbf{1}\{X_i'\bar{\beta} + \bar{\lambda}_\tau(V_i) > 0\})\big] \\
&\quad \equiv \mathbb{E}\Big[\tau\Big|Y_i - \mathbf{1}\{X_i'\bar{\beta} + \bar{\lambda}_\tau(V_i) > 0\}\Big| \cdot \mathbf{1}\{Y_i - \mathbf{1}\{X_i'\bar{\beta} + \bar{\lambda}_\tau(V_i) > 0\} > 0\} \\
&\quad + (1-\tau)\Big|Y_i - \mathbf{1}\{X_i'\bar{\beta} + \bar{\lambda}_\tau(V_i) > 0\}\Big| \cdot \mathbf{1}\{Y_i - \mathbf{1}\{X_i'\bar{\beta} + \bar{\lambda}_\tau(V_i) > 0\} \le 0\}\Big]
\end{aligned}$$

is minimized at $\bar{\beta} = \beta_0$ and $\bar{\lambda}_\tau(\cdot) = \lambda_\tau(\cdot)$, where $\rho_\tau(\cdot)$ is the check function such that $\rho_\tau(u) = |u| + (2\tau - 1)u$ for $0 < \tau < 1$. Simple algebra yields

$$\mathbb{E}\big[\rho_\tau(Y_i - \mathbf{1}\{X_i'\beta + \lambda_\tau(V_i) > 0\})\big] = \mathbb{E}\big[(1 - \tau - Y)\mathbf{1}\{X'\beta + \lambda_\tau(V) > 0\} + Y\big]. \tag{1.15}$$

For an i.i.d. random sample $\{(Y_i, X_i, Z_i) : i = 1, \ldots, n\}$, an (infeasible) estimator of β_0 can be formed by minimizing the sample analogue of the population moment representation (1.15), namely $\tilde{\beta}^* = \arg\max_{\beta \in B} \tilde{Q}_n^*(\beta)$, where

$$\tilde{Q}_n^*(\beta) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - (1-\tau)]\mathbf{1}\{X_i'\beta + \lambda_\tau(V_i) > 0\}, \tag{1.16}$$

with $B \subset \mathbb{R}^{d_x - 1}$ being the compact parameter space.

If the control variable V and the function $\lambda_\tau(\cdot)$ were known, the estimator $\tilde{\beta}^*$ is a generalization of Manski's (1975, 1985) maximum score estimator to allow for endogeneity. However, the maximum score estimator has a complicated asymptotic distribution that is hard to use for inference. Characterizing the asymptotic distribution of the estimator of β_0 based on (1.16) in the presence of a nonparametric quantile function λ_τ and the generated regressor V is even much more challenging. We do not pursue this complication in the present chapter and leave this as an interesting direction for future research. One solution proposed in the literature to deal with this difficulty is to follow Horowitz's (1992) strategy of replacing the discontinuous function $\mathbf{1}\{X'\beta + \lambda_\tau(V) > 0\}$ with its smoothed version $K\left(\frac{X'\beta + \lambda_\tau(V)}{h}\right)$ where the smoothing parameter h satisfies $h = h_n \to 0$ as $n \to \infty$ and $K(\cdot)$ is chosen to be a cumulative distribution function with $\lim_{s \to -\infty} K(s) = 0$, $\lim_{s \to \infty} K(s) = 1$, and $K(t) = \int_{-\infty}^{t} K^{(1)}(s)ds$ where $K^{(1)}$ is a kernel function. Using a smoothed maximum score approach, Horowitz

(1992) improves rates of convergence under certain smoothness conditions. This smoothing strategy works in our context since as $h \to 0$,

$$K\left(\frac{X_i'\beta_0 + \lambda_\tau(V_i)}{h}\right) \to \begin{cases} 1 & \text{if } X_i'\beta_0 + \lambda_\tau(V_i) > 0 \\ 0 & \text{if } X_i'\beta_0 + \lambda_\tau(V_i) < 0 \end{cases}$$

for each $i = 1, \cdots, n$. We impose the condition $\Pr\big(X_i'\beta_0 + \lambda_\tau(V_i) = 0\big) = 0$ and therefore with probability one, $K(\cdot)$ is arbitrarily close to $\mathbf{1}\{\cdot\}$.

Following this approach, we obtain the semiparametric estimator for β_0 by maximizing the following smoothed criterion function: $\tilde{\beta} \equiv \arg\max_{\beta \in B} \quad \tilde{Q}_n(\beta)$, where

$$\tilde{Q}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\big[Y_i - (1-\tau)\big]K\left(\frac{X_i\beta + \lambda_\tau(V_i)}{h_n}\right). \tag{1.17}$$

Compared with Horowitz's (1992) smoothed maximum score estimator, there are several distinctive features in estimation based on (1.17) that are worth pointing out. The sample objective function in (1.17) contains a linear index and a nonparametric component that is a unknown conditional quantile function $\lambda_\tau(\cdot)$ with a nonparametrically generated regressor V. In econometrics, there are many estimation problems involving the partially linear structure with generated regressors. Examples are econometric models with endogeneity or sample selection, rational expectation models, and error correction models. See Li and Wooldridge (2002) for a related discussion. Specifically, previous studies involving partially linear estimation procedures with generated regressors include Ahn and Powell (1993), Chen and Khan (2003), Blundell and Powell (2007), Lee (2007), Newey (2009), and Li and Wooldridge (2002), among others. Ahn and Powell (1993) and Chen and Khan (2003) study semiparametric sample selection models with the propensity scores that are nonparametrically estimated in the first stage. Blundell and Powell (2007) analyze censored regression quantiles with nonparametric estimation of the control variable. The first three aforementioned papers use a pairwise-differencing estimation strategy with kernel weights to eliminate the nuisance nonparametric components in the partially linear structure. Following Robinson's (1988) estimation strategy, Li and Wooldridge (2002) consider semiparametric estimation of a partially linear model with (parametrically) generated regressors.

In practice, the estimator $\tilde{\beta}$ is infeasible since $\lambda_\tau(\cdot)$ and V are typically unknown. The type of estimator we propose is a natural two-step procedure, in which we approximate the infeasible objective function by substituting the unknown objects by their consistent nonparametric estimators. Specifically, the first step is to construct estimated conditional CDF as a control variable. Since the conditional CDF is a regression with $X_1 = x_1$ fixed, $F_{X_1|Z}(x_1 \mid Z = z)$ can be estimated using nonparametric regression methods such as kernel and series approaches. In this section, we are not specific about the estimation procedure for the control variable V. Instead, we simply assume that a consistent estimator $\hat{V}$ is available and satisfies a certain uniform convergence property, as discussed in Section 3.7. In the second step, we then use the nonparametrically estimated control variable $\hat{V}_i$ obtained in the first step in replace of V_i and approximate $\lambda_\tau(\cdot)$ and estimate β_0 simultaneously using the series method by maximizing the sample objective function.[23]

To describe the feasible estimator of β_0, let the number of the approximating function be κ and $\bar{P}_\kappa(v) \equiv (p_1(v), p_2(v), \ldots, p_\kappa(v))$ denote the $\kappa \times 1$ vector of the first κ approximating functions such as polynomials or splines with the property that for large κ a linear combination of $\bar{P}_\kappa(V)$ can approximate the unknown function of V well. Denote $\hat{\lambda}_\tau$ as an estimator of the true function λ_τ. The regularity conditions that κ and h must satisfy are given in Section 1.5.1. As shorthand notation, let $W = (X, V)'$, $\hat{W} = (X, \hat{V})'$, $W_i = (X_i, V_i)'$, $\hat{W}_i = (X_i, \hat{V}_i)'$, $P_{\kappa i} = P_\kappa(W_i)$, and $\hat{P}_{\kappa i} = P_\kappa(\hat{W}_i)$. Using the aforementioned notation, we define for any positive integer κ,

$$P_\kappa(w) = (x', \bar{P}_\kappa(v)) = (x', p_1(v), \ldots, p_\kappa(v)).$$

Then for $\theta_{\kappa 0} \equiv (\beta_0', \alpha_{\kappa 0}')' \in \mathbb{R}^{d_x+\kappa-1}$, $P_\kappa(w)'\theta_{\kappa 0}$ is a series approximation to $x'\beta_0 + \lambda_\tau(v)$. Also, let A denote the $((d_x - 1) \times (d_x + \kappa - 1))$ matrix such that $A = (I_{(d_x-1)}, \mathbf{0}_{(d_x-1)\times\kappa})$, where I_{d_x-1} is the $((d_x - 1) \times (d_x - 1))$ identity matrix and $\mathbf{0}_{(d_x-1)\times\kappa}$ is the $((d_x - 1) \times \kappa)$ matrix of zeros. The feasible series estimator $\hat{\beta}$ of β_0 is given by

$$\hat{\beta} = A\hat{\theta}_{n\kappa},$$

[23] Following Chen (2007), in the absence of generated regressors, this approach belongs to a type of sieve simultaneous M-estimation.

where $\hat{\theta}_{n\kappa} \equiv (\hat{\beta}', \hat{\alpha}_{n\kappa}')'$ is the series estimator of $\theta_{\kappa 0} \equiv (\beta_0', \alpha_{\kappa 0}')'$ obtained by maximizing a sample objective function $\hat{Q}_{n\kappa}(\theta)$, as defined below, with a plug-in preliminary estimate of V. Specifically, $\hat{\theta}_{n\kappa} \equiv \arg \max_{\theta \equiv (\beta', \alpha')' \in \Theta_\kappa} \hat{Q}_{n\kappa}(\theta)$, where

$$\hat{Q}_{n\kappa}(\theta) = \frac{1}{n} \sum_{i=1}^{n} t(\hat{W}_i)[Y_i - (1-\tau)] K\left(\frac{X_i'\beta + \bar{P}_\kappa(\hat{V}_i)'\alpha}{h} \right), \tag{1.18}$$

$\Theta_\kappa = B \times \Lambda_\kappa \subset \mathbb{R}^{d_x + \kappa - 1}$ is a compact parameter space, and $t(w) = \mathbf{1}\{(x_1, z) \in \mathcal{C}\}$ is a trimming function with a compact set $\mathcal{C}$ that is appropriately chosen. Note that the estimator of $\lambda_\tau(v) \equiv Q_\tau(U \mid V = v)$ is given simultaneously by $\bar{P}_\kappa(v)'\hat{\alpha}_{n\kappa}$.[24]

For the purpose of comparison in Section 1.5.2, we also define an estimator $\bar{\beta}^\star \equiv A\bar{\theta}^*_{n\kappa}$ of β_0 in a partially linear binary regression quantiles: $Q_\tau(Y \mid X, V) = \mathbf{1}\{X'\beta_0 + \lambda_\tau(V) \geq 0\}$, where the regressor V is assumed to be observed and the series estimator $\bar{\theta}^*_{n\kappa}$ is obtained by maximizing the sample objective function $\bar{\theta}^*_{n\kappa} \equiv \arg \max_{\theta \equiv (\beta', \alpha')' \in \Theta_\kappa} \bar{Q}^*_{n\kappa}(\theta)$, where

$$\bar{Q}^*_{n\kappa}(\theta) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - (1-\tau)] K\left(\frac{X_i'\beta + \bar{P}_\kappa(V_i)'\alpha}{h} \right). \tag{1.19}$$

We note that the advantage of using series estimation in our context is the relative ease of implementation: when the series method is implemented, the estimation problem reduces to a parametric one and, moreover, we get estimates of parametric component β_0 and nonparametric component λ_τ simultaneously. The series method is also convenient for imposing additive separability restrictions by simply excluding interaction terms.

1.5 Asymptotic Theory

This section presents the main results of the chapter. We give asymptotic theory for the estimator $\hat{\beta}$ of β_0 in models (1.1)-(1.4), as described in Section 1.4. To best convey the issues surrounding estimation, we start with deriving an asymptotic expansion of the series

[24] Although in our setting the parameter of primary interest is β_0, it should be noted that the estimate of the control function $\lambda_\tau(v)$ could be of potential interest. For example, since the control function $\lambda_\tau(v)$ is constant if all components of X are exogenous, the estimate $\hat{\lambda}_\tau$ may be further used for a specification test for endogeneity by testing whether or not $\lambda_\tau(v)$ is a constant with probability one. We leave this topic for future research.

estimator $\hat{\theta}_{n\kappa}$ in the presence of the nonparametrically generated regressor, which is essential to characterize the contributions of the first-stage estimation error of the generated regressor to the asymptotic behavior of the final estimator, including its rates of convergence and the asymptotic distribution. This asymptotic representation result, along with consistency and convergence rates of the estimator, are presented in Section 1.5.1. Section 1.5.2 states asymptotic normality results for the semiparametric model, with a focus on characterizing the influence of the first-stage estimation error on the first-order asymptotic distribution of $\hat{\beta}$.

Notation Define $m_\tau \equiv X'\beta_0 + \lambda_\tau(V)$, $m_{\tau i} \equiv X_i'\beta_0 + \lambda_\tau(V_i)$, $\varepsilon_\tau \equiv Y^* - m_\tau$, $\tilde{X} \equiv (X_2, \ldots, X_{d_x})'$, $\tilde{W} = (\tilde{X}', V)'$, $\bar{W} = (\tilde{X}', \hat{V})'$, and $\tilde{w} = (\tilde{x}', v)'$. Let $F_{\varepsilon_\tau}(\cdot \mid X'\beta_0, \tilde{W})$ and $F^{(1)}_{\varepsilon_\tau}(\cdot \mid X'\beta_0, \tilde{W})$, respectively, denote the cumulative distribution function and the probability density function of ε_τ conditional on $X'\beta_0$, $\tilde{W}$,[25] and $f_{m_\tau}(\cdot \mid \tilde{W})$ the probability density function of m_τ conditional on $\tilde{X}$ and V. Also note that $F_{\varepsilon_\tau}(0 \mid X'\beta_0, \tilde{W}) = \tau$ by the QER in Assumption SPID (a). In what follows, for the sake of simplifying notation, denote $P_\kappa(\tilde{W}_i) = (\tilde{X}_i', \bar{P}_\kappa(V_i))$, $P_\kappa(\bar{W}_i) = (\tilde{X}_i', \bar{P}_\kappa(\hat{V}_i))$, $F^{(1)}_{\varepsilon_{\tau i}} = F^{(1)}_{\varepsilon_\tau}(0 \mid X_i'\beta_0, \tilde{X}_i, V_i)$, $f_{m_{\tau i}} = f_{m_\tau}(0 \mid \tilde{X}_i, V_i)$. Also denote $K_h(\cdot) \equiv K(\cdot/h)$, $K_h^{(1)}(\cdot) \equiv K^{(1)}(\cdot/h)/h$, and $K_h^{(2)}(\cdot) \equiv K^{(2)}(\cdot/h)/h^2$.

1.5.1 Consistency, Stochastic Expansions, and Convergence Rates

Under Assumption 1.3(b) below, $\hat{Q}_{n\kappa}(\theta)$ from 1.18 is twice differentiable with respect to θ. Differentiate $\hat{Q}_{n\kappa}(\theta)$ with respect to θ and under Assumption 1.2 discussed below, with probability tending to 1, $\hat{\theta}_{n\kappa}$ satisfies the following first order condition

$$0 = \frac{\partial \hat{Q}_{n\kappa}(\hat{\theta}_{n\kappa})}{\partial \theta} = n^{-1} \sum_{i=1}^{n} \hat{t}_i [Y_i - (1-\tau)] K_h^{(1)}(P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}) P_\kappa(\bar{W}_i). \tag{1.20}$$

[25] Note that $(X'\beta_0, \tilde{X}')$ and X have one-to-one relation for fixed values of β since $|\beta_1| = 1$ by normalization, we therefore write

$$\mathbb{E}(Y \mid X, V) = \Pr(\varepsilon_\tau > -X'\beta_0 - \lambda_\tau(V) \mid X, V) = 1 - F_{\varepsilon_\tau}(-m_\tau \mid X'\beta_0, \tilde{W}).$$

Define

$$\hat{G}_{n\kappa}(\theta) = \hat{H}_{n\kappa}^{-1} n^{-1} \sum_{i=1}^{n} \hat{t}_i [Y_i - (1-\tau)] K_h^{(1)}\big(P_\kappa(\hat{W}_i)'\theta\big) P_\kappa(\bar{W}_i) \text{ and}$$
$$\tilde{G}_{n\kappa}(\theta) = \tilde{H}_{n\kappa}^{-1} n^{-1} \sum_{i=1}^{n} t_i [Y_i - (1-\tau)] K_h^{(1)}\big(P_\kappa(W_i)'\theta\big) P_\kappa(\tilde{W}_i)$$

with $\hat{H}_{n\kappa} = n^{-1} \sum_{i=1}^{n} \hat{t}_i [Y_i - (1-\tau)] K_h^{(2)}(P_\kappa(\hat{W}_i)'\theta_{\kappa 0}) P_\kappa(\bar{W}_i) P_\kappa(\bar{W}_i)'$ and $\tilde{H}_{n\kappa} = n^{-1} \sum_{i=1}^{n} t_i [Y_i - (1-\tau)] K_h^{(2)}(P_\kappa(W_i)'\theta_{\kappa 0}) P_\kappa(\tilde{W}_i) P_\kappa(\tilde{W}_i)'$.

It is useful to rewrite the term $P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}$ in (1.20) in a form involving the true values of parameters $\theta_{\kappa 0}$ and the estimation errors $\hat{\theta}_{n\kappa} - \theta_{\kappa 0}$ and $\hat{V}_i - V_i$. That is, we decompose $P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}$ into the following three terms

$$\begin{aligned} P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa} &= P_\kappa(W_i)'\theta_{\kappa 0} + P_\kappa(\hat{W}_i)' \underbrace{(\hat{\theta}_{n\kappa} - \theta_{\kappa 0})}_{\text{Channel 1}} + \underbrace{\big(P_\kappa(\hat{W}_i) - P_\kappa(W_i)\big)'}_{\text{Channel 2}} \theta_{\kappa 0} \\ &\equiv P_\kappa(W_i)'\theta_{\kappa 0} + \Delta_{\kappa i}(\hat{\theta}_{n\kappa}). \end{aligned} \tag{1.21}$$

Expression (1.21) clearly implies that there are two channels through which the first-stage estimation error $(\hat{V}_i - V_i)$ of the generated regressor affects the asymptotic distribution of $\hat{\beta}$. The first contribution is derived from indirect effects through *Channel 1* accounting for the estimation error of the conditional quantile function $Q_\tau(U \mid V = v) \equiv \lambda_\tau(v)$ by using $\hat{V}_i$ instead of V_i in forming $\hat{\lambda}_\tau(\cdot)$. In other words, using the estimated control variable $\{\hat{V}_i\}_{i=1}^n$ changes the functional form of the quantile type of nonparametric projection of U onto V, i.e., $Q_\tau(U \mid V)$. In our setting, the functional form of $Q_\tau(U \mid V = v)$ is determined by the series coefficients $\theta_{n\kappa}(v)$.[26] In other words, the first role of the generated regressor V through Channel 1 changes the shape of the function $Q_\tau(U \mid V)$. On the other hand, the second contribution comes from the direct effect through *Channel 2* reflecting the additional uncertainty due to estimating β_0 via $Q_\tau(U \mid V = \hat{V}_i) \equiv \lambda_\tau(\hat{V}_i)$, i.e, the first-stage estimate in $\hat{V}_i$ enters the argument at which the conditional quantile is evaluated.[27] Note that the first

[26] We express the series coefficient vector $\hat{\theta}_{n\kappa}(v)$ as a function of v to make explicit the first role that v plays.

[27] See Hahn and Ridder (2010) and Mammen, Rothe, and Schienle (2011) for related discussion on the dual roles played by the first-stage estimate of generated regressors.

role of the estimation error of $\{V_i\}_{i=1}^n$ impacts the final estimator through the second-order term in the expansion, as will be clear later.

We need the following conditions to analyze the asymptotic properties of the series estimator involving a nonparametric quantile regression function with the nonparametrically generated regressor.

Assumption 1.1. *The data* $\{(Y_i, X_i, Z_i) : i = 1, \ldots, n\}$ *is a random sample.*

Assumption 1.2. *The true value of the parameter* $\beta_0 = (\beta_{20}, \ldots, \beta_{d_x 0})'$ *is an element of the interior of the compact parameter space* $B \subset \mathbb{R}^{d_x - 1}$.

Assumption 1.3 (Smoothing Functions)**.**

(a) Let the smoothing function $K(\cdot)$ *be a continuous function of the real line such that* $|K(u)| < M$ *for some finite* M *and all* u *in* $(-\infty, \infty)$ *and* $\lim_{u \to -\infty} K(u) = 0$ *and* $\lim_{u \to \infty} K(u) = 1$.

(b) The function $K(\cdot) : \mathbb{R} \to \mathbb{R}$ *is twice continuously differentiable everywhere with uniformly bounded first and second derivatives denoted by* $K^{(1)}(\cdot)$ *and* $K^{(2)}(\cdot)$ *and satisfies and following conditions:* $\int_{-\infty}^{\infty} K^{(1)}(u)du < \infty$, $\int_{-\infty}^{\infty} K^{(2)}(u)du < \infty$, *and* $\int_{-\infty}^{\infty} |u^2 K^{(2)}(u)|du < \infty$.

(c) Let $K^{(1)}(\cdot)$ *be* ν*-times continuously differentiable and satisfies the following conditions:* $\int u^j K^{(1)}(u)du = 0$ *for each integer* $1 \leq j < \nu$ *and* $\int |u^\nu K^{(1)}(u)|du < \infty$.

Assumption 1.1 describes the data. Assumption 1.2, along with Assumption SPID(b), contains normalization restrictions that are standard assumptions in the semiparametric estimation literature on binary response models and models with index restrictions. Assumption 1.3 places regularity conditions on the smoothing function $K(\cdot)$ for smoothed maximum score estimators. It implies that $K^{(1)}(\cdot)$ is a νth order kernel function.

Assumption 1.4 (Series Approximation)**.**

(a) The support of V *is* $\mathcal{V} = [0, 1]$ *on which* V *has an absolutely continuous probability density function that is bounded above by a positive constant, bounded away from zero, and is twice continuously differentiable.*

(b) The conditional quantile regression function $\lambda_\tau(\cdot)$ is r-times continuously differentiable on $\mathcal{V}$ for some $r \geq 2$.

Assumption 1.4 collects regularity conditions that are commonly imposed in the series approximation literature. Part (a) imposes smoothness restrictions on the conditional density function, implying that for all κ the smallest eigenvalue of H_κ is bounded away from zero and the largest eigenvalue of H_κ is bounded, and that there exists a sequence of constants $\zeta_0(\kappa)$ that satisfy $\sup_{v\in\mathcal{V}}\|\bar{P}_\kappa(v)\| \leq \zeta_0(\kappa)$ and $\kappa = \kappa_n$ such that $\zeta_0(\kappa)^2\kappa/n \to 0$ as $n \to \infty$, where $\zeta_0(\kappa)$ will increase with κ and therefore the sample size. Namely, part (a) imposes a uniform bound on the magnitude of the series term in the support of V. As shown in Newey (1997), $\zeta_0(\kappa) = O(\kappa)$ for polynomials and $\zeta_0(\kappa) = O(\kappa^{1/2})$ for splines. In addition, differentiability of the density of V is used to ensure that the bias of the series estimator converges to zero sufficiently rapidly. The smoothness condition imposed in part (b) is required to control the bias of the series estimator and depends on the type of the approximating function. It follows from part (b) and Lorentz (1966) that for both power series and splines, there exists a unique series representation $P_\kappa(W)'\theta_{\kappa 0}$, where $\theta_{\kappa 0} \equiv (\beta_0, \alpha_{\kappa 0}')' \in \mathbb{R}^{d_x+\kappa-1}$ such that the uniform approximation error to the function $X'\gamma + \lambda_\tau(\cdot)$ and its first derivative shrink at the rates $\sup_w |x'\beta_0 + \lambda_\tau(v) - P_\kappa(w)'\theta_{\kappa 0}| = O(\kappa^{-r})$ and $\sup_v \left| \frac{\partial \lambda_\tau(v)}{\partial v} - \sum_{k=1}^{\kappa} \left(\frac{\partial p_k(v)}{\partial v} \right) \alpha_{\kappa 0}^k \right| = O(\kappa^{-r+1})$, respectively as $\kappa \to \infty$.

Assumption 1.5 (Smoothing Parameters)**.** *The smoothing parameters $\kappa = \kappa_n$ and $h = h_n$ satisfy that*

(a) $\kappa = C_\kappa n^{\rho_\kappa}$ for some constant C_κ satisfying $0 < C_\kappa < \infty$ and some ρ_κ satisfying $\frac{\nu+1}{(2\nu+1)r} < \rho_\kappa < \frac{\nu}{2(2\nu+1)}$ for power series and $\frac{\nu+1}{(2\nu+1)r} < \rho_\kappa < \frac{2\nu}{3(2\nu+1)}$ for splines.
(b) The bandwidth satisfies $h = h_n \to 0$ and $\frac{\log n}{nh_n^4} \to 0$ as $n \to 0$.

Assumption 1.5(a) states the rates at which $\kappa = \kappa_n \to \infty$ and $h = h_n \to 0$ as $n \to \infty$. The required rate for κ guarantees that the asymptotic bias and variance of the series estimation of the conditional quantile function $\lambda_\tau(\cdot)$ are sufficiently small to achieve an $(nh)^{-1/2}$ rate of convergence of the final estimator of β_0. To be precise, the left inequality for ρ_κ is used to

make the estimator have no asymptotic bias introduced by the series approximation, whereas the right inequality is required to make the remainder terms of the stochastic expansion derived below negligible, as will be clear later. The necessary smoothness conditions implied by this assumption is that $r \geq 4$ and $r \geq 3$ for power series and splines, respectively. Analogous to kernel density estimation, the bandwidth h is assumed to shrink with the sample size n at a certain rate.

Assumption 1.6 (First-stage Estimation Accuracy)**.** *The support of Z is a compact set $\mathcal{Z} \equiv \bigotimes_{i=1}^{d_z}[\underline{z}_i, \bar{z}_i]$. The rate of convergence of the estimator $\hat{V} = \hat{F}_{X_1|Z}(X_1 \mid Z)$ is assumed to be*

$$\sup_{x_1 \in \mathbb{R}} \sup_{z \in \mathcal{Z}} \left| \hat{F}_{X_1|Z}(x_1 \mid z) - F_{X_1|Z}(x_1 \mid z) \right| = O_p\left(\Delta_v\right),$$

satisfying $\Delta_v = o_p\left(n^{1/4}h^{5/4}\right)$.

Assumption 1.6 is a high-level assumption about the accuracy of the first-stage estimator $\hat{V}$. Throughout this section, we do not treat the specific form of estimation of $V = F_{X_1|Z}(X_1 \mid Z)$, instead the asymptotics are derived for any consistent estimate $\hat{V}_i$ of V_i satisfying a given uniform rate of convergence. The required rate on Δ_v reflects the nonparametric rate of convergence of $\hat{\beta}$ and the appearance of the bandwidth h in the denominator in the argument of the smoothing function K. This assumption is an analogous requirement to the well-known necessary condition: the difference between $\hat{V}$ and V vanishes at a rate satisfying $o_p(n^{-1/4})$ that arises frequently in standard semiparametric estimation with $\sqrt{n}$-consistency. This assumption would be a necessary condition to truncate the expansion that will be clear later. We note that several uniform convergence results are available for the standard nonparametric regression estimation. See, for example, Masry (1996) for Nadaraya-Watson, local linear, and local polynomial estimators and Newey (1997) and de Jong (2002) for series estimators. However, to derive uniform convergence results for conditional CDF $F_{X_1|Z}(x_1 \mid Z = z)$ estimation, one needs to deal with a potential difficulty that arises due to the fact that $F_{X_1|Z}(x_1 \mid Z = z)$ is the step function rather than continuous in x_1. It is known that estimation of the conditional CDF can converge uniformly over the conditioning variables Z.

Furthermore, the uniform convergence result jointly over X_1 and Z of the conditional CDF estimator using kernel methods has been established by Rothe (2010). As far as we know, joint uniformity over (x_1, z) for estimation of $F_{X_1|Z}(x_1 \mid Z = z) = \mathbb{E}(\mathbf{1}\{X_1 \leq x_1 \mid Z = z\})$ using series methods is not yet available. This is a topic for future research.

Assumption 1.7. *Assume that*

(a) The first absolute moments $\mathbb{E}|P_\kappa(\tilde{W})|$, $\mathbb{E}|P_\kappa(\tilde{W})P_\kappa(\tilde{W})'|$, *and* $\mathbb{E}|P_\kappa(\tilde{W})P_\kappa(\tilde{W})'P_\kappa(\tilde{W})P_\kappa(\tilde{W})'|$ *are finite.*

(b) For every integer k, $0 < k < \nu$, *any* $\epsilon > 0$, *and any sequence* $h \longrightarrow 0$, $\lim_{n\to\infty} h^{k-\nu} \int_{|hu|>\epsilon} |u^k K^{(1)}(u)| du = 0$ *and* $\lim_{n\to\infty} h^{-1} \int_{|hu|>\epsilon} |K^{(2)}(u)| du = 0$.

(c) For each integer k, $0 < k < \nu$, *all* m_τ *in a neighborhood of 0, almost every* $\tilde{w}$, *and some finite constant* M, $f^{(k)}_{m_\tau}(m_\tau \mid \tilde{W})$ *exists and is a continuous function of* m_τ *satisfying* $|f^{(k)}_{m_\tau}(m_\tau \mid \tilde{W})| < M$.

(d) For each integer k, $0 < k \leq \nu$, *all* m_τ *in a neighborhood of 0, almost every* $\tilde{w}$, *and some finite constant* M, $F^{(k)}_{\varepsilon_\tau}(-m_\tau \mid m_\tau, \tilde{W})$ *exists and is a continuous function of* m_τ *satisfying* $|F^{(k)}_{m_\tau}(-m_\tau \mid m_\tau, \tilde{W})| < M$.

Assumptions 1.7 along with Assumption 1.3 collect standard regularity conditions for smoothed maximum score estimators,[28] as in Horowitz (1992) and Kordas (2006). Assumption 1.7 ensures that B, H, and Σ exist and certain sequences of integrals converge, as shown in the Appendix.

The following theorem presents the results of consistency, the convergence rate, and the stochastic expansion for the series estimator $\hat{\theta}_{n\kappa}$, accounting for the presence of the nonparametrically generated regressor. This expansion result is an important ingredient for accurately describing the asymptotic distribution of the final estimator of the finite-dimensional parameter vector β_0, as will be shown in Section 1.5.2.

Theorem 1.2. *Suppose Assumptions SPID, 1.1-1.7 below hold. Then as* $n \to \infty$,

(a) $\lim_{n\to\infty} \|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\| = 0$.

[28] These assumptions are analogous to those made in kernel density estimation.

(b) $\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\| = O_p\Big(\big(\frac{\kappa}{nh}\big)^{1/2} + \frac{\kappa^{-r}}{h^{1/2}} + \big(\frac{\kappa}{nh^3}\big)^{1/2}\Delta_v\Big)$.

(c) The series estimator $\hat{\theta}_{n\kappa}$ *has the asymptotic expansion*

$$\begin{aligned}\hat{\theta}_{n\kappa} - \theta_{\kappa 0} = {} & H_\kappa^{-1} n^{-1} \sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_h^{(1)}(m_{\tau i})P_\kappa(\tilde{W}_i) \\ & + H_\kappa^{-1} n^{-1} \sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_h^{(2)}(m_{\tau i})\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)P_\kappa(\tilde{W}_i) \\ & + H_\kappa^{-1} n^{-1} \sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_h^{(2)}(m_{\tau i})b_{\kappa 0}(V_i)P_\kappa(\tilde{W}_i) + \hat{R}_n, \qquad (1.22)\end{aligned}$$

where the remainder term $\hat{R}_n$ *satisfies*

$$\|\hat{R}_n\| = O_p\big(\zeta_0(\kappa)\kappa/n + \kappa^{-r} + \zeta_0(\kappa)\|\theta - \theta_{\kappa 0}\|^2 + \zeta_0(\kappa)\kappa^{-2r} + \zeta_0(\kappa)\Delta_v^2\big) + o_p\big((nh)^{-1/2}\big).$$

Remark 1.1. Based on Theorem 1.2(c), it is clear that the asymptotic characterization of $\hat{\theta}_{n\kappa}$ depends on three terms. The first term on the right hand side of equation (1.22) is the main one that reflects the uncertainty if we knew λ_τ and V. By adapting Horowitz's (1992) argument to the partially linear structure, this term can be normally handled by a suitable central limit theorem and law of large numbers. The second term corresponds to the first-stage estimation error. To characterize the influence of the first-step estimation error of $\{\hat{V}_i\}_{i=1}^n$ on the asymptotic normality of the final estimator, we will further investigate the second term on the right hand side of equation (1.22) in Section 1.5.2. Also note that there is the other term that accounts for the sampling variation in the series estimate $\hat{\lambda}_\tau$ which is smaller than $(nh)^{-1/2}$ by Assumption 1.5(a) and therefore is asymptotically negligible.

Remark 1.2. The message from Theorem 1.2(a) is that the uniform convergence rate of the series estimator $\hat{\theta}_{n\kappa}$ is the sum of three terms associated with the standard deviation, bias, and the first-stage convergence rate, respectively. If κ is chosen faster than $n^{1/(2r+1)}$ (i.e., undersmoothing) and V did not have to be estimated in the first stage, then the uniform convergence rate reduces to $O_p\Big(\big(\frac{\kappa}{nh}\big)^{1/2}\Big)$.

The following corollary states the rate of convergence of the proposed estimator $\hat{\beta}$.

Corollary 1.3. *Under Assumptions SPID and 1.1-1.6, as* $n \to \infty$

$$\|\hat{\beta} - \beta_0\| = O_p\big((nh)^{-1/2} + (nh^3)^{-1/2}\Delta_v\big).$$

The convergence rate of $\hat{\beta}$ is the sum of two terms, depending on the bandwidth h and the first-stage convergence rate Δ_v. In particular, if h is chosen proportional to $n^{-1/(2\nu+1)}$ and Δ_v is assumed to satisfy $\Delta_v \propto n^{-\rho_v}$, then the conclusion of Corollary 1.3 is

$$\|\hat{\beta} - \beta_0\| = O_p\Big(\max\Big\{n^{-\nu/(2\nu+1)},\, n^{-\rho_v+(1-\nu)/(2\nu+1)}\Big\}\Big).$$

Analogous to kernel density estimation, the rate of the bandwidth h that minimizes mean square error if we knew λ_τ and V is $n^{-1/(2\nu+1)}$. It then follows that the optimal rate of the second stage estimator $\hat{\beta}$ will be attained when either $\{V_i\}_{i=1}^n$ did not have to be estimated from the data or the following condition holds

$$\rho_v > \frac{1}{2\nu+1},$$

implying that ρ_v depends on the order of the kernel function $K^{(1)}$. We will return to this point on the required rate of convergence of the first-stage estimator later.

1.5.2 Asymptotic Normality

This subsection presents the result of the distribution theory for the semiparametric estimator $\hat{\beta}$. We now outline the heuristics leading to the asymptotic normality result.

Note that $\hat{\beta} - \beta_0 = A(\hat{\theta}_{n\kappa} - \theta_{\kappa 0})$. Based on the expansion given in Theorem 1.2(c), the series estimator $\hat{\beta}$ has the following asymptotic expansion

$$\begin{aligned}\hat{\beta} - \beta_0 &= AH_\kappa^{-1} n^{-1} \sum_{i=1}^n t_i[Y_i - (1-\tau)]K_h^{(1)}\big(m_{\tau i}\big)P_\kappa(\tilde{W}_i) \\ &\quad + AH_\kappa^{-1} n^{-1} \sum_{i=1}^n t_i[Y_i - (1-\tau)]K_h^{(2)}\big(m_{\tau i}\big)\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)P_\kappa(\tilde{W}_i) + o_p\big((nh)^{-1/2}\big) \\ &\equiv AH_\kappa^{-1}T_{1n\kappa} + AH_\kappa^{-1}T_{2n\kappa} + o_p\big((nh)^{-1/2}\big). \qquad (1.23)\end{aligned}$$

In order to obtain the asymptotic variance of $\hat{\beta}$, it is essential to characterize the asymptotic behavior of the stochastic components on the right-hand side of (1.23): the first term $T_{1n\kappa}$ is the one that reflects the uncertainty induced by replacing the expectation with a sample average if the nonparametric components λ_τ and $\{V_i\}_{i=1}^n$ were known and did not

have to be estimated from the data; The second term $T_{2n\kappa}$ captures the sampling variation in the estimate of the generated regressor V. It is the key to characterizing the influence of the estimation uncertainty of $\hat{V}$ on the final estimator $\hat{\beta}$.

Let $q(\tilde{w})$ denote the residual from $t(w)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{w})f_{m_\tau}(0 \mid \tilde{w})$-weighted mean square projection of $\tilde{X}$ on the function $\lambda_\tau(V)$, i.e.,

$$q(\tilde{w}) = \tilde{x} - \frac{\mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})f_{m_\tau}(0 \mid \tilde{W})\tilde{X} \mid V = v]}{\mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})f_{m_\tau}(0 \mid \tilde{W}) \mid V = v]}.$$

Define

$$\varphi(\tilde{w}) = \left(\mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})f_{m_\tau}(0 \mid \tilde{W})q(\tilde{W})q(\tilde{W})']\right)^{-1} q(\tilde{w}).$$

It can be shown that β_0 and A can be represented as

$$\beta_0 = \mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})f_{m_\tau}(0 \mid \tilde{W})\varphi(\tilde{W})m_\tau(W)] \text{ and}$$
$$A = \mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})f_{m_\tau}(0 \mid \tilde{W})\varphi(\tilde{W})P_\kappa(\tilde{W})'].$$

To characterize the influence of the first-stage estimator on $\hat{\beta}$, we take a further step by specifying the bias and variance of the first-stage estimator $\{\hat{V}_i\}_{i=1}^n$. The asymptotic distribution will be derived for any consistent estimator $\{\hat{V}_i\}_{i=1}^n$ with given bias and variance. As we choose the kernel estimator for the generated regressor V_i in simulation experiments and empirical applications, we follow Sperlich (2009) by giving the bias and variance of $\hat{V}_i$ in terms of a bandwidth g, satisfying $g \to 0$ as $n \to \infty$. The following assumption states the bias and variance of the first-stage estimator $\{\hat{V}\}_{i=1}^n$.

Assumption 1.8 (Bias and Variance of the First-stage Estimator)**.**

(a) The function $F_{X_1|Z}(x_1, z)$ is r-times differentiable with respect to z, and the derivatives are uniformly continuous and bounded.

(b) For the first-stage estimator $\hat{V}$ using the smoothing parameter g that satisfies $g \to 0$ as $n \to \infty$, the bias, denoted by b_v, is assumed to be of order $O(g^r)$, whereas the variance $\sigma^2_{\zeta_v}$ is of order $O\left((ng^{d_z})^{-1}\right)$, both uniformly over the two arguments x_1 and z.

(c) Both $b_v(\cdot)$ and $\sigma_{\zeta_v}(\cdot)$ are Lipschitz continuous and the covariance $\mathbb{E}\left[\zeta_i\sigma_{\zeta_v}(V_i)\zeta_j\sigma_{\zeta_v}(V_j)\right] = O(1/n)$ uniformly for $i \neq j = 1, \ldots, n$.

(d) The conditional quantile function $Q_\tau(U|V) \equiv \lambda_\tau(V)$ is assumed to be strictly monotonic.

The order of magnitude of the bias and variance specified in Assumption 1.8 (b) are directly based on Rothe (2010), who shows that for the Nadaraya-Watson-type estimator $\hat{F}_{X_1|Z}(x_1, z)$, the uniform rate of convergence can be given by

$$\sup_{x_1 \in \mathbb{R}} \sup_{z \in \mathcal{Z}} |\hat{F}_{X_1|Z}(x_1, z) - F_{X_1|Z}(x_1, z)| = O_p\left(g^r + \left(\frac{\log n}{n g^{d_z}}\right)^{1/2}\right).$$

Furthermore, given Assumption 1.8, it will be useful to rewrite the estimation error of the first-stage estimator $\{\hat{V}_i\}_{i=1}^n$ as

$$\hat{V}_i - V_i = b_v(V_i) + \zeta_i \sigma_{\zeta_i}(V_i), \quad i = 1, \ldots, n, \tag{1.24}$$

where ζ_v satisfies $\mathbb{E}(\zeta_v \mid V = v) = 0$ and $\mathrm{Var}(\zeta_v \mid V = v) = 1$.[29]

To formally state the asymptotic results for the estimator, we need the following additional assumptions to prove asymptotic normality.

Assumption 1.9. *As function of v, $\mathbb{E}[t(W) F_{\varepsilon_\tau}^{(1)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}(0 \mid \tilde{W}) \tilde{X} \mid v]$ and $\mathbb{E}[t(W) F_{\varepsilon_\tau}^{(1)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}(0 \mid \tilde{W}) \mid v]$ are continuously differentiable.*

Assumption 1.9 is similar to Assumption 3.9 in Lee (2007), implying that for both power series and splines, there exists a sequence of $((d_x - 1) \times \kappa)$ matrices Ψ_κ such that $\mathbb{E}[t(W) F_{\varepsilon_\tau}^{(1)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}(0 \mid \tilde{W}) \|\varphi(\tilde{W}) - \Psi_\kappa P_\kappa(\tilde{W})\|^2] \to 0$ as $\kappa \to \infty$.

Assumption 1.10. *The matrix $H = \mathbb{E}[t(W) q(\tilde{W}) q(\tilde{W})' F_{\varepsilon_\tau}^{(1)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}(0 \mid \tilde{W})]$ is negative definite.*

Assumption 1.10 guarantees the nonsigularity of the asymptotic covariance matrix.

The following theorem gives the main results concerning the asymptotic bias, asymptotic variance, and asymptotic normality of the proposed two-step estimator of the finite-dimensional parameter vector β_0.

[29] As pointed out by Sperlich (2009), the expression (1.24) just assumes that the estimator $\hat{V}$ has the additive bias and stochastic error rather than assuming an additive error of the original model that generates V. This additivity property is satisfied for almost all nonparametric and semiparametric estimators.

Theorem 1.3. *Suppose Assumptions SPID, 1.1-1.10 hold.*

(a) Let $\alpha_\nu(L) = \int_{-\infty}^{\infty} u^\nu L(u)du$. *Then the bias,* $\mathbb{E}(\hat{\beta} - \beta_0)$, *can be expressed as* $h^\nu B_1 + h^{\nu-2}B_2 + o(h^\nu + h^{\nu-2}g^r)$ *where the two leading terms* B_1 *and* B_2 *are given by*

$$B_1 = -\alpha_\nu(K^{(1)}) \times \sum_{j=1}^{\nu} \frac{1}{j!(\nu-j)!}\mathbb{E}\big[F_{\varepsilon_\tau}^{(j)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}^{(\nu-j)}(0 \mid \tilde{W})\varphi(\tilde{W})\big],$$

$$B_2 = -\alpha_{\nu-1}(K^{(2)}) \times \sum_{j=1}^{\nu-1} \frac{1}{j!(\nu-j-1)!}\mathbb{E}\big[t(W)F_{\varepsilon_\tau}^{(j)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}^{(\nu-j-1)}(0 \mid \tilde{W})\frac{d\lambda_\tau(V)}{dv} b_v\big(\lambda_\tau^{-1}(-X'\beta_0)\big)\varphi(\tilde{W})\big].$$

(b) Let $R(L) = \int_{-\infty}^{\infty} L(u)^2 du$. *Then the variance of* $\hat{\beta}$, $Var(\hat{\beta})$, *is* $\frac{\Omega_1}{nh} + \frac{\Omega_2}{nh^3} + o\big(\frac{1}{nh} + \frac{1}{n^2h^3g^{d_z}}\big)$, *where*

$$\Omega_1 = R(K^{(1)})\tau(1-\tau)\mathbb{E}[t(W)\varphi(\tilde{W})\varphi(\tilde{W})' f_{m_\tau}(0 \mid \tilde{W})] \text{ and}$$

$$\Omega_2 = R(K^{(2)})\mathbb{E}\left[t(W)F_{\varepsilon_\tau}^{(1)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}(0 \mid \tilde{W})\frac{d\lambda_\tau(V)}{dv}\varphi(\tilde{W})\right]\mathbb{E}\big[\sigma_\zeta^2\big(\lambda_\tau^{-1}(-X'\beta_0)\big)\big] \times \mathbb{E}\left[t(W)F_{\varepsilon_\tau}^{(1)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}(0 \mid \tilde{W})\frac{d\lambda_\tau(V)}{dv}\varphi(\tilde{W})'\right].$$

(c) Let $\delta = \min\{nh, n^2h^3g^{d_z}\}$. *Then the asymptotic distribution of* $\hat{\beta}$ *is*

$$\Omega^{-1/2}\left[\sqrt{\delta}(\hat{\beta} - \beta_0) - B\right] \xrightarrow{d} N\big(0, I\big)$$

with $B = \sqrt{\delta}\big[h^\nu B_1 + h^{\nu-2}B_2\big]$ *and* $\Omega = \frac{\Omega_1}{nh} + \frac{\Omega_2}{nh^3}$.

Theorem 1.3(a) and (b) are analogous to the asymptotic results for estimation of kernel density of predicted (or generated) variables studied in Sperlich (2009). For kernel density estimation in which the variables are nonparametrically predicted, Sperlich (2009) shows that the bias and variance of the density estimator are augmented by an additive factor that is proportional to the bias and variance of the generated predictor, respectively. In other words, the bias of the predictor only affects the bias of the density estimator, the contribution of the variance of the predictor to the density estimator is only through the variance. Analogous results apply to our estimator of β_0. The deterministic bias of $\hat{\beta}$ is of order $O(h^\nu + h^{\nu-2}g^r)$ and the variance is of order $O_p\big(\frac{1}{nh} + \frac{1}{n^2h^3g^{d_z}}\big)$.

Theorem 1.3(c) is a generalization of Theorem 2 in Horowitz (1992) to the partially linear model with a nonparametrically generated regressor for general quantiles. The form of the first component Ω_1 of the asymptotic variance has intuitive interpretations: (1) The terms $\tau(1-\tau)$ and $F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})$ are usual in standard quantile regression, implying that the estimator is more precise in the tails and less precise in regions of low density; (2) The presence of $f_{m_\tau}(0 \mid \tilde{W})$ follows from the fact that the dependent variable Y is binary and $f_{m_\tau}(0 \mid \tilde{W})$ is used to approximate $\mathbb{E}[\mathbf{1}\{X'\beta_0 + \lambda_\tau(V) = 0\}] = \Pr(X'\beta_0 + \lambda_\tau(V) = 0)$. This is the main reason that leads to the nonparametric rate of convergence of $\hat{\beta}$ to β_0. As a consequence, the terms $R(K^{(1)})$ and $f_{m_\tau}(0 \mid \tilde{W})$ are analogous to those in standard kernel density estimation; (3) The $q(\tilde{W})q(\tilde{W})'$ term (implicitly through $\varphi(\tilde{W})\varphi(\tilde{W})'$) arises stemming from the fact that the model implies a partially linear form $X'\beta_0 + \lambda_\tau(V)$ for the regression function.

Observe that, under the assumption that the control variable $\{V_i\}_{i=1}^n$ is observed, based on Theorem 1.3 we can derive the asymptotic distribution of the estimator $\tilde{\beta}$ defined in (1.17) for the finite-dimensional parameters in partially linear binary regression quantiles.[30] This new result, as stated in Corollary 1.4 below, extends the single-index specification in smoothed binary regression quantiles in Horowitz (1992) and Kordas (2006) to the partially linear specification, which serves as an important compromise between the desire for parametric estimation precision and for nonparametric flexibility. Moreover, if the control variables $\{V_i\}_{i=1}^n$ are *parametrically* generated or if the bandwidths h and g are appropriately chosen satisfying the conditions in Assumption 1.12 below, it can be shown that the replacement of $\{V_i\}_{i=1}^n$ with its estimate $\{\hat{V}_i\}_{i=1}^n$ does not affect the first-order asymptotic distribution of the final estimator. Namely, the final estimator $\hat{\beta}$ has the same asymptotic distribution as it would have if $\{V_i\}_{i=1}^n$ were known. We summarize these results in the following corollaries below.

[30]Partially linear models have received much attention in the mean, quantile, and censored regression contexts. But we are not aware of the asymptotic results in the literature for estimating binary response models using partially linear specifications.

Corollary 1.4 (Partially Linear Binary Regression Quantiles)**.** *Define* $\mu \equiv \lim_{n\to\infty} nh^{2\nu+1}$. *Suppose Assumptions SPID, 1.1-1.10 hold and, additionally,* $\{V_i\}_{i=1}^n$ *were observed. Then the smoothed estimator* $\bar{\beta}^* \equiv A\bar{\theta}^*_{n\kappa}$ *of the finite-dimensional parameters* β_0 *based on the estimation procedure* (1.19) *in the partially linear binary regression quantiles:* $Q_\tau(Y_i \mid X_i, Z_i) = \mathbf{1}\{X_i'\beta_0 + \lambda_\tau(V_i) > 0\}$ *has the asymptotic distribution*

$$\sqrt{nh}(\bar{\beta}^* - \beta_0) \xrightarrow{d} N(\mu^{1/2}B_1, \Omega_1),$$

where B_1 *and* Ω_1 *are given in Theorem 1.3(a) and (b).*

The asymptotic distribution of the estimator $\tilde{\beta}$ has a similar form to the estimator $\hat{\beta}$, as stated in Theorem 1.3(c), except for the absence of the B_2 and Ω_2 components since no first-stage estimator of $\{V_i\}_{i=1}^n$ is needed.

Assumption 1.11 (Parametrically Generated Regressor)**.** *The first-stage estimator* $\hat{V}_i$ *is parametrically estimated, namely, the terms* b_v *and* σ_{ζ_v} *are both uniformly bounded by* $O(n^{-1/2})$.

Assumption 1.12 (Relative Rates of Bandwidths)**.** *The smoothing parameter sequences* $h = h_n$ *and* $g = g_n$ *go to zero as* $n \to \infty$ *and satisfy that* $nh^{2\nu-3}g^{2r} \to 0$ *and* $n^3h^5g^{2d_z} \to \infty$.

Assumption 1.12 insures that the bias and variance of the first-stage estimator of $\{V_i\}_{i=1}^n$ are sufficiently small, i.e., the first-stage estimation error is of smaller order than $(nh)^{-1/2}$. Furthermore, if the bandwidths h and g are assumed to satisfy $h \propto n^{-1/(2\nu+1)}$ and $g \propto n^{-\rho_g}$, then this assumption requires that

$$\frac{2}{r(2\nu+1)} < \rho_g < \frac{3\nu-1}{(2\nu+1)d_z}, \tag{1.25}$$

which in turn requires that the order of the kernel function $K^{(1)}$, ν, and the dimension of Z, d_z, have to satisfy the condition: $d_z < 3r - r/\nu$ (leading to $d_z < 5$ if $\nu = r = 2$ for instance), so that the interval in (1.25) is not empty. Since the optimal bandwidth of the kernel estimator of V satisfies $\rho_g = 1/(2r + d_z)$, it is interesting to note that there is a variety of combinations of ν, r, and d_z, for example, when $\nu = 3$, $r = 2$, and $d_z = 2$ such that the following inequalities hold

$$\frac{2}{r(2\nu+1)} < \frac{1}{2r+d_z} < \frac{3\nu-1}{(2\nu+1)d_z},$$

implying that in such cases undersmoothing, which is common in the semiparametric estimation literature, is not needed to insure the bias of the first-stage estimator to vanish sufficiently rapidly. On the other hand, when $\nu = 4$, $r = 2$, and $d_z = 5$ for example, the undersmoothing first-stage bandwidth can be chosen to satisfy $\rho_g = 1/(2r + d_z - 1)$.

Corollary 1.5. *Suppose Assumptions SPID, 1.1-1.10, and, additionally, either 1.11 or 1.12 hold. Then* $\sqrt{nh}(\hat{\beta} - \beta_0)$ *has the same asymptotic distribution as that in Corollary 1.4.*

Remark 1.3. Comparing the result of Theorem 1.3(c) to that of Corollary 1.5, we can see that when the smoothing parameters h and g are chosen such that $n^3h^5g^{2d_z} = O(1)$, Ω contains an additional term $ng^{d_z}\Omega_2$.

Remark 1.4. It is also interesting to compare Theorem 1.3 with the existing asymptotic theory in the literature. It is known that the rate of convergence of the estimator for binary response models with the median restriction is slower than the $n^{-1/2}$ rate of semiparametric single-index estimators under the independence assumption, i.e., the error terms are independent of the explanatory variables.[31] For these semiparametric single-index estimators converging at a parametric $\sqrt{n}$ rate, the generated regressor problem in general affects the asymptotic variance but not the $\sqrt{n}$ convergence rate of the estimator, unless the asymptotic orthogonality conditions between the preliminary and final estimators are satisfied (e.g., Assumption N(c) in Andrews (1994) for the "MINPIN" estimators).

Remark 1.5. Under Assumption 1.12, our asymptotic results show that the first-stage estimation error of the generated regressor $\{V_i\}_{i=1}^n$ does not appear in the asymptotic distribution of the final estimator. Specifically, the centered, suitably scaled estimator of the finite-dimensional parameters β_0 has the same asymptotic distribution that it would have if the nonparametric components λ_τ and $\{V_i\}_{i=1}^n$ were known. This property resembles many other multi-stage nonparametric procedures involving generated regressors in which the estimation error of the generated regressors does not contribute to the asymptotic variance of

[31] Chamberlain (1986) has proven that the semiparametric efficiency bound is zero for the exogenous binary response model if only the median restriction is assumed.

the final estimator, provided the smoothing parameters in the first and second steps are well chosen, see, for example, Su and Ullah (2008) and Sperlich (2009).

Asymptotic normality given in Theorem 1.3 can be used to carry out asymptotic inference on β_0. To this end, the value of μ has to be determined and consistent estimators of H, B, and Ω must be constructed. The (infeasible) optimal bandwidth is of the form $h^* = C \times n^{-1/(2\nu+1)}$. Then under Assumption 1.12 the optimal bandwidth $h^* = \mu^{*1/(2\nu+1)} n^{-1/(2\nu+1)}$ is set by minimizing the asymptotic mean square error:

$$\begin{aligned}
\text{AMSE} &\equiv \mathbb{E}[(\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0)] \\
&= \text{trace}\Big[\mathbb{E}[(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)']\Big] \\
&= \text{trace}\Big[n^{-2\nu/(2\nu+1)} \mu^{-1/(2\nu+1)} \Omega_1 + \mu^{2\nu/(2\nu+1)} B_1 B_1'\Big].
\end{aligned}$$

Analogous to Horowitz (1992) and Kordas (2006), it can be shown that

$$\mu^* = \text{trace}(\Omega_1)/(2\nu B_1' B_1).$$

Then in applications one can choose the bandwidth h by the plug-in method, as suggested by Horowitz (1992). Namely, the estimate $\hat{\mu}$ is obtained based on μ^* by replacing Ω_1 and B_1 with their estimates. We turn to consistent estimators of H, B_1, and Ω_1. Define

$$\hat{g}_{n\kappa i} = t(\hat{W}_i)[Y_i - (1-\tau)] K_h^{(1)}\Big(P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}\Big) P_\kappa(\bar{W}_i).$$

Let $h_\epsilon = h_{\epsilon n} \propto n^{-\epsilon/(2\nu+1)}$ for some $0 < \epsilon < 1$ and $\hat{\sigma}_v^2$ a consistent estimator of σ_v^2. One can consistently estimate the asymptotic bias B_1 by $-\hat{\mu}^{1/2} A \hat{\hat{H}}_{n\kappa}^{-1} \hat{B}_{n\kappa}$ and the asymptotic variance Ω by

$$\hat{\Omega}_{n\kappa} \equiv A \hat{\hat{H}}_{n\kappa}^{-1} \left(\hat{\Sigma}_{n\kappa} + \hat{\Gamma}_{n\kappa} \hat{\sigma}_v^2 \hat{\Gamma}_{n\kappa}'\right) \hat{\hat{H}}_{n\kappa}^{-1} A',$$

where

$$\hat{\hat{H}}_{n\kappa} = \frac{1}{n}\sum_{i=1}^{n} t(\hat{W}_i)[Y_i - (1-\tau)]K_h^{(2)}\big(P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}\big)P_\kappa(\bar{W}_i)P_\kappa(\bar{W}_i)',$$

$$\hat{B}_{n\kappa} = (h_\epsilon)^{-\nu} n^{-1}\sum_{i=1}^{n} t(\hat{W}_i)[Y_i - (1-\tau)]K_{h_\epsilon}^{(1)}\big(P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}\big)P_\kappa(\bar{W}_i),$$

$$\hat{\Sigma}_{n\kappa} = hn^{-1}\sum_{i=1}^{n} \hat{g}_{n\kappa i}\hat{g}'_{n\kappa i}, \text{ and}$$

$$\hat{\Gamma}_{n\kappa} = n^{-1}\sum_{i=1}^{n} t(\hat{W}_i)[Y_i - (1-\tau)]K_{h_\epsilon}^{(2)}\big(P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}\big)\frac{\mathrm{d}\hat{\lambda}_\tau(\hat{V}_i)}{\mathrm{d}v}P_\kappa(\bar{W}_i).$$

Assumption 1.13. *For power series* $\frac{\nu+1}{(2\nu+1)r} < \rho_\kappa < \frac{2\nu+3}{9(2\nu+1)}$ *and for splines* $\frac{\nu+1}{(2\nu+1)r} < \rho_\kappa < \frac{2\nu+3}{5(2\nu+1)}$.

The following theorem establishes the consistency of the estimators for the asymptotic bias and variance.

Theorem 1.4. *Under Assumptions SPID, 1.1-1.10, and 1.12,* $\hat{\hat{H}}_{n\kappa} \xrightarrow{p} H$, $-\hat{\mu}^{\nu/(2\nu+1)} A\hat{\hat{H}}_{n\kappa}^{-1}\hat{B}_{n\kappa} \xrightarrow{p} B_1$ *and* $\|\hat{\Sigma}_{n\kappa} - \Sigma_\kappa\| = o_p(1)$. *It follows that* $\|\hat{\Omega}_{n\kappa} - \Omega\| = o_p(1)$ *as* $n \to \infty$.

1.6 Extension to Nonparametric Binary Response Models with Endogeneity

In this section, we discuss a straightforward extension to address the endogeneity problem in a fully nonparametric binary response model. Formal asymptotic property results are left for future research. Consider a nonparametric binary response model with a latent outcome specification

$$Y^* = g_0(X) + U, \tag{1.26}$$

where g_0 is assumed to satisfy mild regularity conditions but does not belong to a known, finite-dimensional parametric family. We emphasize here that models (1.2)-(1.4) and (1.26) do not impose any parametric structure either on the systematic component g_0 or the distribution of the random term U.

While imposing the additive separability structure on the model, latent outcome specification (1.26) is still important because of the sensitivity of the parametric and semiparametric

estimators to misspecification of functional form. However, fully nonparametric identification and estimation of binary response models have received relatively little attention in the literature. Matzkin (1992) is the first to deal with nonparametric and distribution-free binary response models, which can also be viewed as a special case of Han's (1987) generalized regression model: $Y = H(g_0(X), U)$, where H is a known function. Interestingly, in the settings of binary regressions, Nadaraya-Watson (with $\mathbb{E}(Y \mid X = x) = \theta_x$) and local linear (with $\mathbb{E}(Y \mid X = x) = x'\theta_x$) approaches should be regarded as local parametric methods where θ_x is a unknown coefficient vector that is allowed to vary with x.[32]

The model (1.2)-(1.4) and (1.26) implies that the τth conditional quantile of Y^* given X and V reduces to an additive function with a generated regressor, namely, $m^*_\tau(X, V) = g_0(X) + \lambda_\tau(V)$.[33] To recover the additive components g_0 and λ_τ, we conjecture that, in addition to the typical exclusion restriction to separate the influence of g_0 from the influence of λ_τ on the object $m^*_\tau(X, V)$, more structure on g_0 is needed. For example, a continuous and additive "special regressor," say X_c, with a full support (conditional on V), i.e., $g_0(X) = X_c + \tilde{g}_0(\tilde{X}_{-c})$ may be sufficient for identification of $\tilde{g}_0$.

If g_0 is identified, we can then develop an estimator of the nonparametric structural function g_0. The first-stage control variable estimator $\hat{V}$ is the same as before. To describe the second-stage estimator, let $P_{g\kappa}(w) = \big(p_1(x), \ldots, p_\kappa(v)\big)$ be a $\kappa \times 1$ vector of approximating functions of $w = (x', v)'$ with each component depending either on x or on v, but not both. That is, we use series methods to impose additive separability, as previously mentioned.[34] Analogously, the series coefficient estimates are obtained by maximizing the sample objective

[32]The usual Nadaraya-Watson or local linear methods are not well suited for nonparametric regression with a binary dependent variable, as the same spirit as linear probability models usually performing poorly compared to probit or logit models. Frandouml (2006) studies estimation of local likelihood logit regression with $\mathbb{E}(Y \mid X = x) = 1/(1+e^{x'\theta_x})$ and shows that the local logit estimator has better finite-sample performance than Nadaraya-Watson, local linear, and Klein-Spady alternatives.

[33]Analogous to the semiparametric case, as we discussed before, if the latent variable Y^* were known, the model $Q_\tau(Y^* \mid X, Z) = g_0(X) + \lambda_\tau(V)$ becomes the quantile version of nonparametric triangular simultaneous equations models analyzed by Newey, Powell, and Vella (1999).

[34]Alternatively, one can use the partial mean (or marginal integration) approach to estimate the additive component function $g_0(x)$ of $m^*_\tau(x, v)$. That is, $\int m^*_\tau(x, v) l(v) dv = g_0(x) + \int \lambda_\tau(v) l(v) dv$ where l is a nonnegative function satisfying $\int l(v) dv = 1$.

function (1.18) with $P_{g\kappa}(\hat{W}_i)'\theta$ in place of $P_\kappa(\hat{W}_i)'\theta$. By collecting those terms that depend only on x and those that depend only on v, estimators of the additive components $g_0(x)$ and $\lambda_\tau(v)$ can be constructed.

As before, we can deal with the generated regressor problem by using an asymptotic expansion analogous to that given in Theorem 1.2. Yet, the derivation of the rate of convergence of $\hat{g}$ to g_0 is more complicated than the semiparametric case when g_0 is unknown up to finite-dimensional parameters. This is because the estimation variance and bias caused by nonparametrically estimating g_0 invalidate asymptotic properties in the semiparametric setting. Specifically, since g_0 is nonparametric, the number of regressors κ in series estimation grows with the sample size and hence decreases the approximation error at a cost of slowing down the convergence rate of the estimation variance. This is in contrast to the semiparametric case which treats κ fixed and therefore changes the asymptotic properties of the estimator.

Using series methods without imposing additivity, the proposed estimation procedure can be possibly further extended to the triangular model without additivity, as described in (1.12). Specifically, it is straightforward to construct an estimator of m_τ^* in which the control variable V enters the model in a nonseparable way. The estimator of m_τ^* may serve as an intermediate object in forming estimators of some parameters of interest such as the average structural function. We leave this possibility to future research.

1.7 Monte Carlo Simulation

1.7.1 Simulation Designs

To illustrate the implementation of the proposed estimation procedure and evaluate the finite-sample properties of the estimator $\hat{\beta}$ in the semiparametric model, this section conducts a simulation study and reports small-scale Monte Carlo results. We then consider an empirical illustration in the next section. To begin, we are concerned with estimating the

scalar parameter β_0 based on the following data generating process (DGP)

$$
\begin{aligned}
Y_1 &= \mathbf{1}\{X + \beta_0 Z_1 > U\}, \\
X &= \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \sigma(Z_2)\eta, \\
U &= \tilde{\lambda}_\tau(\eta) + U^*,
\end{aligned}
$$

with the exogenous explanatory variable $Z_1 \sim N(1,1)$, the excluded instrument $Z_2 \sim N(0,1)$, the first-stage error $\eta \sim N(0, 4/9)$, and the control function $\tilde{\lambda}_\tau(\eta)$. We consider two different DGPs, depending on the distribution of U^* and the functional forms of $\sigma(z_2)$ and the control function:

$$
\text{DGP 1}: \begin{cases} \tilde{\lambda}_\tau(\eta) = \eta, \\ U^* \sim N(0,1), \\ \sigma(z_2) = 1, \end{cases}
$$

and

$$
\text{DGP 2}: \begin{cases} \tilde{\lambda}_\tau(\eta) = -0.4 + 0.5\eta + \phi(\eta), \\ \phi(\eta) = \exp(-(\eta - 1)^2), \\ U^* \sim N(0, \exp(0.1 + 0.5 Z_1)/\sqrt{4.5}), \\ \sigma(z_2) = \exp(z_2/2). \end{cases}
$$

The first design implies that (U, η) are bivariate normally distributed and the control function is linear: $\tilde{\lambda}_\tau(\eta) = \eta$ under which the model is correctly specified for the 2SProbit estimator. The second design allows for the presence of heteroskedasticity in both outcome and first-stage equations and a control function with a nonlinear component $\phi(\eta)$. The function $\tilde{\lambda}_\tau(\eta) = -0.4 + 0.5\eta + \phi(\eta)$ is taken with slight modification from Lee (2007). These two designs are chosen to share some common features. In particular, it holds that $\mathbb{E}(X) \approx 0$, $\text{Var}(X) \approx 1$, $\mathbb{E}(U^*) \approx 0$, $\text{Var}(U^*) \approx 1$, $\mathbb{E}(U) \approx 0$, and $\text{Var}(U) \approx 1.5$.

To investigate the potential gain from using our approach, we compare the proposed estimator with the following two estimators available in the literature:

(a) The two-stage probit estimator of Smith and Blundell (1986) or Rivers and Vuong (1988).

(b) The two-stage least square (i.e. linear probability) estimator that is frequently used in applied work.[35]

The above alternative estimators and the estimator introduced in this chapter are referred to in this section as 2SProbit, 2SLS, and TBRQ, respectively. For each DGP, each estimator is calculated using three sample sizes: $n = 250$, 500 and $1,000$. We set the true values of parameters: $\beta_0 = 1$ and $(\alpha_0, \alpha_1, \alpha_2) = (1, 2/3, 2/3)$. The number of Monte Carlo replications per experiment is set to $B = 500$. We consider the median case: $\tau = 0.5$ throughout this section.

For the TBRQ estimator, we assume that all instruments Z are continuously distributed[36] and in the first stage estimate the conditional CDF $F_{X_1|Z}(X_{1i} \mid Z_i)$ at X_{1i} and Z_i by using the Nadaraya-Watson estimator (smoothing the dependent variable and covariates)[37] of the form

$$\hat{V}_i = \hat{F}_{X_1|Z}(X_{1i} \mid Z_i) = \frac{\sum_{j=1, j\neq i}^{n} G_g(Z_j - Z_i) G_{g_0}(X_{1j} - X_{1i})}{\sum_{j=1, j\neq i}^{n} G_g(Z_j - Z_j)},$$

where $G_{g_0}(u_i) = G(u_i/g_0)$ with G and g_0 being a univariate kernel function and the bandwidth associated with X_{1i}, respectively and $G_g(u) = \Pi_{i=1}^{d_z=2}(G(u_i/g_i)/g_i)$ is a d_z-dimensional kernel function constructed by the product of the univariate kernel function G and a vector of bandwidth $g = (g_1, g_2)$. The second step carries out the series smoothed maximum score estimation procedure using the regressors $(X_1, Z_1, \hat{V})$ and requiring the choice of the basis function and the number of series terms κ to approximate the unknown function $Q_{0.5}(U \mid V) \equiv \lambda_{0.5}(V)$. In the experiments, we use piece-wise quadratic B-splines as base functions to approximate the unknown control function $\lambda_{0.5}(V)$. The smoothing function K is taken

[35] Theoretically speaking, applying the two-stage least square to endogenous binary response models is not appropriate since its consistency requires orthogonality conditions that arise in the linear models. Moreover, 2SLS is incompatible with the nature of the observed data.

[36] In applied settings, the presence of a mix of discrete and continuous instruments is frequently encountered. In that case, to deal with discrete covariates, one can use the conventional frequency method or the new approach suggested by Li and Racine (2008) by smoothing the discrete covariates. See Li and Racine (2008) for details.

[37] From the theoretical point of view, the advantage of the smoothed version of the CDF estimator is the improvement of the higher-order estimation efficiency in the sense that when using smoothing along with the optimal bandwidths, the higher order terms associated with the bandwidth for the dependent variable in the integrated mean square error are of the smaller order as the covariates are of higher dimension.

to be the cumulative normal distribution function. Our estimation procedure involves three kinds of smoothing parameters: one (g) for nonparametric conditional CDF estimation, one (κ) for the series approximation to the conditional quantile λ_τ, and one (h) for the smoothing function K. When estimating the conditional CDF in the first stage, to reduce the computational cost we follow Silverman (1986) by using the bandwidths (g_0, g_1, g_2) of the form $g_i = 1.06 \times \hat{\sigma}_i \times n^{-\rho_g}$, $i = 0, 1, 2$, where $\hat{\sigma}$ is a robust measure of spread given by $\hat{\sigma} = \min(\text{sample standard deviation, interquartile range}/1.349)$ for the respective variable and $\rho_g = 1/7$ is the convergence rate of the bandwidth g_i, $i = 0, 1, 2$. Data-based methods for choosing the first-stage bandwidth are precluded in the experiments since they entail lengthy computing times. Note that in our simulation designs, the combination of $\nu = 4$, $r = 2$, and $d_z + 1 = 3$ due to smoothing the X_1-direction satisfies equation (1.25) implied by Assumption 1.12, so that undersmoothing the first-stage bandwidth g_i is not required. We choose the values of the number of the series expansion terms κ that satisfy Assumption 1.5(a) and allow 4 different values of κ for each sample size. Based on the formula $\kappa = C_\kappa n^{\rho_\kappa}$, we choose $C_\kappa = 1$ and $\rho_\kappa = 1/4$, leading to approximately $\kappa = 4,\ 5,\ 6$ for $n = 250$, 500, and 1000, respectively.

Regarding the computational methods, it is known that the smoothed score function (1.18) can have many local optima and the use of a global optimization method is necessary. The method to compute the parameter estimates is based on generalized simulated annealing (GSA), proposed by Tsallis and Stariolo (1996), which generalizes both the fast simulated annealing and classical simulated annealing procedures. All experiments were carried out in R.

1.7.2 Simulation Results

The performance of each estimator of $\beta_0 = 1$ for each design, with the sample size of 100, 500, and 1000, is summarized in Tables 1.1-1.2, which report the bias, standard deviation (SD), root-mean-square error (RMSE), median absolute deviation (MAD), and the 25%, 50%, and 75% sample quantiles.

Tables 1.1-1.2 indicate some general findings. First, in Design 1 where the parametric model is correctly specified, 2SProbit performs best, as expected, whereas the TBRQ exhibits the nonnegligible bias and has larger RMSE than the 2SProbit for all values of κ under consideration when $n = 250$. This is to be expected as the TBRQ uses several nonparametric estimates, which can be inaccurate for the small sample size. However, the bias of the TBRQ decreases dramatically as the sample size increases from $n = 250$ to 500 and the performance of the TBRQ seems to be satisfactory for $n = 1000$. In addition, the biases of TBRQ accounts for relatively small fraction of MSE. Also note that for some certain range of κ in the sample size considered, the choice of κ has an important effect on the performance of the TBRQ, in the sense that both bias and variance decrease as the number of approximating functions increases. This implies that the performance of the TBRQ estimator can be improved by choosing the number of series terms effectively. This is indeed an important topic for future research. For Design 2, the results in Table 1.2 reflect the benefits of the TBRQ when heteroskedasticity is present in both outcome and first-stage equations. It's somewhat surprising that, unlike the 2SProbit, the 2SLS in Design 2 performs relatively well in terms of its RMSE and MAD. A possible explanation for this result is that there may be a certain offsetting effect on the imprecision caused by the outcome and first-stage model misspecifications, because in the experiments not reported here, the 2SLS exhibits large opposite biases when heteroskedasticity is present either in the outcome or in the first-stage model but not both. Moreover, the 2SLS has low variance but its bias does not vanish as the sample size increases. This is similar to the simulation results in Rothe (2009). On the other hand, the 2SProbit estimator performs badly, exhibiting the biases in the neighborhood of 0.5 regardless of the sample size, indicating its inconsistency under misspecifications such as heteroskedasticity and the nonlinear control function. In contrast, in terms of RMSE, our TBRQ outperforms 2SProbit uniformly over different values of κ in sample sizes of $n = 500$ and $n = 1000$ in Design 2. In summary, the results in simulation experiments indicate that the proposed TBRQ estimator works reasonably well in finite samples.

We end this section by discussing some further efforts needed to investigate the proposed TBRQ estimator. First, it is desirable to compare the TBRQ with the other existing semiparametric estimators in endogenous binary response models, proposed by Blundell and Powell (2004), Rothe (2009), and Krief (2011), under the designs of DER and QER. Secondly, more exercises are needed to investigate the sensitivity of the TBRQ to the first- and second-stage bandwidth selection, different choices of base functions, and different distributional specifications of U^* such as mixed or asymmetric distributions.

1.8 Conclusion

Endogeneity in the nonlinear econometric models is an important but difficult problem. The primary objective of this chapter is to develop a semiparametric (or distribution-free) estimator for the binary response model under the presence of endogeneity and general forms of unknown heteroscedasticity. To do so we have presented a two-step control function procedure by imposing a weak nonparametric quantile restriction. Following the insights of Newey, Powell, and Vella (1999), Lee (2007), Newey (2009) and Horowitz (1992), our estimation procedure is based upon series approximation and kernel-based smoothing techniques to impose additive separability restrictions and to facilitate the asymptotic analysis, respectively. The semiparametric estimator is shown to be consistent and asymptotically normally distributed. We have also given the conditions under which the estimator has the same asymptotic distribution that it would have if the nonparametric components were known. Our approach can be extended in a straightforward fashion to the fully nonparametrically binary response model with endogeneity, whereas existing approaches are not applicable to such cases.

The work here leaves open several important directions for future research. One of these is to find a way to carry out consistent model specification tests for endogeneity and tests for additivity. Specifically, the former is to test whether or not the control function varies with the control variable. For the latter, several tests of additivity have been proposed in the context of nonparametric conditional mean models (with a link function), see Horowitz (2012)

for a review. It would be interesting to develop analogous tests for a partially linear and additive specification against a general nonparametric nonseparable alternative in our model. Such tests will be useful to test if the control variable enters the model in an additive fashion. Secondly, following the idea that was initially suggested by Koenker and Bassett (1982), imposing quantile restrictions here could provide a Kolmogorov-Smirnov or Cramer-von-Mises type testing procedure for the DER or the presence of heteroskedasticity by comparing quantile coefficients on different parts of the conditional distribution. Additionally, it would be desirable to investigate the efficiency gain by (optimally) combining information over different quantiles. Another challenging and important task is to find data-based methods for optimally choosing the number of the series expansion terms κ in practice. Finally, it would also be useful to generalize the approach of this chapter to the model with multinomial responses.

Table 1.1 Simulation results: Design 1

		Bias	SD	RMSE	MAD	25%	50%	75%
n=250	2SProbit	0.031	0.307	0.309	0.276	0.822	0.999	1.212
	2SLS	0.032	0.339	0.340	0.302	0.817	0.987	1.229
	TBRQ ($\kappa = 4$)	0.903	2.216	2.393	0.595	0.875	1.295	2.037
	TBRQ ($\kappa = 5$)	0.558	1.165	1.292	0.578	0.864	1.277	1.842
	TBRQ ($\kappa = 6$)	0.497	0.966	1.086	0.589	0.870	1.309	1.832
	TBRQ ($\kappa = 7$)	0.452	0.732	0.860	0.606	0.932	1.320	1.773
n=500	2SProbit	0.008	0.206	0.206	0.204	0.860	0.998	1.133
	2SLS	0.013	0.222	0.223	0.227	0.850	1.004	1.156
	TBRQ ($\kappa = 5$)	0.266	0.616	0.671	0.472	0.866	1.140	1.516
	TBRQ ($\kappa = 6$)	0.254	0.581	0.634	0.459	0.855	1.155	1.479
	TBRQ ($\kappa = 7$)	0.257	0.472	0.537	0.423	0.933	1.186	1.510
	TBRQ ($\kappa = 8$)	0.271	0.417	0.498	0.369	0.975	1.205	1.471
n=1000	2SProbit	0.004	0.132	0.132	0.136	0.910	1.003	1.092
	2SLS	0.007	0.153	0.154	0.152	0.900	0.994	1.110
	TBRQ ($\kappa = 6$)	0.116	0.334	0.353	0.319	0.871	1.070	1.310
	TBRQ ($\kappa = 7$)	0.133	0.328	0.354	0.296	0.894	1.082	1.311
	TBRQ ($\kappa = 8$)	0.175	0.299	0.347	0.254	0.974	1.133	1.327
	TBRQ ($\kappa = 9$)	0.235	0.272	0.359	0.230	1.051	1.196	1.380

Table 1.2 Simulation results: Design 2

		Bias	SD	RMSE	MAD	25%	50%	75%
n=250	2SProbit	0.581	0.573	0.816	0.507	1.189	1.478	1.888
	2SLS	0.064	0.358	0.363	0.317	0.702	0.888	1.138
	TBRQ ($\kappa = 4$)	0.603	1.554	1.667	0.543	0.803	1.149	1.824
	TBRQ ($\kappa = 5$)	0.361	0.939	1.007	0.512	0.804	1.133	1.622
	TBRQ ($\kappa = 6$)	0.321	0.777	0.841	0.522	0.806	1.143	1.659
	TBRQ ($\kappa = 7$)	0.331	0.666	0.743	0.530	0.889	1.193	1.652
n=500	2SProbit	0.535	0.357	0.643	0.313	1.305	1.492	1.725
	2SLS	0.078	0.223	0.237	0.206	0.769	0.904	1.048
	TBRQ ($\kappa = 5$)	0.152	0.549	0.570	0.363	0.819	1.050	1.316
	TBRQ ($\kappa = 6$)	0.153	0.536	0.557	0.364	0.818	1.057	1.315
	TBRQ ($\kappa = 7$)	0.148	0.445	0.469	0.357	0.849	1.071	1.331
	TBRQ ($\kappa = 8$)	0.171	0.370	0.407	0.328	0.915	1.116	1.369
n=1000	2SProbit	0.521	0.228	0.569	0.214	1.371	1.511	1.659
	2SLS	0.079	0.161	0.180	0.159	0.804	0.911	1.020
	TBRQ ($\kappa = 6$)	0.019	0.289	0.289	0.252	0.824	0.978	1.188
	TBRQ ($\kappa = 7$)	0.035	0.281	0.284	0.262	0.846	0.999	1.210
	TBRQ ($\kappa = 8$)	0.066	0.264	0.272	0.226	0.879	1.021	1.195
	TBRQ ($\kappa = 9$)	0.117	0.222	0.251	0.214	0.956	1.098	1.245

1.9 Appendix

The Interval Estimator of Response Probabilities

In addition to the finite-dimensional coefficients β_0, the choice probability $\Pr(Y = 1 \mid X)$ and marginal effects of the covariates say x_1 , $\partial \Pr(Y = 1 \mid X)/\partial x_1$, are also key parameters of interest in the binary response model. General speaking, to point identify and consistently estimate $\Pr(Y = 1 \mid X)$, one needs to strengthen the QER to the DER. To see this, recall that under the QER the model imply

$$\Pr\Big(Y = 1 \Big| X'\beta_0 + \lambda_\tau(V) \begin{array}{l} > \\ = 0 \\ < \end{array}\Big) \begin{array}{l} > \\ = 1 - \tau. \\ < \end{array}$$

The intermediate conditional probability $\Pr(Y = 1 \mid X, V)$ is given by

$$\int_{\mathbb{R}} \mathbf{1}\{X'\beta_0 + U > 0\} \mathrm{d}F_{U|X,V} = \int_0^1 \mathbf{1}\{X'\beta_0 + \lambda_\tau(V) > 0\} \mathrm{d}\tau,$$

where the equality follows from a change-of-variable: $U \to F^{-1}_{U|X,Z}$ by assuming that the QER holds for *all* $\tau \in (0, 1)$. Thus it is clear that $\Pr(Y = 1 \mid X, V)$ is point identified only under the DER. On the other hand, $\Pr(Y = 1 \mid X, V)$ is not identified if QER holds for just a single quantile. Between these two extreme cases, partial identification arises by looking at a number of different quantiles.

Following Blundell and Powell (2003), we define the average structural function, $\mathrm{ASF}(x)$, that summarizes the effect on the whole distribution by integrating $\Pr(Y = 1 \mid X = x, V)$ over the marginal distribution of V, i.e.,

$$\mathrm{ASF}(x) \equiv \int \Pr(Y = 1 \mid X = x, V) \mathrm{d}F_V.$$

Using the idea similar to that of Kordas (2006), we summarize below how to partially identify and consistently estimate the interval of the choice probability by looking at a number of different quantiles. Fix $X = x$ and let a grid $T = \{\tau^1, \ldots, \tau^m : \tau^1 < \cdots < \tau^m\}$. Define

$$\hat{\tau}_i^{\min}(x) \equiv \arg\min_{\tau \in T} \{\tau : \hat{m}^*_\tau(x, \hat{V}_i) > 0\},\ i = 1, \ldots, n,$$

where $\hat{m}^*_\tau(x, \hat{V}_i) = x'\hat{\beta} + \hat{\lambda}_\tau(\hat{V}_i)$.

An interval estimate of $F(x, V_i) \equiv \Pr(Y_i = 1 \mid X = x, V_i)$ can be formed by

$$1 - \hat{\tau}_i^{\min}(x) \leq \hat{\Pr}(Y_i = 1 \mid X = x, \hat{V}_i) < 1 - \hat{\tau}_i^{\min -1}(x).$$

Then given the interval estimate of $\Pr(Y_i = 1 \mid X = x, V_i)$, the interval estimate of $\mathrm{ASF}(x)$ is immediately obtained by

$$1 - \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i^{\min}(x) \leq \frac{1}{n}\sum_{i=1}^{n} \hat{\Pr}\big(Y_i = 1 \mid X = x, \hat{V}_i\big) < 1 - \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i^{\min -1}(x).$$

Despite the $\mathrm{ASF}(x)$ is not point identified under the weak QER, the parameter that may be of interest is the average quantile structural function (AQSF), which is an immediate extension of Hoderlein's (2009) average median structural function, defined as

$$\mathrm{AQSF}(x) = \int Q_\tau(Y \mid X = x, V) \mathrm{d}F_V = \int \mathbf{1}\{x'\beta_0 + \lambda_\tau(V) > 0\} \mathrm{d}F_V.$$

The identification of $\mathrm{AQSF}(x)$ may allow the researcher to examine and compare the structural effect for different quantiles.

The following appendix gathers the proofs of theorems in the main text and of lemmas used to prove the theorems. For the following proof, let C denote a generic positive constant that may be different in different uses. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a symmetric matrix A. For the sake of simplifying notation, we will write $\simeq$ for up to higher-order terms.

Proof of Theorem 1.1

The proof of identification is similar to the argument used in the exogenous binary response model by Manski (1988) and Horowitz (2009). Let $\bar{\gamma} \neq \gamma_0$ and $\bar{\lambda}_\tau \neq \lambda_\tau$ be any respective alternatives of parametric and nonparametric components satisfying the scale normalization, i.e., $\gamma_0 = (1, \beta_0')'$ and $\bar{\gamma} = (1, \bar{\beta}')'$. Define the set $S(\bar{\gamma}, \bar{\lambda}_\tau) \equiv \big\{(x, z) : \{x'\gamma_0 + \lambda_\tau(v) < 0 \leq x'\bar{\gamma} + \bar{\lambda}_\tau(v)\} \cup \{x'\bar{\gamma} + \bar{\lambda}_\tau(v) < 0 \leq x'\gamma_0 + \lambda_\tau(v)\}\big\}$. Since the first components of γ_0 and γ equal 1, the set $S(\bar{\gamma}, \bar{\lambda}_\tau)$ can be rewritten as $\big\{(x, z) : \{x_1 + \tilde{x}'\beta_0 + \lambda_\tau(v) < 0 \leq$

$x_1 + \tilde{x}'\bar{\beta} + \bar{\lambda}_\tau(v)\} \cup \{x_1 + \tilde{x}'\bar{\beta} + \bar{\lambda}_\tau(v) < 0 \leq x_1 + \tilde{x}'\beta_0 + \lambda_\tau(v)\}\big\}$. Since the distribution of X_1 is assumed to have everywhere positive density conditional on $\tilde{X} = \tilde{x}$ and $V = v$, the set $S(\bar{\gamma}, \bar{\lambda}_\tau)$ has positive probability whenever $-\tilde{x}'\bar{\beta} - \bar{\lambda}_\tau(v) < -\tilde{x}'\beta_0 - \lambda_\tau(v)$ or $-\tilde{x}'\beta_0 - \lambda_\tau(v) < -\tilde{x}'\bar{\beta} - \bar{\lambda}_\tau(v)$. This would occur if we can show $\Pr\big(\tilde{x}'\beta_0 + \lambda_\tau(v) = \tilde{x}'\bar{\beta} + \bar{\lambda}_\tau(v)\big) < 1$. It is equivalent to showing that $\Pr\big(\tilde{x}'(\beta_0 - \bar{\beta}) + (\lambda_\tau(v) - \bar{\lambda}_\tau(v)) = 0\big) = 1$ implies $\beta_0 = \bar{\beta}$ (and therefore $\gamma_0 = \bar{\gamma}$) and $\lambda_\tau = \bar{\lambda}_\tau$. To show this, by the differentiability assumption on λ_τ and v_0 and $\Delta\bar{\lambda}_\tau(V) \equiv \lambda_\tau(V) - \bar{\lambda}_\tau(V)$ is continuously differentiable, differencing the identity $\tilde{X}'(\beta_0 - \bar{\beta}) + \Delta\bar{\lambda}_\tau(V) = 0$ with respect to Z_{21} and X yields

$$\frac{\partial \Delta\bar{\lambda}_\tau(V)}{\partial V}\frac{\partial v_0(X_1, Z)}{\partial Z_{21}} = 0$$

$$\text{and } \frac{\partial \tilde{X}'(\beta_0 - \bar{\beta})}{\partial X} + \frac{\partial \Delta\bar{\lambda}_\tau(V)}{\partial V}\frac{\partial v_0(X_1, Z)}{\partial X} = 0,$$

respectively. By part (c) of Assumption 1.3, we have $\partial v_0(X_1, Z)/\partial Z_{21} \neq 0$ implying $\partial\Delta\bar{\lambda}_\tau(V)/\partial V = 0$. It then follows from the second equation that $\partial\tilde{X}'(\beta_0 - \bar{\beta})/\partial X = 0$. This along with the linear index restriction must be the case that $\tilde{X}'(\beta_0 - \bar{\beta}) = 0$. Finally, Assumption 1.3 (b) implies that this is only possible if $\beta_0 = \bar{\beta}$ and therefore $\lambda_\tau = \bar{\lambda}_\tau$. This completes the proof of the theorem.

Proof of Corollary 1.1

Recall that the model implies

$$Q_\tau(Y \mid X, Z) = \begin{cases} 0 & \text{if } P(Y = 1 \mid X, Z) \leq 1 - \tau \iff P(U > -X'\beta_0 \mid X, Z) \leq 1 - \tau \iff X'\beta_0 \leq -\lambda_\tau(V) \\ 1 & \text{if } P(Y = 1 \mid X, Z) > 1 - \tau \iff P(U > -X'\beta_0 \mid X, Z) > 1 - \tau \iff X'\beta_0 > -\lambda_\tau(V). \end{cases}$$

Define the population objective function $Q(\bar{\beta}, \bar{\lambda}_\tau) \equiv \mathbb{E}\big[\big((1 - \tau) - Y\big)\mathbf{1}\{X'\bar{\beta} + \bar{\lambda}_\tau(V) > 0\}\big]$ where $(\bar{\beta}, \bar{\lambda}_\tau)$ is a pair of generic elements in the respective parameter spaces. We want to show that $Q(\bar{\beta}, \bar{\lambda}_\tau)$ is uniquely minimized at $\bar{\beta} = \beta_0$ and $\bar{\lambda}_\tau = \lambda_\tau$. Note first that by the law of iterated expectations,

$$\begin{aligned} Q(\bar{\beta}, \bar{\lambda}_\tau) &= \mathbb{E}[((1 - \tau) - \mathbb{E}(Y \mid X, Z)) \cdot \mathbf{1}\{X'\bar{\beta} + \bar{\lambda}_\tau(V) > 0\}] \\ &= \mathbb{E}[((1 - \tau) - \Pr(Y = 1 \mid X = x, Z = z))] \cdot \mathbf{1}\{x'\bar{\beta} + \bar{\lambda}_\tau(v) > 0\}] \\ &= \mathbb{E}[((1 - \tau) - \Pr(U > -X'\beta_0 \mid X = x, Z = z))] \cdot \mathbf{1}\{x'\bar{\beta} + \bar{\lambda}_\tau(v) > 0\}] \end{aligned}$$

where $x = (x_1, z_1')'$ and $z = (z_1', z_2')'$ are arbitrarily chosen and $v = F_{X_1|Z}(x_1 \mid z)$. Notice that $\bar{\lambda}_\tau(v)$ is just a point given v (or x and z). Next consider $Q(\bar{\beta}, \bar{\lambda}_\tau)$ evaluated at $\bar{\beta} = \beta_0$ and $\bar{\lambda}_\tau(\cdot) = \lambda_\tau(\cdot)$. Since

$$\mathbf{1}\{x'\beta_0 + \lambda_\tau(v) > 0\} = \begin{cases} 1 & \iff x'\beta_0 > -\lambda_\tau(v) \iff \Pr(U > -X'\beta_0 \mid X = x, Z = z) > 1 - \tau \\ 0 & \iff x'\beta_0 \leq -\lambda_\tau(v) \iff \Pr(U > -X'\beta_0 \mid X = x, Z = z) \leq 1 - \tau \end{cases}$$

$$\iff \begin{cases} Q(\beta_0, \lambda_\tau) \text{ is negative} \\ Q(\beta_0, \lambda_\tau) \text{ is zero,} \end{cases} \tag{1.27}$$

the population objective function evaluated at the truth values of parameters, i.e., $Q(\beta_0, \lambda_\tau)$, must be nonpositive. Next Consider $Q(\beta, \lambda)$ being evaluated at any $\bar{\beta} \neq \beta_0$ or $\bar{\lambda}_\tau \neq \lambda_\tau$. We want to show that if the identification condition: $\Pr[\mathbf{1}\{X'\bar{\beta} + \bar{\lambda}_\tau(V) \geq 0\} \neq \mathbf{1}\{X'\beta_0 + \lambda_\tau(V) \geq 0\}] > 0$ holds for all $\bar{\beta} \neq \beta_0$ or $\bar{\lambda}_\tau \neq \lambda_\tau$, then we have

$$\mathbb{E}\big[\big((1-\tau) - Y\big)\mathbf{1}\{X'\bar{\beta} + \bar{\lambda}_\tau(V) > 0\}\big] > \mathbb{E}\big[\big((1-\tau) - Y\big)\mathbf{1}\{X'\beta_0 + \lambda_\tau(V) > 0\}\big].$$

To see this, fix $X = x$ and $Z = z$ such that $\mathbf{1}\{x'\beta_0 + \lambda_\tau(v) > 0\} = 1$ and $\mathbf{1}\{x'\bar{\beta} + \bar{\lambda}_\tau(v) > 0\} = 0$ for $\bar{\beta} \neq \beta_0$ or $\bar{\lambda}_\tau \neq \lambda_\tau$. We can do so because by the assumption stated in the theorem there is a subset of the support of (X, Z) that occurs with nonzero probability and on which the condition $x'\bar{\beta} + \bar{\lambda}_\tau(v) \leq 0 < x'\beta_0 + \lambda_\tau(v)$ holds. Since the condition $\mathbf{1}\{x'\beta_0 + \lambda_\tau(v) > 0\} = 1$ results in some negative value of $Q(\beta_0, \lambda_\tau)$ as previously shown in (1.27) and $Q(\bar{\beta}, \bar{\lambda}_\tau) = 0$ under the condition of $\mathbf{1}\{x\bar{\beta} + \bar{\lambda}(v) > 0\} = 0$ by using the same argument, it then follows that $Q(\bar{\beta}, \bar{\lambda}_\tau) = 0 > Q(\beta_0, \lambda_\tau)$. Similarly, we also have $Q(\bar{\beta}, \bar{\lambda}_\tau) > Q(\beta_0, \lambda_\tau) = 0$ for the case of $\mathbf{1}\{x'\beta_0 + \lambda_\tau(v) > 0\} = 0$ and $\mathbf{1}\{x'\bar{\beta} + \bar{\lambda}_\tau(v) > 0\} = 1$. Putting things together, we conclude that β_0 and λ_τ are the unique joint minimizers of $Q(\bar{\beta}, \bar{\lambda}_\tau)$. The desired result then follows from the fact that the choices of x_1 and z and hence v are arbitrary and that measurable separability between X and V imposed in Assumption 1, ensuring that $X'\beta$ and $V = F_{X_1|Z}(X_1, Z)$ can vary sufficiently independently.

Proof of Theorem 1.2(a)

To prove consistency, define

$$\hat{Q}_{n\kappa}(\theta) = \frac{1}{n}\sum_{i=1}^{n} t(\hat{W}_i)[Y_i - (1-\tau)]K\left(\frac{P_\kappa(\hat{W}_i)'\theta}{h}\right) \qquad \text{and } \hat{\theta}_{n\kappa} \equiv \arg\max_{\theta\in\Theta_\kappa} \hat{Q}_{n\kappa}(\theta),$$

$$\tilde{Q}_{n\kappa}(\theta) = \frac{1}{n}\sum_{i=1}^{n} t(W_i)[Y_i - (1-\tau)]K\left(\frac{P_\kappa(W_i)'\theta}{h}\right) \qquad \text{and } \tilde{\theta}_{n\kappa} \equiv \arg\max_{\theta\in\Theta_\kappa} \tilde{Q}_{n\kappa}(\theta),$$

$$Q_{n\kappa 0}(\theta) = \frac{1}{n}\sum_{i=1}^{n} t(W_i)[Y_i - (1-\tau)]K\left(\frac{P_\kappa(W_i)'\theta + b_{\kappa 0}(V_i)}{h}\right) \quad \text{and } \tilde{\theta}_{n\kappa 0} \equiv \arg\max_{\theta\in\Theta_\kappa} Q_{n\kappa 0}(\theta),$$

$$Q^*_{n\kappa 0}(\theta) = \frac{1}{n}\sum_{i=1}^{n} t(W_i)[Y_i - (1-\tau)]\mathbf{1}\{P_\kappa(W_i)'\theta + b_{\kappa 0}(V_i)\} \qquad \text{and } \theta^*_{n\kappa 0} \equiv \arg\max_{\theta\in\Theta_\kappa} Q^*_{n\kappa 0}(\theta),$$

$$Q^*_{\kappa 0}(\theta) = \mathbb{E}\big[t(W)[Y - (1-\tau)]\mathbf{1}\{P_\kappa(W)'\theta + b_{\kappa 0}(V) \geq 0\}\big] \qquad \text{and } \theta^*_{\kappa} \equiv \arg\max_{\theta\in\Theta_\kappa} Q^*_{\kappa 0}(\theta).$$

Write

$$\begin{aligned}\left|\hat{Q}_{n\kappa}(\theta) - Q^*_{\kappa 0}(\theta)\right| = {} & \left|\hat{Q}_{n\kappa}(\theta) - \tilde{Q}_{n\kappa}(\theta)\right| + \left|\tilde{Q}_{n\kappa}(\theta) - Q_{n\kappa 0}(\theta)\right| \\ & + \left|Q_{n\kappa 0}(\theta) - Q^*_{n\kappa 0}(\theta)\right| + \left|Q^*_{n\kappa 0}(\theta) - Q^*_{\kappa 0}(\theta)\right|. \qquad (1.28)\end{aligned}$$

Uniform convergence of each term on the right hand side of (1.28) is shown by the following lemmas. Then by the usual consistency argument in Theorem 2.1 in Newey and McFadden (1994), the consistency result in part (a) follows since $\hat{\theta}_{n\kappa}$ and θ^*_κ uniquely maximize $\hat{Q}_{n\kappa}(\theta)$ and $Q^*_{\kappa 0}(\theta)$, respectively, and uniqueness of the series representation of the function $\lambda_\tau(V)$ implying that $\|\theta^*_\kappa - \theta_{\kappa 0}\| \to 0$ as $n \to 0$. This completes the proof of the theorem.

Lemma 1.9.1. *For any $\eta > 0$, $\hat{Q}_{n\kappa}(\hat{\theta}_{n\kappa}) - \tilde{Q}_{n\kappa}(\hat{\theta}_{n\kappa}) < \eta$ for all sufficiently large n.*

Proof. First it can be shown that the feasible objective function $\hat{Q}_{n\kappa}(\theta)$ and the infeasible one $\tilde{Q}_{n\kappa}(\theta)$ get close to each other uniformly in θ as $n \to \infty$, i.e.,

$$
\begin{aligned}
&\sup_{\theta\in\Theta_\kappa} \left|\tilde{Q}_{n\kappa}(\theta) - \hat{Q}_{n\kappa}(\theta)\right| \\
&= \sup_{\theta\in\Theta_\kappa} \left| \frac{1}{n}\sum_{i=1}^{n} [Y_i - (1-\tau)] K\left(\frac{P_\kappa(\hat{W}_i)'\theta}{h}\right) - \frac{1}{n}\sum_{i=1}^{n} [Y_i - (1-\tau)] K\left(\frac{P_\kappa(W_i)'\theta}{h}\right) \right| \\
&\le C \sup_{\theta\in\Theta_\kappa} \left| K\left(\frac{P_\kappa(\hat{W}_i)'\theta}{h}\right) - K\left(\frac{P_\kappa(W_i)'\theta}{h}\right) \right| \\
&\le C \sup_{\theta\in\Theta_\kappa} \left(\frac{1}{h} |K^{(1)}(d_i^*)| \times \max_{1\le i\le n} \left| \left(P_\kappa(\hat{W}_i) - P_\kappa(W_i)\right)'\theta \right| \right) \\
&\le C \sup_{\theta\in\Theta_\kappa} \left(\frac{1}{h} |K^{(1)}(d_i^*)| \times \max_{1\le i\le n} \left| \frac{\mathrm{d}\lambda_\tau(\tilde{V}_i)}{\mathrm{d}v} (\hat{V}_i - V_i) \right| \right) \\
&\le C \sup_{\theta\in\Theta_\kappa} \left(\frac{1}{h} |K^{(1)}(d_i^*)| \times \left(\max_{1\le i\le n} \left| \frac{\mathrm{d}\lambda_\tau(\tilde{V}_i)}{\mathrm{d}v} (\hat{V}_i - V_i) \right| \right) \right) \\
&= O_p\left(h^{-1}\zeta_1(\kappa)\Delta_v\right) \\
&= o_p(1)
\end{aligned}
$$

by Lemma A.1 where $\tilde{V}_i$ is some value between V_i and $\hat{V}_i$. □

Lemma 1.9.2. $|\tilde{Q}_{n\kappa}(\theta) - Q_{n\kappa 0}(\theta)| \to 0$ *almost surely uniformly over* $\theta \in \{-1, 1\} \times \mathbb{R}^{d_x+\kappa-1}$.

Proof. This lemma follows since the approximation error $b_{\kappa 0}(v) \to 0$ for almost every v. □

Lemma 1.9.3. $|Q_{\kappa 0}(\theta) - Q_{\kappa 0}(\theta)| \to 0$ *almost surely uniformly over* $\theta \in \{-1, 1\} \times \mathbb{R}^{d_x+\kappa-1}$.

Proof. Observe that $K\left(\frac{P_\kappa(W)'\theta + b_{\kappa 0}(V)}{h}\right)$ converges to $\mathbf{1}\{P_\kappa(W)'\theta + b_{\kappa 0}(V) \ge 0\}$ as $h \to 0$ if $P_\kappa(W)'\theta + b_{\kappa 0}(V) \ne 0$. Thus, $Q_{\kappa 0}(\theta_\kappa)$ can be arbitrarily close to $Q^*_{\kappa 0}(\theta_\kappa)$ for all $\theta \in \Theta_\kappa$ as $n \to \infty$. Formally, for any $\eta > 0$, write

$$
\begin{aligned}
|Q_{\kappa 0}(\theta_\kappa) - Q^*_{\kappa 0}(\theta_\kappa)| \le\ & n^{-1} \sum_{i=1}^{n} \left| \mathbf{1}\{X_i'\beta_0 + \lambda_\tau(V_i) \ge 0\} - K\left((X_i'\beta_0 + \lambda_\tau(V_i))/h\right) \right| \mathbf{1}\{|X_i'\beta_0 + \lambda_\tau(V_i)| \ge \eta\} \\
&+ n^{-1} \sum_{i=1}^{n} \left| \mathbf{1}\{X_i'\beta_0 + \lambda_\tau(V_i) \ge 0\} - K\left((X_i'\beta_0 + \lambda_\tau(V_i))/h\right) \right| \mathbf{1}\{|X_i'\beta_0 + \lambda_\tau(V_i)| < \eta\}.
\end{aligned}
$$

By applying the same argument as that in the proofs of Lemma 4 in Horowitz (1992), one can show that both terms on the right hand side of the equation above converge to zero uniformly over $\theta \in \Theta^*_\kappa = \{1, -1\} \times B \times \Lambda_\kappa$ as $n \to \infty$. □

Lemma 1.9.4. *$Q^*_{n\kappa 0}(\bar\theta) \to Q^*_{\kappa 0}(\bar\theta)$ almost surely uniformly over $\bar\theta \in \mathbb{R}^{d_x+\kappa}$.*

Proof. This proof is given as in the proof in Lemma 4 of Manski (1988). □

Proof of Theorem 1.2(b) and (c)

Let $K^{(2)}_{h\kappa 0i}$ denote $K^{(2)}_h(P_\kappa(W_i)'\theta_{\kappa 0})$. Recall that

$$H_\kappa = \mathbb{E}[t(W)P_\kappa(W)P_\kappa(W)'F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde W) f_{m_\tau}(0 \mid \tilde W)]$$

and define

$$\hat H_{n\kappa} = n^{-1}\sum_{i=1}^n \hat t_i[Y_i - (1-\tau)]K^{(2)}_{h\kappa 0i}\hat P_{\kappa i}\hat P'_{\kappa i},$$

$$\tilde H_{n\kappa} = n^{-1}\sum_{i=1}^n t_i[Y_i - (1-\tau)]K^{(2)}_{h\kappa 0i} P_{\kappa i} P'_{\kappa i},$$

and

$$\hat G_{n\kappa}(\theta) = n^{-1}\hat H^{-1}_{n\kappa}\sum_{i=1}^n \hat t_i[Y_i - (1-\tau)]K^{(1)}_h\big(P_\kappa(\hat W_i)'(\theta-\theta_{\kappa 0}) + P_\kappa(\hat W_i)'\theta_{\kappa 0}\big)\hat P_{\kappa i},$$

$$\tilde G_{n\kappa}(\theta) = n^{-1}\tilde H^{-1}_{n\kappa}\sum_{i=1}^n t_i[Y_i - (1-\tau)]K^{(1)}_h\big(P_\kappa(W_i)'(\theta-\theta_{\kappa 0}) + P_\kappa(W_i)'\theta_{\kappa 0}\big) P_{\kappa i}.$$

The following lemmas are useful to prove Theorem 1.2.

Lemma 1.9.5. *As $n \to \infty$,*

(a) $\max_{1\le i\le n} t_i\|\hat P_{\kappa i} - P_{\kappa i}\|^2 = O_p(\zeta_1(\kappa)^2\Delta_v^2)$.
(b) $n^{-1}\sum_{i=1}^n t_i\|\hat P_{\kappa i} - P_{\kappa i}\|^2 = O_p(\zeta_1(\kappa)^2\Delta_v^2)$.
(c) $n^{-1}\sum_{i=1}^n \|P_{\kappa i}\|^2 = O_p(\kappa)$.
(d) $n^{-1}\sum_{i=1}^n t_i|\hat t_i - t_i| = O_p(\Delta_v)$.
(e) $\|\tilde H_{n\kappa} - H_\kappa\| = o_p(1)$.
(f) $\|\hat H_{n\kappa} - \tilde H_{n\kappa}\| = O_p(\zeta_1(\kappa)^2\Delta_v^2 + \kappa^{1/2}\zeta_1(\kappa)\Delta_v + \zeta_0(\kappa)^2\Delta_v) = o_p(1)$.

Proof. For part (a), a mean value expansion gives $p_\kappa(\hat V_i) - p_\kappa(V_i) = \frac{\mathrm{d}p_\kappa(\tilde V_i)}{\mathrm{d}v}(\hat V_i - V_i)$, where $\tilde V_i$ lies in between $\hat V_i$ and V_i. Hence

$$\max_{1\le i\le n} t_i\|\hat P_{\kappa i} - P_{\kappa i}\|^2 = \max_{1\le i\le n} t_i\sum_{k=1}^\kappa [p_k(\hat V_i) - p_k(V_i)]^2 = \max_{1\le i\le n} t_i\sum_{k=1}^\kappa \left[\frac{\mathrm{d}p_k(\tilde V_i)}{\mathrm{d}v}(\hat V_i - V_i)\right]^2$$
$$= \max_{1\le i\le n} t_i(\hat V_i - V_i)^2\sum_{k=1}^\kappa \left[\frac{\mathrm{d}p_k(\tilde V_i)}{\mathrm{d}v}\right]^2 = O_p(\zeta_1(\kappa)^2\Delta_v^2),$$

where the last equality follows from the fact that $\max_{1\leq i\leq n} t_i|\hat{V}_i - V_i| = O_p(\Delta_v)$ and $\|\mathrm{d}p_\kappa(\tilde{V}_i)/\mathrm{d}v\| = O_p(\zeta_1(\kappa))$. Similarly,

$$n^{-1}\sum_{i=1}^{n} t_i\|\hat{P}_{\kappa i} - P_{\kappa i}\|^2 \leq C\zeta_1(\kappa)^2 n^{-1}\sum_{i=1}^{n} t_i|\hat{V}_i - V_i|^2 = O_p(\zeta_1(\kappa)^2\Delta_v^2),$$

where the first inequality follows from the fact that by the Cauchy-Schwarz inequality, $\|\hat{P}_{\kappa i} - P_{\kappa i}\| \leq C\zeta_1(\kappa)|\hat{V}_i - V_i|$. This completes the proof of part (b).

For part (c), note that

$$n^{-1}\sum_{i=1}^{n}\|P_{\kappa i}\|^2 = n^{-1}\sum_{i=1}^{n}\mathrm{trace}(\hat{P}_{\kappa i}\hat{P}'_{\kappa i}) \leq C\cdot\mathrm{trace}(\hat{H}_{n\kappa}) = O_p(\kappa).$$

For part (d), using equation (14) of the Appendix in Lee (2007) by slightly modifying arguments used in Lemma A3 of Newey, Powell, and Vella (1999) yields

$$n^{-1}\sum_{i=1}^{n} t_i|\hat{t}_i - t_i| = O_p\Big(\max_{1\leq i\leq n} t_i|\hat{V}_i - V_i|\Big) = O_p(\Delta_v).$$

For part (e), define $\ddot{\theta}_{n\kappa} = \frac{\hat{\theta}_{n\kappa}-\theta_{\kappa 0}}{h}$. Let $\{a_n\}$ be a sequence such that $a_n \to \infty$ and $a_n\ddot{\theta}_{n\kappa} \to 0$ as $n \to \infty$. Define $W_{n\kappa} = \{\tilde{w} : \|P_\kappa(\tilde{w})\| \leq a_n\}$ and $\{\theta_{n\kappa}\} = \{\gamma_{n1}, \hat{\theta}_{n\kappa}\}$. For any $\varepsilon > 0$, $\lim_{n\to\infty}\Pr[|\tilde{H}_{n\kappa}(\theta_{n\kappa}; h) - H_\kappa| > \varepsilon] = \lim_{n\to\infty}\Pr[|\tilde{H}_{n\kappa}(\theta_{n\kappa}; h) - H_\kappa| > \varepsilon \mid W_{n\kappa}]$. Then using the argument that is analogous to that of Theorem 3 (c) (or Lemma 9) of Horowitz (1992), one can show that $\mathbb{E}[H_{n\kappa}(\theta_{n\kappa}; h) \mid W_{n\kappa}] \to H_\kappa$ and $\mathrm{Var}[H_{n\kappa}(\theta_{n\kappa}; h) \mid W_{n\kappa}] \to 0$ under Assumption 1.7. Next consider part (f). As in (A.5) of Newey, Powell, and Vella (1999), part (f)

follows that

$$\begin{aligned}
\|\hat{H}_{n\kappa} - \tilde{H}_{n\kappa}\| &= \left\| n^{-1} \sum_{i=1}^{n} \big(\hat{t}_i F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} \hat{P}_{\kappa i} \hat{P}'_{\kappa i} - t_i F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} P_{\kappa i} P'_{\kappa i} \big) \right\| \\
&= \Big\| n^{-1} \sum_{i=1}^{n} \big[t_i \hat{t}_i F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} \big(\hat{P}_{\kappa i} \hat{P}'_{\kappa i} - P_{\kappa i} P'_{\kappa i} \big) \\
&\quad + \big(\hat{t}_i - t_i \hat{t}_i \big) F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} \hat{P}_{\kappa i} \hat{P}'_{\kappa i} + (t_i \hat{t}_i - t_i) F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} P_{\kappa i} P'_{\kappa i} \big] \Big\| \\
&\leq C n^{-1} \sum_{i=1}^{n} t_i \hat{t}_i \big(\|\hat{P}_{\kappa i} - P_{\kappa i}\|^2 + 2 \|\hat{P}_{\kappa i} - P_{\kappa i}\| \|P_{\kappa i}\| \big) + C \zeta_0(\kappa)^2 n^{-1} \sum_{i=1}^{n} t_i |\hat{t}_i - t_i| \\
&\leq C n^{-1} \sum_{i=1}^{n} t_i \|\hat{P}_{\kappa i} - P_{\kappa i}\|^2 + C \Big(n^{-1} \sum_{i=1}^{n} t_i \|P_{\kappa i}\|^2 \Big)^{1/2} \Big(n^{-1} \sum_{i=1}^{n} t_i \hat{t}_i \|\hat{P}_{\kappa i} - P_{\kappa i}\|^2 \Big)^{1/2} \\
&\quad + C \zeta_0(\kappa)^2 n^{-1} \sum_{i=1}^{n} t_i |\hat{t}_i - t_i| \\
&= O_p\big(\zeta_1(\kappa)^2 \Delta_v^2 \big) + O_p\big(\kappa^{1/2} \zeta_1(\kappa) \Delta_v \big) + O_p\big(\zeta_0(\kappa)^2 \Delta_v \big),
\end{aligned}$$

where the last equality follows from Lemma 1.9.5 (b), (c), and (d). □

Let $\mathbf{1}_n$ be the indicator function such that $\mathbf{1}\{\lambda_{\min}(\hat{H}_{n\kappa}) \geq \lambda_{\min}(H_\kappa)/2 \text{ and } \lambda_{\min}(\tilde{H}_{n\kappa}) \geq \lambda_{\min}(H_\kappa)/2\}$.

Lemma 1.9.6. *As $n \to \infty$,*

$$\mathbf{1}_n n^{-1} \sum_{i=1}^{n} \|\hat{t}_i \hat{H}^{-1}_{n\kappa} \hat{P}_{\kappa i} - t_i \tilde{H}^{-1}_{n\kappa} P_{\kappa i}\|^2 = O_p(\zeta_0(\kappa)^2 \Delta_v + \zeta_1(\kappa)^2 \Delta_v^2 + \kappa \zeta_1(\kappa)^4 \Delta_v^4) = o_p(1).$$

Proof. The argument used to prove this lemma is the same as the first equation on page 1145 of Lee (2007) with the adjustment for the nonparametrically generated regressor. That is, the lemma follows from Lemma 1.9.5 (b)-(d) and (f). □

Lemma 1.9.7. *As $n \to \infty$,*

$$\max_{1 \leq i \leq n} \quad t_i \Big| (\hat{P}_{\kappa i} - P_{\kappa i})' \theta_{\kappa 0} - \frac{d\lambda_\tau(V_i)}{dv} (\hat{V}_i - V_i) \Big| = O_p(\kappa^{-r+1} \Delta_v) + o_p(\Delta_v).$$

Proof. Consider

$$
\begin{aligned}
& t_i\left[(\hat{P}_{\kappa i}-P_{\kappa i})'\theta_{\kappa 0}-\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i-V_i)\right] \\
&= t_i\sum_{k=1}^{\kappa}\left[\Big(p_k(\hat{W}_i)-p_k(W_i)\Big)\theta_{\kappa 0}^k-\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}(\hat{V}_i-V_i)\right] \\
&= t_i\sum_{k=1}^{\kappa}\left[\frac{\mathrm{d}p_k(\tilde{W}_i)}{\mathrm{d}v}\theta_{\kappa 0}^k(\hat{V}_i-V_i)-\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}(\hat{V}_i-V_i)\right] \\
&= t_i\left[\sum_{k=1}^{\kappa}\frac{\mathrm{d}p_k(\tilde{V}_i)}{\mathrm{d}v}\alpha_{\kappa 0}^k-\frac{\mathrm{d}\lambda_\tau(\tilde{V}_i)}{\mathrm{d}v}\right](\hat{V}_i-V_i)+t_i\left[\frac{\mathrm{d}\lambda_\tau(\tilde{V}_i)}{\mathrm{d}v}-\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\right](\hat{V}_i-V_i) \\
&= O_p(\kappa^{-r+1}\Delta_v)+\frac{\mathrm{d}^2(\lambda_\tau(\tilde{\tilde{V}}_i))}{\mathrm{d}v^2}(\tilde{V}_i-V_i)(\hat{V}_i-V_i) \\
&\leq O_p(\kappa^{-r+1}\Delta_v+\Delta_v^2),
\end{aligned}
$$

where the second and third equalities follow from the Taylor expansion approximation and the fact that $\sup_v|\mathrm{d}\lambda_\tau(v)/\partial v-(\partial P_\kappa(w)/\partial v)'\theta_{\kappa 0}|=O(\kappa^{-1+r})$, respectively and the last inequality follows from a Taylor expansion with $\mathrm{d}\lambda_\tau(v)/\mathrm{d}v$ being continuously differentiable and $\tilde{\tilde{V}}_i$ between $\hat{V}_i$ and V_i. □

Denote $K_{h\kappa 0i}^{(1)}\equiv K_h^{(1)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big)$ and $K_{h\kappa 0i}^{(2)}\equiv K_h^{(2)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big)$.

Lemma 1.9.8. *As $n\to\infty$,*

(a) $\mathbf{1}_n\left\|n^{-1}AH_{n\kappa}^{-1}\sum_{i=1}^n t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}\frac{d\lambda(V_i)}{dv}(\hat{V}_i-V_i)P_{\kappa i}\right\|=O_p\Big((nh^3)^{-1/2}\Delta_v\Big)$.

(b) $\mathbf{1}_n\left\|n^{-1}H_{n\kappa}^{-1}\sum_{i=1}^n t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}P_{\kappa i}b_{\kappa 0}(V_i)\right\|=O_p(h^{-1/2}\kappa^{-r})$.

Proof. For part (a), consider

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_n\left\|n^{-1}AH_{n\kappa}^{-1}\sum_{i=1}^{n}t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}(\hat{V}_i-V_i)P_{\kappa i}\right\|^2\middle|X_1,\ldots,X_n,Z_1,\ldots,Z_n\right]\\
&\quad\leq \mathbf{1}_n n^{-2}\sum_{i=1}^{n}\left\{\mathbb{E}\big[\big\{[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}\big\}^2|X_i,Z_i\big]\left(\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\right)^2(\hat{V}_i-V_i)^2P_\kappa(W_i)'H_{n\kappa}^{-1}A'AH_{n\kappa}^{-1}P_\kappa(W_i)\right\}\\
&\quad\leq C\mathbf{1}_n n^{-2}O(h^{-3})\max_{1\leq i\leq n}(\hat{V}_i-V_i)^2\sum_{i=1}^{n}\mathrm{trace}[t_iP_\kappa(W_i)'H_{n\kappa}^{-1}A'AH_{n\kappa}^{-1}P_\kappa(W_i)]\\
&\quad\leq C\mathbf{1}_n n^{-2}h^{-3}\Delta_v^2\sum_{i=1}^{n}\left(\frac{t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}}{\min_i\big\{t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}\big\}}\right)\mathrm{trace}[t_iP_\kappa(W_i)'H_{n\kappa}^{-1}A'AH_{n\kappa}^{-1}P_\kappa(W_i)]\\
&\quad\leq C\mathbf{1}_n n^{-1}h^{-3}\Delta_v^2\mathrm{trace}\left[AH_{n\kappa}^{-1}\left\{n^{-1}\sum_{i=1}^{n}t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}P_\kappa(W_i)P_\kappa(W_i)'\right\}H_{n\kappa}^{-1}A'\right]\\
&\quad= C\mathbf{1}_n n^{-1}h^{-3}\Delta_v^2\mathrm{trace}\big[AH_{n\kappa}^{-1}A'\big]\\
&\quad\leq C(nh^3)^{-1}\Delta_v^2
\end{aligned}$$

for some constant $C<\infty$. The desired result follows from Markov's inequality.

For part (b), define

$$B_n=n^{-1}H_{n\kappa}^{-1}\sum_{i=1}^{n}t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}P_{\kappa i}b_{\kappa 0}(V_i).$$

For each $i=1,\cdots,n$, let D_b be the $n\times 1$ vector with ith component $\big(t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}\big)^{1/2}b_{\kappa 0}(V_i)$ and $\mathbf{P}_\kappa$ a $n\times\kappa$ matrix with the ith row $\Big(t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}\Big)^{1/2}P_{\kappa i}'$. Then $B_n=n^{-1}H_{n\kappa}^{-1}\mathbf{P}_\kappa'D_b$, and $\mathbf{1}_n\|B_n\|^2=\mathbf{1}_n n^{-2}D_b'\mathbf{P}_\kappa H_{n\kappa}^{-2}\mathbf{P}_\kappa'D_b$. Also note that $H_{n\kappa}=n^{-1}\sum_{i=1}^{n}t_i[Y_i-(1-\tau)]K_{h\kappa 0i}^{(2)}P_{\kappa i}P_{\kappa i}'=$

$\mathbf{P}_\kappa'\mathbf{P}_\kappa/n$. By exactly the same arguments used to prove part (a), we have,

$$\begin{aligned}
&\mathbf{1}_n \left\| n^{-1} H_{n\kappa}^{-1} \sum_{i=1}^n t_i [Y_i - (1-\tau)] K_{h\kappa 0i}^{(2)} P_{\kappa i} \right\|^2 \\
&= \mathbf{1}_n \| n^{-1} H_{n\kappa}^{-1} \mathbf{P}_\kappa' D_v \|^2 \\
&= \mathbf{1}_n n^{-2} D_v' \mathbf{P}_\kappa H_{n\kappa}^{-2} \mathbf{P}_\kappa' D_v \\
&\leq \mathbf{1}_n n^{-2} \lambda_{\max}(H_{n\kappa}^{-1}) D_v' \mathbf{P}_\kappa H_{n\kappa}^{-1} \mathbf{P}_\kappa' D_v \\
&= \mathbf{1}_n n^{-2} \lambda_{\max}(H_{n\kappa}^{-1}) D_v' \mathbf{P}_\kappa (\mathbf{P}_\kappa'\mathbf{P}_\kappa/n)^{-1} \mathbf{P}_\kappa' D_v \\
&\leq \mathbf{1}_n n^{-1} \lambda_{\max}(H_{n\kappa}^{-1}) \lambda_{\max}(\mathbf{P}_\kappa (\mathbf{P}_\kappa'\mathbf{P}_\kappa)^{-1} \mathbf{P}_\kappa') D_v' D_v \\
&\leq O_p(h^{-1}) \max_{1\leq i\leq n} b_{\kappa 0}(V_i)^2 = O_p(\kappa^{-2r}/h),
\end{aligned}$$

where the first two inequalities follow from the fact that the matrices $H_{n\kappa}^{-1}$ and $\mathbf{P}_\kappa(\mathbf{P}_\kappa'\mathbf{P}_\kappa)^{-1}\mathbf{P}_\kappa'$ are symmetric. This completes the proof of the lemma. □

Lemma 1.9.9. *As $n \to \infty$*

$$\begin{aligned}
\mathbf{1}_n \hat{G}_{n\kappa}(\theta) &= \mathbf{1}_n H_{n\kappa}^{-1} n^{-1} \sum_{i=1}^n t_i [Y_i - (1-\tau)] K_h^{(1)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big) P_{\kappa i}' + \mathbf{1}_n(\theta - \theta_{\kappa 0}) \\
&\quad + \mathbf{1}_n n^{-1} H_{n\kappa}^{-1} \sum_{i=1}^n t_i [Y_i - (1-\tau)] K_h^{(2)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big) P_{\kappa i} b_{\kappa 0}(V_i) \\
&\quad + \mathbf{1}_n n^{-1} H_{n\kappa}^{-1} \sum_{i=1}^n t_i [Y_i - (1-\tau)] K_h^{(2)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big) \frac{d\lambda_\tau(V_i)}{dv} (\hat{V}_i - V_i) P_{\kappa i} + R_{n\kappa}^*(\theta),
\end{aligned}$$

where the remainder term $R_{n\kappa}^*(\theta)$ satisfies

$$\|R_{n\kappa}^*(\theta)\| = O_p\Big(\kappa^{1/2}\Delta_v + \zeta_0(\kappa) h^{-1} \|\theta - \theta_{\kappa 0}\|^2 + \kappa^{1/2} h^{-1} \Delta_v^2 + \kappa^{-2r+1/2}\Big) + o_p(\Delta_v).$$

Proof. Define

$$\begin{aligned}
\mathbf{1}_n \hat{G}_{n\kappa 1}^*(\theta) &= \mathbf{1}_n(\theta - \theta_{\kappa 0}) + \mathbf{1}_n n^{-1} \tilde{H}_{n\kappa}^{-1} \sum_{i=1}^n t_i [Y_i - (1-\tau)] K_{h\kappa 0i}^{(2)} \frac{d\lambda_\tau(V_i)}{dv} (\hat{V}_i - V_i) P_{\kappa i} \\
&\quad + \mathbf{1}_n n^{-1} \tilde{H}_{n\kappa}^{-1} \sum_{i=1}^n t_i [Y_i - (1-\tau)] K_{h\kappa 0i}^{(2)} b_{\kappa 0i} P_{\kappa i}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{1}_n \bar{G}_{n\kappa}^*(\theta) &= \mathbf{1}_n(\theta - \theta_{\kappa 0}) + \mathbf{1}_n n^{-1} \hat{H}_{n\kappa}^{-1} \sum_{i=1}^n \hat{t}_i F_{\varepsilon_{\tau i}}^{(1)} f_{m_{\tau i}} (\hat{P}_{\kappa i} - P_{\kappa i})' \theta_{\kappa 0} \\
&\quad + n^{-1} \hat{H}_{n\kappa}^{-1} \sum_{i=1}^n \hat{t}_i F_{\varepsilon_{\tau i}}^{(1)} f_{m_{\tau i}} b_{\kappa 0i} \hat{P}_{\kappa i}.
\end{aligned}$$

Then the term $\mathbf{1}_n\hat{G}^*_{n\kappa}(\alpha)$ can be expressed as

$$\begin{aligned}\mathbf{1}_n\hat{G}^*_{n\kappa}(\theta) &= \mathbf{1}_n\hat{G}^*_{n\kappa 1}(\theta) + \big[\mathbf{1}_n\bar{G}^*_{n\kappa}(\theta) - \mathbf{1}_n\hat{G}^*_{n\kappa 1}(\theta)\big] + \big[\mathbf{1}_n\hat{G}^*_{n\kappa}(\theta) - \mathbf{1}_n\bar{G}^*_{n\kappa}(\theta)\big] \\ &\equiv \mathbf{1}_n\hat{G}^*_{n\kappa 1}(\theta) + \underbrace{R^*_{n\kappa 1}(\theta) + R^*_{n\kappa 2}(\theta)}_{\equiv R^*_{n\kappa}(\theta)}.\end{aligned}$$

Notice that $X_i'\beta_0 + \lambda_\tau(V_i) = P_{\kappa i}'\theta_{\kappa 0} + b_{\kappa 0 i}$, so that

$$\begin{aligned}\|R^*_{n\kappa 1}(\theta)\| \le{}& \Bigg\|\mathbf{1}_n n^{-1}\hat{H}^{-1}_{n\kappa}\sum_{i=1}^n \hat{t}_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}(\hat{P}_{\kappa i} - P_{\kappa i})'\theta_{\kappa 0}\hat{P}_{\kappa i} \\ &\quad -\mathbf{1}_n n^{-1}\hat{H}^{-1}_{n\kappa}\sum_{i=1}^n \hat{t}_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)\hat{P}_{\kappa i}\Bigg\| \\ &+ \Bigg\|\mathbf{1}_n n^{-1}\hat{H}^{-1}_{n\kappa}\sum_{i=1}^n \hat{t}_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)\hat{P}_{\kappa i} \\ &\quad -\mathbf{1}_n n^{-1}\tilde{H}^{-1}_{n\kappa}\sum_{i=1}^n t_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)P_{\kappa i}\Bigg\| \\ &+ \Bigg\|\mathbf{1}_n n^{-1}\hat{H}^{-1}_{n\kappa}\sum_{i=1}^n \hat{t}_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i} b_{\kappa 0 i}\hat{P}_{\kappa i} - \mathbf{1}_n n^{-1}\tilde{H}^{-1}_{n\kappa}\sum_{i=1}^n t_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i} b_{\kappa 0 i}P_{\kappa i}\Bigg\| \\ \equiv{}& G_{11} + G_{12} + G_{13}.\end{aligned}$$

Consider the first term G_{11}:

$$\begin{aligned}G_{11} &= \Bigg\|\mathbf{1}_n n^{-1}\hat{H}^{-1}_{n\kappa}\sum_{i=1}^n \hat{t}_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}\big[(\hat{P}_{\kappa i} - P_{\kappa i})'\theta_{\kappa 0} - \frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)\big]\hat{P}_{\kappa i}\Bigg\| \\ &\le \mathbf{1}_n n^{-1}\sum_{i=1}^n \|\hat{H}^{-1}_{n\kappa}\hat{P}_{\kappa i}\|\hat{t}_i F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}\left|(\hat{P}_{\kappa i} - P_{\kappa i})'\theta_{\kappa 0} - \frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)\right| \\ &\le \mathbf{1}_n\Bigg(n^{-1}\sum_{i=1}^n \|\hat{H}^{-1}_{n\kappa}\hat{P}_{\kappa i}\|^2\Bigg)^{1/2}\Bigg(n^{-1}\sum_{i=1}^n \hat{t}_i\big(F^{(1)}_{\varepsilon_\tau i} f_{m_\tau i}\big)^2\left|(\hat{P}_{\kappa i} - P_{\kappa i})'\theta_{\kappa 0} - \frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)\right|^2\Bigg)^{1/2} \\ &= O_p\big(h^{-1}\kappa^{1/2}\Delta_v\big) = o_p(1) \qquad (1.29)\end{aligned}$$

as $n \longrightarrow \infty$, where the second inequality follows from the Cauchy-Schwarz inequality and the last equality is by Lemma 1.9.7 and the fact that $\mathbf{1}_n n^{-1}\sum_{i=1}^n \|\hat{H}^{-1}_{n\kappa}\hat{P}_{\kappa i}\|^2 = O_p(\kappa)$.

Similarly, for the term G_{12}:

$$\begin{aligned} G_{12} &\le \mathbf{1}_n n^{-1} \sum_{i=1}^{n} F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} \|\hat{t}_i \hat{H}_{n\kappa} \hat{P}_{\kappa i} - t_i \tilde{H}^{-1}_{n\kappa} P_{\kappa i}\| \left| \frac{d\lambda_\tau(V_i)}{dv} (\hat{V}_i - V_i) \right| \\ &\le C\mathbf{1}_n \left(n^{-1} \sum_{i=1}^{n} \|\hat{t}_i \hat{H}_{n\kappa} \hat{P}_{\kappa i} - t_i \tilde{H}^{-1}_{n\kappa} P_{\kappa i}\|^2 \right)^{1/2} \left(n^{-1} \sum_{i=1}^{n} \left| \frac{d\lambda_\tau(V_i)}{dv} (\hat{V}_i - V_i) \right|^2 \right)^{1/2} \\ &\le o_p(1) O_p(\Delta_v) = o_p(\Delta_v) = o_p(1) \text{ as } n \longrightarrow \infty. \end{aligned} \tag{1.30}$$

Similar arguments give G_{13} satisfying

$$\begin{aligned} G_{13} &= \left\| \mathbf{1}_n n^{-1} \hat{H}^{-1}_{n\kappa} \sum_{i=1}^{n} \hat{t}_i F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} b_{\kappa 0}(V_i) \hat{P}_{\kappa i} - \mathbf{1}_n n^{-1} \tilde{H}^{-1}_{n\kappa} \sum_{i=1}^{n} t_i F^{(1)}_{\varepsilon_{\tau i}} f_{m_{\tau i}} b_{\kappa 0}(V_i) P_{\kappa i} \right\| \\ &\le o_p(1) O_p(\kappa^{-r}) = o_p(1). \end{aligned} \tag{1.31}$$

Summing up equations (1.29), (1.30), and (1.31) gives that as $n \longrightarrow \infty$

$$\|R^*_{n\kappa 2}(\theta)\| = O_p\big(\kappa^{1/2}\Delta_v\big) + o_p(\Delta_v) + o_p(1) = o_p(1).$$

Now consider the term $R^*_{n\kappa 2}(\theta)$. Using the first-order Taylor expansion yields

$$\begin{aligned} \|R^*_{n\kappa 2}(\theta)\| &\equiv \|\mathbf{1}_n \hat{G}^*_{n\kappa}(\theta) - \mathbf{1}_n \bar{G}^*_{n\kappa}(\theta)\| \\ &\le Ch^{-1}[n^{-1} \sum_{i=1}^{n} \hat{t}_i \|\hat{H}^{-1}_{n\kappa} \hat{P}_{\kappa i}\| (P'_{\kappa i}(\theta - \theta_{\kappa o}) + (\hat{P}_{\kappa i} - P_{\kappa i})'\theta_{\kappa 0} + b_{\kappa 0 i})^2] \\ &\le Ch^{-1} \max_{1\le i\le n} \|\hat{H}^{-1}_{n\kappa} \hat{P}_{\kappa i}\| (\theta - \theta_{\kappa 0})' \left\{ \sum_{i=1}^{n} \hat{t}_i \hat{P}_{\kappa i} \hat{P}'_{\kappa i} \right\} (\theta - \theta_{\kappa 0}) \\ &\quad + Ch^{-1} \left\{ n^{-1} \sum_{i=1}^{n} \hat{t}_i \|\hat{H}^{-1}_{n\kappa} \hat{P}_{\kappa i}\| \right\} \max_{1\le i\le n} [((\hat{P}_{\kappa i} - P_{\kappa i})'\theta_{\kappa 0})^2 + b^2_{\kappa 0 i}] \\ &\le Ch^{-1}\zeta_0(\kappa) \lambda_{\max}(\hat{H}_{n\kappa})(\theta - \theta_{n\kappa})'(\theta - \theta_{\kappa 0}) + Ch^{-1}\kappa^{1/2}(\Delta_v^2 + \kappa^{-2r}) \\ &= O_p\big(h^{-1}\zeta_0(\kappa)\|\theta - \theta_{\kappa 0}\|^2 + h^{-1}\kappa^{1/2}\Delta_v^2 + h^{-1}\kappa^{-2r+1/2}\big). \end{aligned}$$

The desired result then follows by combining the last two equations. □

Lemma 1.9.10. *As $n \to \infty$,*

(a) $\|\tilde{G}_{n\kappa}(\theta_{\kappa 0})\| = O_p\big((\kappa/(nh))^{1/2}\big)$.

(b) $\|A\tilde{G}_{n\kappa}(\theta_{\kappa 0})\| = O_p\big((nh)^{-1/2}\big)$.

Proof. By using the kernel smoothing arguments, it is not difficult to show that

$$\mathbb{E}\big[\big\{[Y_i-(1-\tau)]K_h^{(1)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big)\big\}^2|X_i,Z_i\big]=O(h^{-1}). \tag{1.32}$$

Now Consider

$$\begin{aligned}
&\mathbb{E}\big[\|A\tilde{G}_{n\kappa}(\theta_{\kappa 0})\|^2|X_1,\ldots,X_n,Z_1,\ldots,Z_n\big]\\
&\quad\le n^{-2}\sum_{i=1}^n\Big\{t_i\mathbb{E}[([Y_i-(1-\tau)]K_{h0i}^{(1)})^2|X_i,Z_i]P_{\kappa i}'H_{n\kappa}^{-1}A'AH_{n\kappa}^{-1}P_{\kappa i}\Big\}\\
&\quad\le O(h^{-1})n^{-2}\sum_{i=1}^n\mathrm{trace}(t_iP_{\kappa i}'H_{n\kappa}^{-1}A'AH_{n\kappa}^{-1}P_{\kappa i})\\
&\quad= O(h^{-1})n^{-2}\sum_{i=1}^n\mathrm{trace}(t_iAH_{n\kappa}^{-1}P_{\kappa i}P_{\kappa i}'H_{n\kappa}^{-1}A')\\
&\quad\le O(h^{-1})n^{-2}\sum_{i=1}^n\Big\{\min_i F_{\varepsilon_{\tau i}}^{(1)}f_{m_{\tau i}}\Big\}^{-1}F_{\varepsilon_{\tau i}}^{(1)}f_{m_{\tau i}}\mathrm{trace}(t_iAH_{n\kappa}^{-1}P_{\kappa i}P_{\kappa i}'H_{n\kappa}^{-1}A')\\
&\quad= O(h^{-1})n^{-1}\mathrm{trace}(AH_{n\kappa}^{-1}n^{-1}\sum_{i=1}^n t_iF_{\varepsilon_{\tau i}}^{(1)}f_{m_{\tau i}}P_{\kappa i}P_{\kappa i}'H_{n\kappa}^{-1}A')\\
&\quad= O\big((nh)^{-1}\big)\mathrm{trace}(AH_{n\kappa}^{-1}A')\\
&\quad= O\big((nh)^{-1}\big),
\end{aligned}$$

where the first inequality follows from the data being i.i.d. and the second inequality follows from (1.32). This proves part (b) by Markov's inequality. Part (a) then follows from replacing A with the identity matrix and applying Markov's inequality. □

Proof of Theorem 1.2(b) and (c): By Lemma 1.9.9 and the fact that $\hat{G}_{n\kappa}(\hat{\theta}_{n\kappa})=0$, we have

$$\begin{aligned}
\mathbf{1}_n(\hat{\theta}_{n\kappa}-\theta_{\kappa 0}) &= H_{n\kappa}^{-1}n^{-1}\sum_{i=1}^n t_i[Y_i-(1-\tau)]K_h^{(1)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big)P_{\kappa i}'\\
&\quad+\mathbf{1}_nn^{-1}H_{n\kappa}^{-1}\sum_{i=1}^n t_i[Y_i-(1-\tau)]K_h^{(2)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big)P_{\kappa i}b_{\kappa 0}(V_i)\\
&\quad+\mathbf{1}_nn^{-1}H_{n\kappa}^{-1}\sum_{i=1}^n t_i[Y_i-(1-\tau)]K_h^{(2)}\big(P_\kappa(W_i)'\theta_{\kappa 0}\big)\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i-V_i)P_{\kappa i}+R_{n\kappa}^*(\hat{\theta}_{n\kappa}).
\end{aligned} \tag{1.33}$$

To prove part (b), suppose that $\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\| \leq C\Big(\big(\kappa/(nh)\big)^{1/2} + h^{-1/2}\kappa^{-r} + \big(\kappa/(nh^3)\big)^{1/2}\Delta_v\Big)$ for any constant $C > 0$. Then applying Lemmas 1.9.9 and 1.9.10 to equation (1.33), we have

$$\begin{aligned}
\mathbf{1}_n\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\| &\leq \mathbf{1}_n\|\tilde{G}_{n\kappa}(\theta_{\kappa 0})\| + \mathbf{1}_n\left\|n^{-1}\tilde{H}_{n\kappa}^{-1}\sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_{h\kappa 0i}^{(2)}\frac{d\lambda_\tau(V_i)}{dv}(\hat{V}_i - V_i)P_{\kappa i}\right\| \\
&\quad + \mathbf{1}_n\|n^{-1}\tilde{H}_{n\kappa}^{-1}\sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_{h\kappa 0i}^{(2)}b_{\kappa 0i}P_{\kappa i}\| + \|R_{n\kappa}^{*}(\hat{\theta}_{n\kappa})\| \\
&\leq O_p\Big(\big(\kappa/(nh)\big)^{1/2}\Big) + O_p\Big(\kappa^{1/2}(nh^3)^{-1/2}\Delta_v\Big) + O_p\big(h^{-1/2}\kappa^{-r}\big) \\
&\quad + O_p\Big(\kappa^{1/2}\Delta_v + \zeta_0(\kappa)\|\hat{\theta}_{n\kappa} - \theta_{\kappa 0}\|^2 + \zeta_0(\kappa)\Delta_v^2 + \zeta_0(\kappa)\kappa^{-2r}\Big) + o_p(\Delta_v). \qquad (1.34) \\
& \qquad (1.35)
\end{aligned}$$

The desired result follows since the right-hand side of equation (1.34) is less than $C((\kappa/(nh))^{1/2} + h^{-1/2}\kappa^{-r} + \kappa^{1/2}(nh^3)^{1/2}\Delta_v)$ with probability approaching one and $\Pr(\mathbf{1}_n = 1) \longrightarrow 1$ as $n \longrightarrow \infty$.

For part (c), define

$$\bar{G}_{n\kappa}(\theta_{\kappa 0}) = n^{-1}\tilde{H}_{n\kappa}^{-1}\sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_h^{(1)}(m_{\tau i})P_{\kappa i}.$$

By using the arguments similar to those used in the proof of Lemma 1.9.10, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_n\left\|\bar{G}_{n\kappa}(\theta_{\kappa 0}) - n^{-1}\tilde{H}_{n\kappa}^{-1}\sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_h^{(1)}(P_\kappa(W_i)'\theta_{\kappa 0})P_{\kappa i}\right\|^2 \middle| X_1, \ldots, X_n, Z_1, \ldots, Z_n\right] \\
&\quad \leq C(\kappa/n)\sup_i |b_{\kappa 0i}|.
\end{aligned}$$

Hence,

$$\mathbf{1}_n\left\|\bar{G}_{n\kappa}(\theta_{\kappa 0}) - n^{-1}\tilde{H}_{n\kappa}^{-1}\sum_{i=1}^{n}[Y_i - (1-\tau)]K_h^{(1)}(P_\kappa(W_i)'\theta_{\kappa 0})P_{\kappa i}\right\| = o_p\big((nh^{-1/2})\big)$$

by Markov's inequality. It then follows from part (e) of Lemma 1.9.5: $\|\tilde{H}_{n\kappa} - H_\kappa\| = o_p(1)$ that

$$\mathbf{1}_n\left\|\bar{G}_{n\kappa}(\theta_{\kappa 0}) - n^{-1}H_\kappa^{-1}\sum_{i=1}^{n} t_i[Y_i - (1-\tau)]K_h^{(1)}(m_{\tau i})P_{\kappa i}\right\| = o_p\big((nh)^{-1/2}\big).$$

This completes the proof of the theorem since $\Pr(\mathbf{1}_n = 1) \longrightarrow 1$ as $n \longrightarrow \infty$. □

Proof of Theorem 1.3

Define

$$\begin{aligned}
\Gamma_\kappa &= \mathbb{E}\bigg[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{W})f_{m_\tau}(0 \mid \tilde{W})\frac{\mathrm{d}\lambda_\tau(V)}{\mathrm{d}v}P_\kappa(W)\bigg],\\
\Sigma_{1\kappa} &= \mathrm{Var}\big(n^{-1}\sum_{i=1}^{n}[Y_i - (1-\tau)]P_\kappa(W_i)K^{(1)}_h\big(m_{\tau i}\big)\big),\\
\Sigma_{2\kappa} &= \mathrm{Var}\big(n^{-1}\sum_{i=1}^{n}[Y_i - (1-\tau)]P_\kappa(W_i)K^{(2)}_h\big(m_{\tau i}\big)\big)\frac{\mathrm{d}\lambda_\tau(V)}{\mathrm{d}v},\\
\Omega_{1\kappa} &\equiv AH_\kappa^{-1}\Sigma_{1\kappa}H_\kappa^{-1}A', \text{ and}\\
\Omega_{2\kappa} &\equiv AH_\kappa^{-1}\Sigma_{2\kappa}H_\kappa^{-1}A'.
\end{aligned}$$

Then $\Sigma_{1\kappa}$ and $\Sigma_{2\kappa}$ can be approximated by

$$\frac{1}{nh}\tau(1-\tau)\bigg(\int K^{(1)}(u)^2\,\mathrm{d}u\bigg)\mathbb{E}[t(W)P_\kappa(W)P_\kappa(W)'f_{m_\tau}(0 \mid \tilde{X}, Z)]$$

and $\frac{1}{nh^3}\Gamma_\kappa\sigma^2_\zeta(V)\Gamma'_\kappa$, respectively.

Let

$$\begin{aligned}
\varphi_\kappa(w) &= AH_\kappa^{-1}P_\kappa(w)\\
&= \mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{X}, Z)f_{m_\tau}(0 \mid \tilde{X}, Z)\varphi(W)P_\kappa(W)']\times\\
&\quad \big(\mathbb{E}[t(W)F^{(1)}_{\varepsilon_\tau}(0 \mid 0, \tilde{X}, Z)f_{m_\tau}(0 \mid \tilde{X}, Z)P_\kappa(W)P_\kappa(W)']\big)^{-1}P_\kappa(w).
\end{aligned}$$

Then

$$\begin{aligned}
nh\Omega_{1\kappa} &= nhAH_\kappa^{-1}\Sigma_\kappa H_\kappa^{-1}A'\\
&= \tau(1-\tau)\bigg(\int K^{(1)}(u)^2\,\mathrm{d}u\bigg)\mathbb{E}[t(W)\varphi_\kappa(W)\varphi_\kappa(W)'f_{m_\tau}(0 \mid \tilde{X}, Z)]\\
&\longrightarrow \tau(1-\tau)\bigg(\int K^{(1)}(u)^2\,\mathrm{d}u\bigg)\mathbb{E}[t(W)\varphi(W)\varphi(W)'f_{m_\tau}(0 \mid \tilde{X}, Z)],
\end{aligned}$$

since $\mathbb{E}[t(W)\|\varphi(W) - \varphi_\kappa(W)\|^2] \to 0$ as $\kappa \to \infty$.

For part (a), recall that in the text $\varepsilon_\tau \equiv U - \lambda_\tau(V)$, $\hat{\beta} - \beta_0 = AH_\kappa^{-1}T_{1n\kappa} + AH_\kappa^{-1}T_{2n\kappa} + o_p\big((nh)^{-1/2}\big)$, and $\hat{V}_i - V_i = b(V_i) + \zeta_i\sigma_{\zeta_i}(V_i)$, $i = 1, \ldots, n$. Consider

$$\begin{aligned}\mathbb{E}[T_{1n\kappa} + T_{2n\kappa}] &= n^{-1}\sum_{i=1}^{n}\mathbb{E}\big[t_i[Y_i - (1-\tau)]K_h^{(1)}(m_{\tau i})P_\kappa(W_i)\big] \\ &\quad + n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[t_i[Y_i - (1-\tau)]K_h^{(2)}(m_{\tau i})\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}(b(V_i) + \zeta_i\sigma_{\zeta_i}(V_i))P_\kappa(W_i)\right].\end{aligned} \tag{1.36}$$

For the first term on the right hand side in (1.36), standard kernel smoothing arguments yield

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n}\mathbb{E}\big[t_i[Y_i - (1-\tau)]K_h^{(1)}(m_{\tau i})P_\kappa(W_i)\big] \\ &= \mathbb{E}\big[t_i[\mathbb{E}(Y_i \mid X_i, V_i) - (1-\tau)]K^{(1)}(m_{\tau i})P_\kappa(W_i)\big] \\ &= \mathbb{E}\big[t_i[1 - \Pr(U_i \le -X_i'\beta_0 \mid X_i, V_i) - (1-\tau)]K_h^{(1)}(m_{\tau i})P_\kappa(W_i)\big] \\ &= \mathbb{E}\big[t_i[\tau - F_{\varepsilon_\tau}(-m_{\tau i} \mid m_{\tau i}, \tilde{X}_i, V_i)]K_h^{(1)}(m_{\tau i})P_\kappa(W_i)\big] \\ &= \iint t(w)[\tau - F_{\varepsilon_\tau}(-s \mid s, \tilde{w})]K_h^{(1)}(s)P_\kappa(w)f_{m_\tau}(s \mid \tilde{w})\,\mathrm{d}s\,\mathrm{d}F_{\tilde{W}}(\tilde{w}),\end{aligned}$$

where the first equality follows from the i.i.d. assumption and the law of iterated expectation. Now making standard change-of-variables to $s = hu$, expanding and combining both $F_{\varepsilon_\tau}(-hu \mid hu, \tilde{w})$ and $f_{m_\tau}(hu \mid \tilde{w})$ around $hu = 0$ in Taylor expansions, and using the fact that $F_{\varepsilon_\tau}(0 \mid X, V) = \tau$, we have the following expression as a polynomial in hu

$$\begin{aligned}&\iint t(w)\left\{\big[\tau - F_{\varepsilon_\tau}\big(-hu \mid hu, \tilde{w}\big)\big]K^{(1)}(u)\,P_\kappa(w)f_{m_\tau}(hu \mid \tilde{w})\,\mathrm{d}u\,\mathrm{d}F_{\tilde{W}}(\tilde{w})\right\} \\ &= -\iint t(w)\left\{\left(\frac{1}{\nu!}F_{\varepsilon_\tau}^{(\nu)}(-\xi_\nu \mid \xi_\nu, \tilde{w})f_{m_\tau}(0 \mid \tilde{w}) + \sum_{i=1}^{\nu-1}\frac{1}{i!(\nu-i)!}F_{\varepsilon_\tau}^{(i)}\big(0 \mid 0, \tilde{x}, V\big)f_{m_\tau}^{(\nu-i)}\big(\xi_i \mid \tilde{w}\big)\right)h^\nu u^\nu\right. \\ &\quad \left. - \sum_{i=1}^{\nu-1}\sum_{j=0}^{\nu-i-1}\frac{1}{i!j!}F_{\varepsilon_\tau}^{(i)}\big(0 \mid 0, \tilde{w}\big)f_{m_\tau}^{(j)}(0 \mid \tilde{w})h^{i+j}u^{i+j}\right\}K^{(1)}(u)P_\kappa(w)\,\mathrm{d}u\,\mathrm{d}F_{\tilde{W}}(\tilde{w}) \\ &= O(h^\nu),\end{aligned}$$

where $\zeta_1, \ldots, \zeta_\nu$ are scalars with values between 0 and hu.

Then

$$\begin{aligned}
&\mathbb{E}(AH_\kappa^{-1}T_{1n\kappa}) \\
&\quad = h^\nu \times \alpha_\nu(K^{(1)}) \times \sum_{j=1}^{\nu} \frac{1}{j!(\nu-j)!}\mathbb{E}\big[t(W)F_{\varepsilon_\tau}^{(j)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}^{(\nu-j)}(0 \mid \tilde{W}) A H_\kappa^{-1} P_\kappa(W)\big] \\
&\quad = h^\nu \times \alpha_\nu(K^{(1)}) \times \sum_{j=1}^{\nu} \frac{1}{j!(\nu-j)!}\mathbb{E}\big[t(W)F_{\varepsilon_\tau}^{(j)}\big(0 \mid 0, \tilde{W}\big) f_{m_\tau}^{(\nu-j)}(0 \mid \tilde{W}) \varphi_\kappa(W)\big].
\end{aligned}$$

Since $\mathbb{E}[t(W)\|\varphi_\kappa(W) - \varphi(W)\|^2] \to 0$ by Assumption 1.9, we have $\|\mathbb{E}(AH_\kappa^{-1}T_{1n\kappa}) - B_1\| \to 0$. This proves the expression of B_1 in part (a).

By the analogous arguments to those used in Lemma 5 (a) of Horowitz (1992), Assumption 1.7, and Lebsegue's dominated convergence, one can show that $\mathbb{E}[h^{-\nu}AH_\kappa^{-1}T_{1n\kappa}] \to B$ and therefore $\lim_{n\to\infty}\sqrt{nh}h^\nu H_\kappa^{-1}\mathbb{E}(h^{-\nu}AT_{1n\kappa}) = \mu H^{-1}B$ where $\mu = \lim_{n\to\infty} nh^{2\nu+1}$.

Next we turn to the second term on the right hand side in (1.36). Plugging $\hat{V}_i - V_i = b_v(V_i) + \zeta_i\sigma_{\zeta_i}(V_i)$, the expectation of the term $T_{2n\kappa}$ can be approximated by

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[t_i[Y_i - (1-\tau)]\frac{\mathrm{d}\lambda_\tau(v)}{\mathrm{d}v}\frac{b_v(V_i)+\zeta_i\sigma_{\zeta_i}(V_i)}{h^2}K^{(2)}\Big(\frac{m_{\tau i}}{h}\Big)P_{\kappa i}\right] \\
&\quad = \mathbb{E}\left[t_i\big(\tau - F_{\varepsilon_\tau}(-m_{\tau i} \mid m_{\tau i}, \tilde{X}_i, V_i)\big)\frac{\mathrm{d}\lambda_\tau(v)}{\mathrm{d}v}\frac{b_v(V_i)+\zeta_i\sigma_{\zeta_i}(V_i)}{h^2}K^{(2)}\Big(\frac{m_{\tau i}}{h}\Big)P_{\kappa i}\right] \\
&\quad = \iint t(w)\big(\tau - F_{\varepsilon_\tau}(-s \mid s, \tilde{x}, v)\big)\frac{\mathrm{d}\lambda_\tau(v)}{\mathrm{d}v}\frac{b_v(v)+\zeta\sigma_\zeta(v)}{h^2}K^{(2)}\Big(\frac{s}{h}\Big)P_\kappa(w)f_{m_\tau}(s \mid \tilde{w})\mathrm{d}s\,\mathrm{d}F_{\tilde{W}}(\tilde{w}) \\
&\quad = \iint t(w)\big(\tau - F_{\varepsilon_\tau}(-hu \mid hu, \tilde{w})\big)K^{(2)}(u)\frac{\mathrm{d}\lambda_\tau(v)}{\mathrm{d}v}\frac{b_v\big(\lambda_\tau^{-1}(hu - x'\beta_0)\big) + \zeta\sigma_\zeta\big(\lambda_\tau^{-1}(hu - x'\beta_0)\big)}{h} \\
&\qquad\qquad \times P_\kappa(w)f_{m_\tau}(hu \mid \tilde{w})\mathrm{d}u\,\mathrm{d}F_{\tilde{W}}, \qquad (1.37)
\end{aligned}$$

where the third equality follows from the fact that $\mathbb{E}[\sigma_\zeta(V) \mid \tilde{W}] = 0$.

Similarly, using a Taylor expansion of the terms $F_{\varepsilon_\tau}(-hu \mid hu, \tilde{w})$, $b_v\big(\lambda_\tau^{-1}(hu - x'\beta_0)\big)$, and $f_{m_\tau}(hu \mid \tilde{w})$ in the argument hu and the fact that $F_{\varepsilon_\tau}(0 \mid 0, \tilde{w}) = \tau$ yields

$$\begin{aligned}
&\big(\tau - F_{\varepsilon_\tau}(-hu \mid hu, \tilde{W})\big)\frac{b_v\big(\lambda_\tau^{-1}(hu - X'\beta_0)\big)}{h} f_{m_\tau}(hu \mid \tilde{W}) \\
&\quad = -\Bigg\{ \frac{1}{(\nu-1)!} F_{\varepsilon_\tau}^{(\nu-1)}(0 \mid 0, \tilde{W}) f_{m_\tau}(0 \mid \tilde{W}) + \frac{1}{(\nu-2)!1!} F_{\varepsilon_\tau}^{(\nu-2)}(0 \mid 0, \tilde{W}) f_{m_\tau}^{(1)}(0 \mid \tilde{W}) \\
&\qquad \frac{1}{(\nu-3)!2!} F_{\varepsilon_\tau}^{(\nu-3)}(0 \mid 0, \tilde{W}) f_{m_\tau}^{(2)}(0 \mid \tilde{W}) + \cdots + \frac{1}{1!(\nu-2)!} F_{\varepsilon_\tau}^{(1)}(0 \mid 0, \tilde{W}) f_{m_\tau}^{(\nu-2)}(0 \mid \tilde{W}) \Bigg\} \\
&\qquad \times b_v(\lambda_\tau^{-1}(-X'\beta_0)) h^{\nu-2} u^{\nu-1}. \qquad (1.38)
\end{aligned}$$

Substituting (1.38) into (1.37), we end up with the following expression:

$$\begin{aligned}
&h^{\nu-2}\left(\int u^{\nu-1} K^{(2)}(u)\mathrm{d}u\right) \\
&\quad \times \sum_{j=1}^{\nu-1} \frac{1}{j!(\nu-j-1)!} \mathbb{E}\left[t(W) F_{\varepsilon_\tau}^{(j)}(0 \mid 0, \tilde{W}) f_{m_\tau}^{(\nu-j-1)}(0 \mid \tilde{W}) \frac{\mathrm{d}\lambda_\tau(V)}{\mathrm{d}v} b_v(\lambda_\tau^{-1}(-X'\beta_0)) P_\kappa(W) \right]. \qquad (1.39)
\end{aligned}$$

The same arguments as those used for the term $T_{1n\kappa}$ gives $\mathbb{E}[AH_\kappa^{-1}T_{2n\kappa}] \to B_2$. The expression (1.39) makes it clear that the bias of the adjustment term $AH_\kappa^{-1}T_{2n\kappa}$ is of the order $O(h^{\nu-2}g^r)$.

To calculate the variance in part (b), consider

$$\begin{aligned}
&\mathbb{E}[(T_{1n\kappa} + T_{2n\kappa})^2] \\
&\quad = \mathbb{E}\left[\left\{ n^{-1} \sum_{i=1}^{n} t_i[Y_i - (1-\tau)] K_h^{(1)}(m_{\tau i}) P_\kappa(W_i) + t_i[Y_i - (1-\tau)] K_h^{(2)}(m_{\tau i}) \frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v} (\hat{V}_i - V_i) P_\kappa(W_i) \right\}^2 \right] \\
&\quad = I_1 + 2I_2 + I_3,
\end{aligned}$$

where

$$I_1 = \mathbb{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{h^2}t_i t_j[Y_i-(1-\tau)][Y_j-(1-\tau)]K^{(1)}\Big(\frac{m_{\tau i}}{h}\Big)K^{(1)}\Big(\frac{m_{\tau j}}{h}\Big)P_\kappa(W_i)P_\kappa(W_j)'\right],$$

$$I_2 = \mathbb{E}\Bigg[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}t_i t_j[Y_i-(1-\tau)][Y_j-(1-\tau)]\frac{1}{h}K^{(1)}\Big(\frac{m_{\tau i}}{h}\Big)K^{(2)}\Big(\frac{m_{\tau j}}{h}\Big)\frac{\mathrm{d}\lambda_\tau(V_j)}{\mathrm{d}v}\frac{b(V_j)+\zeta_j\sigma_\zeta(V_j)}{h^2}$$
$$\times P_\kappa(W_i)P_\kappa(W_j)'\Bigg],$$

$$I_3 = \mathbb{E}\Bigg[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{h^2}t_i t_j[Y_i-(1-\tau)][Y_j-(1-\tau)]K^{(2)}\Big(\frac{m_{\tau i}}{h}\Big)\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\frac{V_i-\hat{V}_i}{h^2}$$
$$K^{(2)}\Big(\frac{m_{\tau j}}{h}\Big)\frac{\mathrm{d}\lambda_\tau(V_j)}{\mathrm{d}v}\frac{V_j-\hat{V}_j}{h^2}P_\kappa(W_i)P_\kappa(W_j)'\Bigg].$$

For the first term I_1, it is not difficult to show that I_1 is equal to $\beta_0 + 2\beta_0 B_1 + B_1^2 + \Sigma_{1\kappa}$ up to higher-order terms, where

$$\begin{aligned}
\Sigma_{1\kappa} &\equiv \mathrm{Var}\left[n^{-1}\sum_{i=1}^{n}[Y_i-(1-\tau)]K_h^{(1)}\big(m_{\tau i}\big)P_\kappa(W_i)\right] \\
&= n^{-1}\mathrm{Var}\left[[Y_i-(1-\tau)]K_h^{(1)}\big(m_{\tau i}\big)P_\kappa(W_i)\right] \\
&= n^{-1}\mathbb{E}\left[[Y_i-(1-\tau)]^2K_h^{(1)}\big(m_{\tau i}\big)^2P_{\kappa i}P_{\kappa i}'\right] - n^{-1}\left(\mathbb{E}\left[[Y_i-(1-\tau)]K_h^{(1)}\big(m_{\tau i}\big)P_\kappa(W_i)\right]\right)^2 \\
&= n^{-1}\iint K_h^{(1)}(s)^2P_\kappa(w)P_\kappa(w)'f_{m_\tau}(s\mid s,\tilde{w})\,\mathrm{d}s\,\mathrm{d}F_{\tilde{W}}(\tilde{w}) + o(1) \\
&= \frac{\tau(1-\tau)}{nh}\iint K^{(1)}(u)^2P_\kappa(w)P_\kappa(w)'f_{m_\tau}(hu\mid hu,\tilde{w})\,\mathrm{d}u\,\mathrm{d}F_{\tilde{W}}(\tilde{w}) + o(1) \\
&= \frac{\tau(1-\tau)}{nh}\left(\int K^{(1)}(u)^2\,\mathrm{d}u\right)\mathbb{E}\left[P_\kappa(W)P_\kappa(W)'f_{m_\tau}(0\mid\tilde{W})\right] + o_p(1),
\end{aligned} \tag{1.40}$$

where the first equality follows from the *i.i.d.* assumption of (Y_i, X_i, Z_i), the third uses the law of iterated expectations and exploits the fact that $\mathbb{E}[(Y_i-(1-\tau))^2 \mid X_i, V_i] = \tau(1-\tau)$, and the remaining lines follow from the standard change-of-variables to $s = hu$ and Taylor expansion arguments.

For the term I_3, we have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\Bigg[t_i t_j[Y_i-(1-\tau)][Y_j-(1-\tau)]\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\frac{\mathrm{d}\lambda_\tau(V_j)}{\mathrm{d}v}\frac{\zeta_i\zeta_j\sigma_\zeta(V_i)\sigma_\zeta(V_j)+b(V_i)b(V_j)}{h^4}\\
&\qquad\qquad\times K^{(2)}\left(\frac{m_{\tau i}}{h}\right)K^{(2)}\left(\frac{m_{\tau j}}{h}\right)P_\kappa(W_i)P_\kappa(W_j)'\Bigg]\\
&=\frac{\tau(1-\tau)}{n}\int t(w)h^{-4}\left[K^{(2)}\left(\frac{s}{h}\right)\right]^2\left(\frac{\mathrm{d}\lambda_\tau(v)}{\mathrm{d}v}\right)^2 f_{m_\tau}(s\mid\tilde{w})\,\mathrm{d}s\,\mathrm{d}F_{\tilde{W}}(\tilde{w})+O(n^{-1})\\
&\quad+\iiint h^{-4}b\big(\lambda_\tau^{-1}(-x'\beta_0+s_1)\big)b\big(\lambda_\tau^{-1}(-x'\beta_0+s_1)\big)K^{(2)}\left(\frac{s_1}{h}\right)K^{(2)}\left(\frac{s_2}{h}\right)\\
&\qquad\times f_{m_\tau}(s_1\mid\tilde{w})f_{m_\tau}(s_2\mid\tilde{w})\,\mathrm{d}s_1\,\mathrm{d}s_2\,\mathrm{d}F_{\tilde{W}}(\tilde{w})\\
&\simeq\frac{\tau(1-\tau)}{nh^3}\Sigma_{2\kappa}\\
&\quad+h^{2(\nu-2)}\left(\int u^{\nu-1}K^{(2)}(u)\mathrm{d}u\right)^2\\
&\qquad\times\left\{\sum_{j=1}^{\nu-1}\frac{1}{j!(\nu-j-1)!}\mathbb{E}\left[t(W)F_{\varepsilon_\tau}^{(j)}(0\mid 0,\tilde{W})f_{m_\tau}^{(\nu-j-1)}(0\mid\tilde{W})\frac{\mathrm{d}\lambda_\tau(V)}{\mathrm{d}v}b_v(\lambda_\tau^{-1}(-X'\beta_0))P_\kappa(W)\right]\right\}^2.
\end{aligned}$$

Part (b) follows from subtracting $\big(\mathbb{E}[T_{1n\kappa}+T_{2n\kappa}]\big)^2$ from $\mathbb{E}\big[(T_{1n\kappa}+T_{2n\kappa})^2\big]$.

For part (c), using the expressions of $\Sigma_{1\kappa}$ and $\Sigma_{2\kappa}$ derived above, we have

$$\begin{aligned}
\Omega_\kappa&=\min\{nh,n^2h^3g^{d_z}\}\big[\Omega_{1\kappa}+\Omega_{2\kappa}\big]\\
&=\min\{nh,n^2h^3g^{d_z}\}\Bigg\{\frac{R(K^{(1)})\tau(1-\tau)}{nh}\mathbb{E}[t(W)\varphi_\kappa(W)\varphi_\kappa(W)'f_{m_\tau}(0\mid\tilde{X},Z)]\\
&\quad+\frac{R(K^{(2)})}{nh^3}\mathbb{E}\Big[t(W)F_{\varepsilon_\tau}^{(1)}\big(0\mid 0,\tilde{W}\big)f_{m_\tau}(0\mid\tilde{W})\frac{\mathrm{d}\lambda_\tau(V)}{\mathrm{d}v}\varphi_\kappa(W)\Big]\sigma_\zeta^2\big(\lambda_\tau^{-1}(-X'\beta_0)\big)\\
&\qquad\times\mathbb{E}\Big[t(W)F_{\varepsilon_\tau}^{(1)}\big(0\mid 0,\tilde{W}\big)f_{m_\tau}(0\mid\tilde{W})\frac{\mathrm{d}\lambda_\tau(V)}{\mathrm{d}v}\varphi_\kappa(W)'\Big]\Bigg\}.
\end{aligned}$$

Note that

$$\begin{aligned}
\Omega_\kappa^{-1/2}A(\hat{\theta}_{n\kappa}-\theta_{\kappa 0})&=-\Omega_\kappa^{-1/2}AH_\kappa^{-1}n^{-1}\sum_{i=1}^{n}t_i[Y_i-(1-\tau)]K_h^{(1)}\big(m_{\tau i}\big)P_{\kappa i}\\
&\quad+\Omega_\kappa^{-1/2}AH_\kappa^{-1}n^{-1}\sum_{i=1}^{n}t_i[Y_i-(1-\tau)]K_h^{(2)}\big(m_{\tau i}\big)\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\big(\hat{V}_i-V_i\big)P_{\kappa i}+o_p\big((nh)^{-1/2}\big).
\end{aligned}$$

Define

$$\nu_{in} = AH_\kappa^{-1}\Omega_\kappa^{-1/2}t_i[Y_i - (1-\tau)]\left\{K_h^{(1)}\big(m_{\tau i}\big) + K_h^{(2)}\big(m_{\tau i}\big)\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}(\hat{V}_i - V_i)\right\}P_{\kappa i}$$

and

$$W_n = \sqrt{\delta}n^{-1}\sum_{i=1}^{n}[\nu_{in} - \mathbb{E}(\nu_{in})],$$

where $\delta \equiv \min\{nh, n^2h^3g^{d_z}\}$, $\mathbb{E}(\nu_{in}) = B_1 + B_2$, and $\mathbb{E}(\nu_{in}^2) = 1/\min\{nh, n^2h^3g^{d_z}\}$.

For any nonstochastic $(d_x-1)\times 1$ vector c with $c'c = 1$, consider the asymptotic distribution of $c'\Omega_\kappa^{-1/2}W_n$. Using the arguments analogous to those in Lemma 6 of Horowitz (1992), it can be shown that

$$\sqrt{\delta}c'\Omega_\kappa^{-1/2}A(\hat{\theta}_{n\kappa} - \theta_{\kappa 0}) = c'\Omega_\kappa^{-1/2}W_n + o_p(1) \xrightarrow{d} N(0, c'c). \tag{1.41}$$

The desired result follows by combining (1.41) with the expression of B_1 and B_2 derived above, Cramér-Wold device, and the fact that $\|\Omega_\kappa - \Omega\| \to 0$. □

Proof of Theorem 1.4

We begin with the following lemmas that are used to prove the consistency of $\Omega_{n\kappa}$.

Lemma 1.9.11. $\|\hat{H}_{n\kappa} - H_\kappa\| = o_p(1)$.

Proof. Define

$$\bar{H}_{n\kappa} = nh^{-2}\sum_{i=1}^{n}\hat{t}_i[Y_i - (1-\tau)]K^{(2)}\left(\frac{m_{\tau i}}{h}\right)\hat{P}_{\kappa i}\hat{P}_{\kappa i}'$$

and consider

$$\begin{aligned}
\|\hat{H}_{n\kappa} - \bar{H}_{n\kappa}\| &\le h^{-2}\max_{1\le i\le n}\hat{t}_i\left|K^{(2)}\left(\frac{P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}}{h}\right) - K^{(2)}\left(\frac{m_{\tau i}}{h}\right)\right|n^{-1}\sum_{i=1}^{n}\hat{t}_i\|\hat{P}_{\kappa i}\|^2 \\
&\le h^{-2}\max_{1\le i\le n}\hat{t}_i\left\{\left|K^{(2)}\left(\frac{P_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa}}{h}\right) - K^{(2)}\left(\frac{P_\kappa(W_i)'\hat{\theta}_{n\kappa}}{h}\right)\right|\right. \\
&\quad \left. + \left|K^{(2)}\left(\frac{P_\kappa(W_i)'\hat{\theta}_{n\kappa}}{h}\right) - K^{(2)}\left(\frac{m_{\tau i}}{h}\right)\right|\right\}n^{-1}\sum_{i=1}^{n}\hat{t}_i\|\hat{P}_{\kappa i}\|^2 \\
&\le Ch^{-3}\kappa\Big[\max_{1\le i\le n}\hat{t}_i|\hat{V}_i - V_i| + (\kappa/n)^{1/2}\Big] \\
&= O_p\big(h^{-3}\kappa\big[\Delta_v + \kappa^{1/2}n^{-1/2}\big]\big) = o_p(1),
\end{aligned}$$

where the last equality follows from Assumption 1.13.

Next consider

$$\Xi_{n\kappa 1} \equiv n^{-1}h^{-2}\sum_{i=1}^{n}\hat{t}_i\left\{[Y_i-(1-\tau)]K^{(2)}\left(\frac{m_{\tau i}}{h}\right)-\mathbb{E}\left[[Y_i-(1-\tau)]K^{(2)}\left(\frac{m_{\tau i}}{h}\right)\middle|\tilde{W}_i\right]\right\}\hat{P}_{\kappa i}\hat{P}'_{\kappa i}$$

$$\text{and } \Xi_{n\kappa 2} \equiv n^{-1}\sum_{i=1}^{n}\hat{t}_i\left\{\mathbb{E}\left[[Y_i-(1-\tau)]K^{(2)}\left(\frac{m_{\tau i}}{h}\right)\middle|\tilde{W}_i\right]-F^{(1)}_{\varepsilon_\tau}(0\mid 0,\tilde{W})f_{m_\tau}(0\mid\tilde{W}_i)\right\}\hat{P}_{\kappa i}\hat{P}'_{\kappa i}.$$

Following the same argument as that in the proof of Lemma A.9 of Lee (2007), it can be shown that $\|\Xi_{n\kappa 1}\| = o_p(1)$ and $\|\Xi_{n\kappa 2}\| = O_p(h^3\kappa) = o_p(1)$ by Assumption 1.13. The desired result then follows from the triangle inequality. □

Lemma 1.9.12. $\|\hat{\Sigma}_{n\kappa}-\Sigma_\kappa\| = o_p(1)$.

Proof. The arguments used to prove this theorem are identical to those used to prove Lemma 1.9.5(e) and (f). □

Lemma 1.9.13. $\|\hat{\Gamma}_{n\kappa}-\Gamma_\kappa\| = o_p(1)$.

Proof. Note that

$$\begin{aligned}
&\max_{1\le i\le n}\hat{t}_i\left|\frac{\mathrm{d}\hat{\lambda}_\tau(\hat{V}_i)}{\mathrm{d}v}-\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\right|\\
&\le \max_{1\le i\le n}\hat{t}_i\left|\frac{\mathrm{d}\hat{\lambda}_\tau(\hat{V}_i)}{\mathrm{d}v}-\frac{\mathrm{d}\lambda_\tau(\hat{V}_i)}{\mathrm{d}v}\right|+\max_{1\le i\le n}\hat{t}_i\left|\frac{\mathrm{d}\lambda_\tau(\hat{V}_i)}{\mathrm{d}v}-\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\right|\\
&\le O_p\big(\zeta_1(\kappa)(\kappa/n)^{1/2}\big)+O_p\big(\max_{1\le i\le n}t_i|\hat{V}_i-V_i|\big)\\
&\le O_p\big(\zeta_1(\kappa)(\kappa/n)^{1/2}+\Delta_v\big).
\end{aligned}$$

Define

$$\bar{\Gamma}_{n\kappa} = n^{-1}\sum_{i=1}^{n}\hat{t}_i[Y_i-(1-\tau)]K_h^{(2)}(\hat{P}_\kappa(\hat{W}_i)'\hat{\theta}_{n\kappa})\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\hat{P}'_{\kappa i}.$$

Then

$$\begin{aligned}
&\|\hat{\Gamma}_{n\kappa}-\bar{\Gamma}_{n\kappa}\|\\
&\le \max_{1\le i\le n}\hat{t}_i\left|\frac{\mathrm{d}\hat{\lambda}_\tau(\hat{V}_i)}{\mathrm{d}v}-\frac{\mathrm{d}\lambda_\tau(V_i)}{\mathrm{d}v}\right|\zeta_0(\kappa)n^{-1}h^{-2}\sum_{i=1}^{n}\hat{t}_i\left\|[Y_i-(1-\tau)]K^{(2)}\left(\frac{P'_\kappa(W_i)\hat{\theta}_{n\kappa}}{h}\right)\right\|\\
&= O_p\big(h^{-1}\zeta_0(\kappa)\zeta_1(\kappa)(\kappa/n)^{1/2}\big) = o_p(1),
\end{aligned}$$

provided that for power series $\rho_\kappa < \frac{2\nu+3}{9(2\nu+1)}$ and for splines $\rho_\kappa < \frac{2\nu+3}{5(2\nu+1)}$, as described in Assumption 1.13. Using the same argument as that in Lemma 1.9.11, one can show that $\|\bar{\Gamma}_{n\kappa} - \Gamma_\kappa\| = o_p(1)$ and the desired result follows by the triangle inequality. □

Proof of Theorem 1.4. This theorem can be proven by combining Lemmas 1.9.11-1.9.13 with the identical argument to that of Theorem 3.2 of Lee (2007). □

Bibliography

Abrevaya, J., J. A. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model With Endogenous Regressors," *Econometrica*, 78(6), 2043–2061.

Ahn, H., and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58(1-2), pp. 3–29.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), pp. 43–72.

Blundell, R., and J. L. Horowitz (2007): "A Non-Parametric Test of Exogeneity," *The Review of Economic Studies*, 74(4), pp. 1035–1058.

Blundell, R., and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. by L. P. H. Mathias Dewatripont, and S. J. Turnovsky, pp. 312–357. Cambridge University Press.

——— (2007): "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, 141(1), pp. 65–83.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, pp. 655–679.

Brown, B. W., and M. B. Walker (1989): "The Random Utility Hypothesis and Inference in Demand Systems," *Econometrica*, 57(4), pp. 815–829.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), pp. 189–218.

Chen, S., and S. Khan (2003): "Semiparametric Estimation of a Heteroskedastic Sample Selection Model," *Econometric Theory*, 19(6), pp. 1040–1064.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 76. Elsevier.

Chernozhukov, V., I. Fernandez-Val, and A. Kowalski (2011): "Quantile Regression with Censoring and Endogeneity," Cowles Foundation Discussion Papers 1797, Cowles Foundation for Research in Economics, Yale University.

Chernozhukov, V., P. Gagliardini, and O. Scaillet (2012): "Nonparametric Instrumental Variable Estimators of Structural Quantile Effects," *Econometrica*.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), pp. 245–261.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139(1), pp. 4–14.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), pp. 1405–1441.

——— (2009): "Single equation endogenous binary reponse models," CeMMAP working papers CWP23/09, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

——— (2010): "Instrumental Variable Models for Discrete Outcomes," *Econometrica*, 78(2), pp. 575–601.

——— (2011): "Semiparametric structural models of binary response: shape restrictions and partial identification," CeMMAP working papers CWP31/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Chesher, A., A. Rosen, and K. Smolinski (2011): "An instrumental variable model of multiple discrete choice," CeMMAP working papers CWP06/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70(1), pp. 33–58.

de Jong, R. M. (2002): "A note on convergence rates and asymptotic normality for series estimators: uniform convergence rates," *Journal of Econometrics*, 111(1), 1–9.

D'Haultfoeuille, X., and P. Fvrier (2012): "Identification of Nonseparable Modes with Endogeneity and Discrete Instruments," Working Papers 2011-28, Centre de Recherche en Economie et Statistique.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76(5), pp. 1191–1206.

Frandouml, M. (2006): "Non-parametric regression for binary dependent variables," *Econometrics Journal*, 9(3), 511–540.

Haan, P. (2006): "Much ado about nothing: conditional logit vs. random coefficient models for estimating labour supply elasticities," *Applied Economics Letters*, 13(4), 251–256.

Hahn, J., and G. Ridder (2010): "The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors," Textos para discusso 575, Department of Economics PUC-Rio (Brazil).

Han, A. K. (1987): "Non-parametric analysis of a generalized regression model : The maximum rank correlation estimator," *Journal of Econometrics*, 35(2-3), pp. 303–316.

Hoderlein, S. (2009): "Endogenous semiparametric binary choice models with heteroscedasticity," CeMMAP working papers CWP34/09, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Hoderlein, S., and Y. Sasaki (2011): "Necessary and Sufficient Conditions for Identification of Causal Effects in Endogenous Nonseparable Models," Working papers.

Hong, H., and E. Tamer (2003): "Endogenous binary choice model with median restrictions," *Economics Letters*, 80(2), pp. 219–225.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), pp. 505–31.

Horowitz, J. L. (2009): *Semiparametric and Nonparametric Methods in Econometrics (Springer Series in Statistics)*. Springer, corrected edn.

——— (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79(2), pp. 347–394.

——— (2012): "Nonparametric additive models," CeMMAP working papers CWP20/12, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Horowitz, J. L., and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75(4), pp. 1191–1208.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), pp. 1481–1512.

Kasy, M. (2011): "Identification In Triangular Systems Using Control Functions," *Econometric Theory*, 27(03), pp. 663–671.

——— (2012): "Identification in General Triabgular Systems," Discussion paper, Harvard University.

Klein, R. W., and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), pp. 387–421.

Koenker, R., and J. Bassett, Gilbert (1982): "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50(1), pp. 43–61.

Kordas, G. (2006): "Smoothed Binary Regression Quantiles," *Journal of Applied Econometrics*, 21(3), pp. 387–407.

Krief, J. M. (2011): "Kernel Weighted Smoothed Maximum Score Estimation for Applied Work," Departmental working papers, Department of Economics, Louisiana State University.

Lee, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141(2), pp. 1131–1158.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97(1), 145 – 177.

——— (2012): "An Overview of the Special Regressor Method," Boston college working papers in economics, Boston College Department of Economics.

Lewbel, A., Y. Dong, and T. T. Yang (2012): "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors," Boston College Working Papers in Economics 789, Boston College Department of Economics.

Li, Q. (2000): "Efficient Estimation of Additive Partially Linear Models," *International Economic Review*, 41(4), 1073–1092.

Li, Q., and J. S. Racine (2008): "Nonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous Data," *Journal of Business and Economic Statistics*, 26, 423–434.

Li, Q., and J. M. Wooldridge (2002): "Semiparametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors," *Econometric Theory*, 18(3), pp. 625–645.

Lorentz, G. G. (1966): *Approximation of Functions*. New York: Chelsea.

Mammen, E., C. Rothe, and M. Schienle (2011): "Semiparametric Estimation with Generated Covariates," SFB 649 Discussion Papers SFB649DP2011-064, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

——— (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annuals of Statistics*.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3(3), pp. 205–228.

——— (1985): "Semiparametric analysis of discrete response : Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), pp. 313–333.

——— (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83(403), pp. 729–738.

Masry, E. (1996): "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17(6), 571–599.

Matzkin, R. L. (1992): "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and The Binary Choice Models," *Econometrica*, 60(2), pp. 239–270.

——— (2007): "Nonparametric identification," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 73. Elsevier.

Maurer, J., R. W. Klein, and F. Vella (2011): "Subjective Health Assessments and Active Labor Market Participation of Older Men: Evidence from a Semiparametric Binary

Choice Model with Nonadditive Correlated Individual-specific Effects," *The Review of Economics and Statistics*, 93(3), 764–774.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), pp. 147–168.

——— (2009): "Two-step series estimation of sample selection models," *Econometrics Journal*, 12, S217–S229.

Newey, W. K., and D. McFadden (1994): "Chapter 36 Large sample estimation and hypothesis testing," vol. 4 of *Handbook of Econometrics*, pp. pp. 2111–2245. Elsevier.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), pp. 565–603.

Pacifico, D. (2012): "On the role of unobserved preference heterogeneity in discrete choice models of labour supply," *Empirical Economics*, pp. 1–35, 10.1007/s00181-012-0637-6.

Rivers, D., and Q. H. Vuong (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39(3), 347–366.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), pp. 931–954.

Rothe, C. (2009): "Semiparametric estimation of binary response models with endogenous regressors," *Journal of Econometrics*, 153(1), pp. 51–64.

——— (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, 155(1), pp. 56–70.

Shaikh, A. M., and E. Vytlacil (2008): "Endogenous binary choice models with median restrictions: A comment," *Economics Letters*, 98(1), pp. 23–28.

Shaikh, A. M., and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79(3), pp. 949–955.

Silverman, B. W. (1986): *Density estimation: for statistics and data analysis*. London.

Smith, R. J., and R. W. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54(3), 679–85.

Song, K. (2008): "Uniform Convergence Of Series Estimators Over Function Spaces," *Econometric Theory*, 24(06), 1463–1499.

Sperlich, S. (2009): "A note on non-parametric estimation with predicted variables," *Econometrics Journal*, 12(2), 382–395.

Su, L., and A. Ullah (2008): "Local polynomial estimation of nonparametric simultaneous equations models," *Journal of Econometrics*, 144(1), 193–218.

Tsallis, C., and D. A. Stariolo (1996): "Generalized simulated annealing," *Physica A: Statistical Mechanics and its Applications*, 233(12), 395 – 406.

Vytlacil, E., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75(3), pp. 757–779.

Chapter 2

An Empirical Analysis of Female Labor Market Participation and Endogenous Non-labor Income

2.1 Introduction

This chapter illustrates estimation of a triangular binary response model for labor market participation with endogenous income. To motivate our approach, we begin by reviewing the background of the (static) labor supply analysis. In the area of labor economics, labor supply modelling has received a great deal of attention in the literature, with a particular interest in implications for responsiveness of labor supply to wages, taxes, and transfers, see Blundell, MaCurdy, and Meghir (2007) and Keane (2011) for recent surveys of the literature on labor supply models. In a static labor supply model, in contrast to the utility function approach that is typically employed in dynamic settings, it is convenient to directly specify labor supply regressions of hours of work on wages and non-labor income. As discussed in Keane (2011, p.971), there are several econometric issues in estimating a static labor supply model. Among the most obvious problem is the endogeneity of wages and non-labor income arising from correlation with unobserved heterogeneity in preferences for work. For example, preferences for work would be positively correlated with wages through productivity. We discuss the endogeneity source of non-labor income in more detail later. To address the endogeneity problem, one can use the IV (or 2SLS) approach in linear models or adopt a fixed effects specification in the panel data context. When the labor supply function is nonlinear, nonparametric IV and CF methods are available, e.g., Newey and Powell (2003) and Newey, Powell, and Vella (1999). On the other hand, due to the selection problem, incorporating

labor market participation is essential for modelling female labor supply decisions. However, the presence of endogenous regressors in a binary response participation model makes the analysis fundamentally different from that in continuous models for hours of work. In the context of modelling labor market participation with endogenous non-labor income, the CF method is particularly useful. But unlike the IV assumption, in practice it may be difficult to justify the CF assumption from economic theory, as pointed out by Blundell, MaCurdy, and Meghir (2007). This motivates the use of our methodology proposed in Chapter 1 under weak restrictions in the current application.

The literature on binary response models of labor market participation is large. Most of the previous work assume that explanatory variables are exogenous, e.g., ,among others. More recently, the endogeneity problem in labor market participation has been addressed in the literature. For example, Blundell and Powell (2004) allow for the other income to be endogenous and demonstrate the importance of accounting for endogenous other income in their empirical findings. Carrasco (2001) considers estimating the effect of endogenous fertility on female labor participation in the panel data setting. Maurer, Klein, and Vella (2011) use a semiparametric binary choice panel data model to estimate the effect of endogenous subjective health assessment on the labor market participation of older men.

This chapter is organized as follows. Section 2.2 presents the empirical model. Section 2.3 briefly describe the 2011 U.S. Current Population Survey used to estimate the relationship between wives' non-labor income and their labor market participation. Section 2.4 presents estimation results using CPS data. Section 2.5 concludes.

2.2 Empirical Model

We turn to formulating the empirical models. Observed hours of work h_i of individual i can be represented by

$$h_i = \max\{h^*(W_i, Z_{1i}, I_i, \zeta_i) - h_i^r,\ 0\},$$

where h_i^* and h_i^r represent desired and reservation hours of work, respectively, W_i is the hourly wage rate, Z_{1i} are observable social demographic variables, I_i is wife's non-labor

income including the earned income of the spouse, and ζ_i is unobserved heterogeneity.[1] Let Y_i denote a binary variable that is unity if individual i participates and zero otherwise. The participation decision may be formulated by comparing h_i^* and h_i^r:[2]

$$Y_i = \mathbf{1}\{h_i^* > h_i^r\}. \tag{2.1}$$

The empirical specification of the labor market participation model is based on Blundell and Powell (2004):

$$h_i^* = \delta_0 + Z_{1i}'\delta_1 + \ln W_i\delta_2 + X_{1i}\delta_3 + \zeta_i, \tag{2.2}$$

$$h_i^r = \pi_0 + Z_{1i}'\pi_1 + \xi_i, \tag{2.3}$$

where $X_{1i} = \ln I_i$ and ζ_i is unobserved heterogeneity. Combining (2.1), (2.2), and (2.3) and replacing the unobserved wage rate for non-participants with the wage equation: $\ln W_i = \theta_0 + Z_{1i}'\theta_1 + \omega_i$ yields the model for labor force participation

$$Y_i = \mathbf{1}\{\beta_{00} + Z_{1i}'\beta_{10} + X_{1i}\beta_{20} + U_i > 0\}. \tag{2.4}$$

The key parameter of interest is β_{20} (expected sign is negative). It may be well be the case that wife's non-labor income is endogenous to labor force participation. As discussed in Blundell, MaCurdy, and Meghir (2007, section 2.2), the main estimation issue of the participation model (2.4) is the endogeneity of non-labor income X_{1i}, in the sense that X_{1i} is correlated with the unobserved term U_i. There are at least three reasons why non-labor income is likely to be endogenous. First, if non-labor income partly represents asset income, it may be correlated with unobserved heterogeneity in preferences for work since individuals

[1] In order to take into account fixed costs upon entry into the labor market, Cogan (1981) defines reservation hours of work h^r as $\mathcal{U}(H - h^r, I - F + Wh^r) = \mathcal{U}(H, I)$ where $\mathcal{U}$ is the utility function, H is time endowment, and F is a fixed cost.

[2] As pointed out by Cogan (1981), this formulation is equivalent to the market wage-reservation wage characterization. That is, $h_i > 0 \iff h_i^* > h_i^r \iff W_i > W_i^r$ where W_i^r is defined implicitly by $h^*(W_i^r, Z_{1i}, I_i, \zeta_i) = h_i^r$. Note that if there are fixed costs of labor market entry, the reservation hours would be positive and can be described by the discontinuity of the labor supply function at the reservation wage. Another advantage of introducing the reservation hours equation into the model is that, relative to the classic Heckman's labor supply model, it leads to fewer constraints between the parameters in equations of participation and hours-to-work. For details see Zabel (1993).

who worked more in the past are likely to have higher levels of non-labor income today. Second, hard-working individuals may tend to marry due to positive assortative mating. As such, non-labor income including the earned income of the spouse may well be correlated with unobservables affecting taste for work. Third, it is often the case that measurement error in the income measure, inducing correlation between X_1 and U, exists. To handle this endogeneity problem, wife's non-labor income is assumed to be determined by the following general reduced-form model

$$X_{1i} = f(Z_i, \eta_i), \tag{2.5}$$

where $Z_i = (Z_{1i}', Z_{2i})'$ and Z_{2i} is the education level of the spouse, which is the excluded instrument.

2.3 Data

To estimate the model we use the data set extracted from the 2011 March Supplement to the US Current Population Survey. The sample comprises $6,645$ white married women aged between 22 and 65 with non-hispanic origin residing in the Midwest region.[3] The binary dependent variable Y is the indicator variable for labor force participation, defined as 1 if wife's usual hours worked were positive and 0 otherwise. We select the following explanatory and excluded instrumental variables: the endogenous regressor (X_1) is the logarithm of wife's non-labor income, where wife's non-labor income is computed by subtracting wife's wage and salary earnings from total family income; the Z_1 matrix contains educational attainment, potential labor market experience, squared potential labor market experience, and the presence and age of children in the household. Following Blundell and Powell (2004),[4] we use as an instrument spouse's education level (Z_2) since it should affect wife's non-labor

[3]To justify treating schooling as exogenously determined, we follow Eckstein and Lifshitz (2011) by drawing married women starting at age 22, at which schooling is implicitly assumed to be given.

[4]Another instrumental variable that is used in the empirical application in Blundell and Powell (2004) is a welfare benefit entitlement variable, for the construction of this variable see Blundell and Powell (2004, p.669).

income but has no direct influence on wife's labor force participation. Some descriptive statistics for these variables are summarized in Tables 2.1 and 2.2.

2.4 Estimation Results

To facilitate comparisons, we exclude the intercept and normalize the coefficient on the education variable to unity. We first estimate the participation equation (2.4) by parametric approaches, including standard probit and two-stage probit (2SProbit) accounting for endogeneity. The latter uses the residual from the linear reduced form model for non-labor income as an additional regressor, which does not directly account for heteroskedasticity in the reduced form model. The coefficient estimates are reported in columns 1-3 of Table **??**, respectively, with standard errors in parentheses. The reduced form results in column 1 indicate that the spouse's education level plays an important role in the determination of wife's non-labor income. The coefficient of wife's non-labor income is of the expected sign (negative), showing that non-labor income influences negatively on the probability of working. This evidence is consistent with a positive income effect on a wife's demand for leisure. This effect becomes more negative when endogeneity is accounted for via the 2SProbit procedure, being more than twice as large as the probit counterpart. Such estimates indicate that wife's non-labor income is positively correlated with unobserved heterogeneity in tastes for work, as we discussed in Section 2.2. The substantial difference between coefficient estimates of non-labor income, combined with statistical significance of the coefficient estimate for the first-stage residual, suggests that non-labor income is endogenous to labor market participation. As expected, wives with more years of schooling are likely to work than less educated wives. Labor market experience has a quadratic effect on labor force participation. Wives with children under the age of six have a larger effect on labor force participation than the number of older children over the age of six.

While there are some striking difference between Probit and 2SProbit estimates, the 2SProbit estimator is highly biased and inconsistent in the presence of heteroskedasticity in both outcome and reduced form equations, as shown in simulation work in the preceding

section. To address this issue, we perform a simple likelihood ratio (LR) test for homoskedasticity based on Probit estimates. The form of heteroskedasticity tested is $\text{Var}(U) = \exp(\bar{Z}'c)$, where $\bar{Z}$ is the vector of variables suspected of causing heteroskedasticity and c is a conformable vector of parameters. We conduct the LR test for each of the variables in Z and X_1 and report the results in Table 2.3. The test results indicate that heteroskedasticity is present through several of the variables, including non-labor income, education, and education of husband at the 5% level and the number of children aged between 6 and 18 at the 10% level. In addition, it is expected that the conditional variance of the non-labor income variable may depend on the education level of the spouse, as indicated in Figure 2.1. The studentized Breusch-Pagan test strongly rejects homoskedasticity (with the p-value 0.0043) in the linear reduced form model for non-labor income. Evidence of the presence of heteroskedasticity in both participation and non-labor income reduced form models strengthens the value of our TBRQ approach. Moreover, Figure 2.2 gives a graph of kernel density estimates for log non-labor income using the Epanechnikov kernel function and the least square cross-validation bandwidth (0.140).

We now turn to estimating coefficients in the participation equation (2.4) by Horowitz's smoothed maximum score (SMS) (without taking the endogeneity problem into account) and by the TBRQ. Analogous to simulation experiments, we use the smoothed version of the Nadaraya-Watson estimator and rule-of-thumb bandwidths when estimating the conditional CDF in the first stage. To allow the control function λ_τ to be nonlinear, capturing the possibly nonlinear relationship between errors U and V, we employ quadratic B-spline base functions to approximate the unknown control function λ_τ. Following the simulation results in the preceding section, we consider $\kappa = 7, 8, 9, 10, 11$, and 12 and $h = C_h \times n^{-1/5} = 0.172$ (where $C_h = 1$ and $n = 6645$) in the second stage. For the range of κ considered here, varying C_h from 1 to 0.5 does not change estimation results dramatically and the coefficients estimates are not very sensitive to the choice of κ, as presented in Table 2.7. On the other hand, TBRQ estimates seem somewhat sensitive to the values of the starting point for this data set. To address this issue, for each κ, we run the TBRQ estimation procedure 100 times using the

GSA global optimization algorithm with different starting values that are randomly chosen from the feasible ranges, and then report the best of these solutions. Implementation of 100 runs took approximately 23 ~ 36 hours (depending on the values of κ) to complete in R 2.15.1 on a 1.70 GHz and 4 GB RAM personal computer. We only report estimation results for $\kappa = 7$ and 9 in Table 2.4. The standard errors of the TBRQ estimates are obtained using asymptotic approximation based on Theorem 1.4 in Chapter 1. In addition to the case $\tau = 0.5$, in Table 2.5 we perform the TBRQ for different quantiles: $\tau = 0.1, 0.25, 0.45, 0.55, 0.75, 0.9$ and find that coefficients at different quantiles are significantly different, suggesting that in our application the DER is unlikely to be satisfied, which is consistent with evidence of the presence of heteroskedasticity in the binary outcome (participation) equation, and the estimators using the DER are inconsistent. This also explains why estimation results obtained from the TBRQ and 2SProbit based on strong assumptions are quite different.

In summary, all TBRQ coefficients (for $\kappa = 7$ and 9) have expected signs with statistical significance, but with large differences in the magnitude of coefficients when comparing to 2SProbit and SMS counterparts. However, qualitatively similar to the probit estimates, accounting for endogeneity leads to a substantial increase in the magnitude of the non-labor income coefficient, being 62% ~ 77% larger than that in the SMS. This appears to be qualitatively similar for quantiles between $\tau = 0.45$ and $\tau = 0.55$ (see Table 2.6). This finding is analogous to that in Blundell and Powell (2004) using British data. Another noteworthy difference is that, in contrast to probit estimates, older children have roughly equal negative effect on the probability of participation to younger children.

The sixth column of Table 2.4 corresponds to the case in which $\lambda_\tau(v)$ is specified as a linear function. For this approach we obtain the point estimate for the non-labor income coefficient being -0.094, with a standard error 0.164. The coefficient estimate of the control variable is -0.745 and not significantly different from zero. These results provide evidence that a linear specification of $\lambda_\tau(v)$ may not be sufficient to capture the endogeneity from wife's non-labor income. In contrast, an important aspect of our estimation procedure is the ability to allow the control function to be nonlinear. To examine this we plot point estimates of the

control function $\lambda_{\tau=0.5}(v)$ for the linear case and κ = 7 and 9 in Figure 2.3 for comparison purposes. The graph shows that the control function λ_τ is unlikely to be constant and the estimated relationship between errors U and V is highly nonlinear, indicating that a model that allows for more flexible specification of the control function is needed.

Further interesting extensions of this application will obtain interval estimates of the participation probabilities and examine marginal effects by using a set of quantiles, as discussed in Appendix Section 1.9 in Chapter 1. We leave this for future research.

2.5 Conclusion

We have applied the methodology to deal with the endogeneity problem of non-labor income in female labor market participation. Three main conclusions emerged from this empirical analysis. First, there is evidence of the presence of heteroskedasticity in the outcome and reduced form equations, indicating that 2SProbit estimates are biased and inconsistent. Second, qualitatively similar to the probit estimates, accounting for endogeneity leads to a substantial increase in the magnitude of the non-labor income coefficient, being 62% ~ 77% larger than that in the smoothed maximum score estimation. For the range of the number of series expansion terms (κ) considered here, varying the second-stage bandwidth by increasing C_h from 1 to 0.5 does not change estimation results dramatically and the coefficients estimates are not very sensitive to the choice of κ. However, estimation results are somewhat sensitive to the number of series expansion term when C_h = 1.5. This suggests that a challenging and important task is to find data-based methods for optimally choosing the bandwidths and the number of the series expansion terms κ in practice.

2.6 Appendix: Figures and Tables

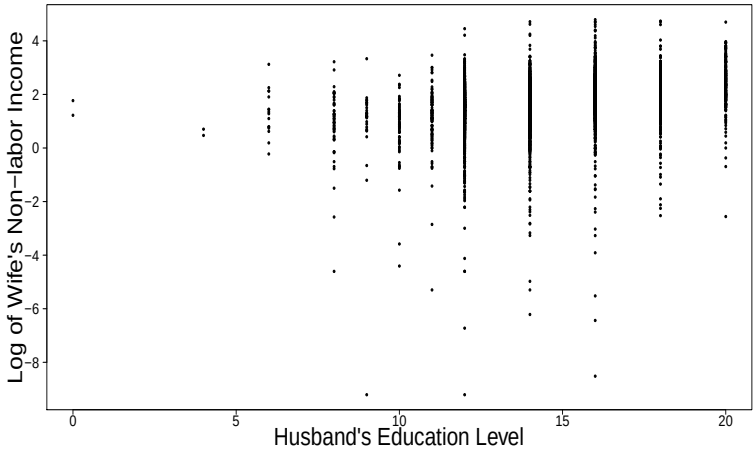

Figure 2.1 Log wife's non-labor Income and husband's education level

Table 2.1 Descriptive statistics

Variable	Definition	Mean	Std. dev.	Min	Max
Labor Force Participation	Participation indicator	0.702	0.458	0	1
Age	-	44.161	10.867	22	65
Education of husband	Husband's years of schooling/10	1.422	0.245	0	2
Education	Years of schooling/10	1.435	0.224	0	2
Experience	Potential labor force experience =(Age/10-Education-0.6)	2.382	1.132	-0.1	5.6
No. of Children < 6 yrs old	-	0.363	0.695	0	5
No. of Children 6 – 18 yrs old	-	0.782	1.038	0	7
Non-labor income	Log of wife's non-labor income (in $10,000)	1.637	0.897	-9.210	4.791

Notes: Non-labor income is defined as total family income excluding wife's wage and salary earnings.

Table 2.2 Descriptive statistics grouped by LFP

	LFP=1		LFP=0	
Variable	Mean	Std. dev.	Mean	Std. dev.
Age	43.535	10.324	45.635	11.918
Education of husband	1.425	0.234	1.412	0.267
Education	1.455	0.218	1.386	0.230
Experience	2.299	1.081	2.577	1.223
No. of Children < 6 yrs old	0.346	0.673	0.403	0.741
No. of Children 6 – 18 yrs old	0.783	1.007	0.781	1.108
Non-labor income	1.613	0.893	1.691	0.904
No. of Observations	4663		1982	

Table 2.3 p-values for LR tests for heteroskedasticity

Variable	p-values
Non-labor income	0.0069
Education of husband	0.0500
Education	0.0079
Experience	0.9750
No. of Children < 6 yrs old	0.1218
No. of Children 6 – 18 yrs old	0.0988

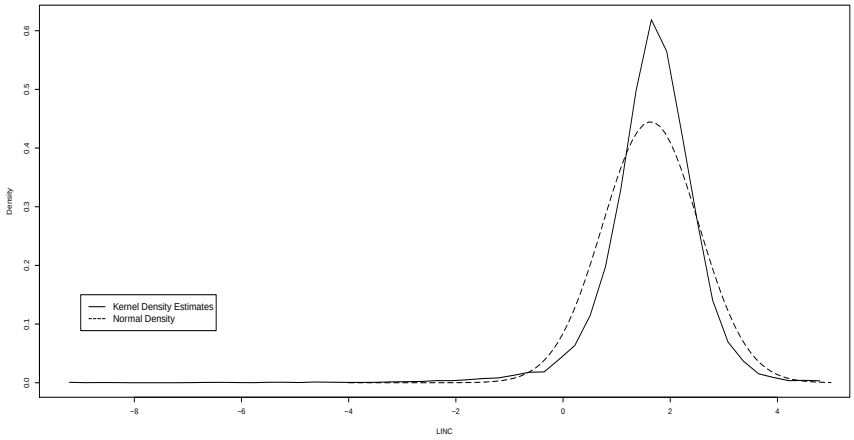

Figure 2.2 Kernel density estimates of log wife's non-labor income

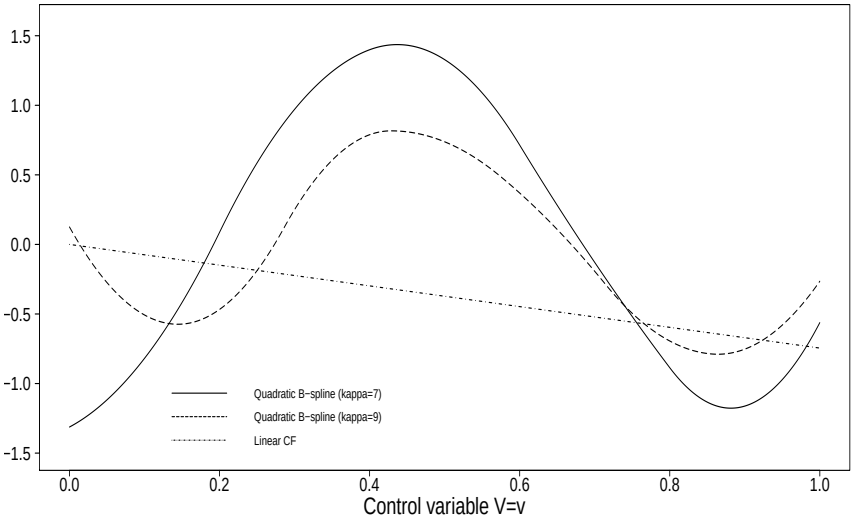

Figure 2.3 Point estimates of $\lambda_{0.5}(v)$

Table 2.4 Estimation results

	Parametric Estimation			Semiparametric Estimation			
	Linear RF	Probit	2SProbit	SMS	TBRQ ($\tau = 0.5$)		
Variable	for LINC				Linear $\lambda_\tau(v)$	$\kappa = 7$	$\kappa = 9$
Non-labor income	-	-0.189	-0.385	-0.261	-0.094	-0.424	-0.463
	-	(0.034)	(0.098)	(0.107)	(0.164)	(0.109)	(0.225)
Experience	0.472	0.630	0.611	2.696	2.857	6.809	4.948
	(0.043)	(0.101)	(0.091)	(1.057)	(0.569)	(2.893)	(0.983)
Exp. Square	-0.074	-0.201	-0.180	-0.681	-0.708	-1.669	-1.207
	(0.009)	(0.021)	(0.019)	(0.251)	(0.126)	(0.704)	(0.245)
No. of Children	0.037	-0.426	-0.347	-0.614	-0.619	-1.417	-1.069
< 6 yrs old	(0.018)	(0.044)	(0.039)	(0.117)	(0.113)	(0.537)	(0.229)
No. of Children	0.043	-0.210	-0.162	-0.717	-0.735	-1.530	-1.107
$6 - 18$ yrs old	(0.011)	(0.029)	(0.025)	(0.199)	(0.102)	(0.662)	(0.226)
Educ. of husband	1.13	-	-	-	-	-	-
	(0.050)	-	-	-	-	-	-
Education	0.240	1.000	1.000	1.000	1.000	1.000	1.000
	(0.057)	-	-	-	-	-	-
Control	-	-	0.249	-	-0.745	-	-
variable	-	-	(0.102)	-	(0.732)	-	-

Notes: LINC stands for logarithm of wife's non-labor income. The 2SProbit and TBRQ use the linear reduced form residual and the conditional CDF of LINC given instruments as control variables, respectively. The SMS is Horowitz's (1992) smoothed maximum score estimator. Standard errors (based on 500 bootstrap replications for 2SProbit and the asymptotic formulas otherwise) in parentheses.

Table 2.5 TBRQ estimation results for different quantiles ($\kappa = 7$)

	TBRQ						
Variable	$\tau = 0.1$	0.25	0.45	0.5	0.55	0.75	0.9
Non-labor income	1.509	-0.229	-0.548	-0.424	-0.182	-0.731	1.340
	(1.876)	(0.095)	(0.536)	(0.109)	(0.190)	(0.214)	(2.011)
Experience	-14.739	2.110	8.833	6.809	5.822	5.176	20.000
	(17.153)	(0.765)	(4.088)	(2.893)	(1.315)	(2.760)	(30.270)
Exp. Square	2.610	-0.665	-2.182	-1.669	-1.379	-1.145	-4.351
	(2.993)	(0.254)	(1.010)	(0.704)	(0.316)	(0.605)	(6.510)
No. of Children	-2.048	-1.412	-1.894	-1.417	-1.365	6.816	12.102
< 6 yrs old	(2.538)	(0.588)	(0.946)	(0.537)	(0.305)	(4.112)	(5.278)
No. of Children	0.122	-0.706	-1.859	-1.530	-1.497	3.223	-2.276
6 – 18 yrs old	(0.133)	(0.418)	(0.870)	(0.662)	(0.372)	(1.340)	(2.373)
Education	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	-	-	-	-	-	-	-

Notes: Standard errors in parentheses.

Table 2.6 SMS estimation results for different quantiles

	SMS						
Variable	$\tau = 0.1$	0.25	0.45	0.5	0.55	0.75	0.9
Non-labor income	-2.630	-0.830	-0.323	-0.261	-0.136	0.766	-0.463
	(1.223)	(0.108)	(0.117)	(0.107)	(0.072)	(0.086)	(0.225)
Experience	-7.079	2.366	1.594	2.696	6.255	13.445	4.948
	(3.141)	(0.471)	(0.505)	(1.057)	(0.760)	(1.721)	(0.983)
Exp. Square	1.008	-0.761	-0.429	-0.681	-1.474	-2.923	-1.207
	(3.489)	(0.139)	(0.112)	(0.251)	(0.178)	(0.368)	(0.245)
No. of Children	-4.878	-1.040	-0.681	-0.614	-1.022	1.007	-1.069
< 6 yrs old	(2.866)	(0.136)	(0.101)	(0.117)	(0.108)	(0.796)	(0.229)
No. of Children	0.335	-0.600	-0.485	-0.717	-1.404	-2.873	-1.107
$6 - 18$ yrs old	(0.567)	(0.120)	(0.080)	(0.199)	(0.147)	(0.379)	(0.226)
Education	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	-	-	-	-	-	-	-

Notes: Standard errors in parentheses.

Table 2.7 Coefficient estimates of non-labor income for different smoothing parameters ($\tau = 0.5$)

	$C_h = 0.5$	$C_h = 1.0$	$C_h = 1.5$
$\kappa = 7$	-0.400 (0.073)	-0.424 (0.109)	-0.514 (0.373)
$\kappa = 8$	-0.300 (0.039)	-0.192 (0.213)	-0.473 (1.512)
$\kappa = 9$	-0.409 (0.080)	-0.463 (0.225)	0.114 (0.205)
$\kappa = 10$	-0.264 (0.810)	-0.410 (0.154)	0.064 (2.553)
$\kappa = 11$	-0.239 (0.236)	-0.253 (0.489)	0.015 (1.046)
$\kappa = 12$	-0.252 (0.215)	-0.300 (0.839)	-0.161 (0.487)

Notes: Standard errors in parentheses.

Bibliography

Blundell, R., T. MaCurdy, and C. Meghir (2007): "Chapter 69 Labor Supply Models: Unobserved Heterogeneity, Nonparticipation and Dynamics," vol. 6, Part A of *Handbook of Econometrics*, pp. 4667 – 4775. Elsevier.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, pp. 655–679.

Carrasco, R. (2001): "Binary Choice with Binary Endogenous Regressors in Panel Data: Estimating the Effect of Fertility on Female Labor Participation," *Journal of Business and Economic Statistics*, 19(4), pp. 385–394.

Cogan, J. F. (1981): "Fixed Costs and Labor Supply," *Econometrica*, 49(4), pp. 945–963.

Eckstein, Z., and O. Lifshitz (2011): "Dynamic Female Labor Supply," *Econometrica*, 79(6), 1675–1726.

Keane, M. P. (2011): "Labor Supply and Taxes: A Survey," *Journal of Economic Literature*, 49(4), pp. 961–1075.

Maurer, J., R. W. Klein, and F. Vella (2011): "Subjective Health Assessments and Active Labor Market Participation of Older Men: Evidence from a Semiparametric Binary Choice Model with Nonadditive Correlated Individual-specific Effects," *The Review of Economics and Statistics*, 93(3), 764–774.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), pp. 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), pp. 565–603.

Zabel, J. E. (1993): "The Relationship between Hours of Work and Labor Force Participation in Four Models of Labor Supply Behavior," *Journal of Labor Economics*, 11(2), pp. 387–416.

Chapter 3

Identification and Estimation of Sample Selection Models Without Additivity

3.1 Introduction

Sample selection models have long been important in modeling two interrelated discrete and continuous choices made by individual agents (e.g. consumers or firms) in econometrics. It is also a leading area of application of flexible estimation methods in economics (Ichimura and Todd (2007)). In particular, to cope with possible misspecification, a number of semiparametric and nonparametric procedures that are hopefully more robust have generated considerable interest since the mid 1980s[1] (e.g. Cosslett (1991), Newey (2009), Powell (2001), and Ahn and Powell (1993), among others). More recently, by applying the results of Newey, Powell, and Vella (1999) (henceforth NPV), the fully nonparametric estimator in sample selection models (with endogenous regressors) has been suggested by Das, Newey, and Vella (2003) (hereafter DNV), which allows for both regression functions and error distributions to be unknown.

This chapter considers more general models than those mentioned above by introducing nonseparability (in the unobserved component) and endogeneity in the framework of sample selection. Imbens and Newey's (2009) nonseparable triangular system is particularly similar in concept; however, sample selection models deserve a separate treatment because they are widely used in empirical applications. In other words, we apply the nonadditive model of

[1] In the semiparametric cases, the researcher tries to relax the normality assumption while retaining the parametric functional form of the outcome equation. In such settings, the most important task is to estimate the selection correction term which is unknown since the distribution of the errors is not specified.

triangular simultaneous equations of Imbens and Newey (2009) to allowing for selection. The main contribution of this chapter is to extend the existing semi- and non-parametric estimators of sample selection models to allow for nonseparability and endogeneity. Particularly, we extend Newey's (2007) identification results of nonseparable sample selection model to allow for endogeneity and to consider nonparametric estimation (see Table 3.1 for the relation to the literature). We adopt the control function approach for identification and estimation because the control function is particularly convenient for correcting for selection and endogeneity jointly, by simply imposing conditional independence restrictions. Using the control function approach, we propose a three-step nonparametric estimator for the average structural function (ASF) given selection. The first step consists of estimating selection probability (or the propensity score) and the conditional cumulative distribution function of the endogenous variable given the instruments. In the second step, the first-step estimates are added as control variables to the outcome equation to account for selectivity and endogeneity jointly in a nonseparable setting.

Nonparametric structural econometric models have recently received much attention in the literature since they play an important role in combining economic theories with statistical models. Among them, econometric models that admit nonseparable disturbances have been a growing focus in research over the past ten years (e.g. Chesher (2003), Matzkin (2003), Chesher (2005), Chernozhukov and Hansen (2005), Hoderlein and Mammen (2007), Chernozhukov, Imbens, and Newey (2007), Altonji, Ichimura, and Otsu (2012), Rothe (2009) and Imbens and Newey (2009), Chernozhukov, Fernndez-Val, Hahn, and Newey (2013), amongst others). There are at least two important economic meanings for the nonseparable models: (1) nonseparability is likely to generate endogeneity which not only has long been important in econometrics and but is also a key feature in structural models; (2) instead of explaining the unobservable error term as the difference between Y and $\mathbb{E}(Y \mid X)$, the error term may represent individual unobserved heterogeneity, such as tastes, beliefs, ability or productivity. Taking nonseparability in the error term into account provides one way to allow causal effects or responses to policy interventions to vary across individuals with identical observed

characteristics. As for endogeneity, it is another important property that may arise in many economic models in which explanatory variables are choice variables that are related to the error. The control function approach is one of the leading approaches that have been used to deal with endogeneity in econometric models. The basic idea is that conditioning on unobservable variables can purge the dependence of observable and unobservable explanatory variables and thus the endogeneity problem disappears. It has recently been used in a variety of nonparametric and semiparametric models, including triangular nonparametric simultaneous equations models (Ng and Pinkse (1995), NPV, Pinkse (2000)), semiparametric binary response models (Blundell and Powell (2004), Rothe (2009)), semiparametric quantile regression models (Blundell and Horowitz (2007) and Lee (2007)), and nonparametric sample selection models (DNV). Imbens and Newey (2009) apply the control function approach to nonparametric triangular simultaneous equations models without imposing additivity.

We note that, unlike the conventional sample selection models imposing additive separability in the error term on the outcome equation, nonseparable sample selection models considered here only allow us to identify the *average structural function given selection* rather than the unconditional one that is of traditional interest in the sample selection literature. Whether or not the parameter given selection is of interest will depend on the context.[2] Nevertheless, there are at least two approaches to deal with the potential lack of identification of parameters for the full population due to nonseparability. The first approach is to impose additional structures on the model to secure point identification. Following this approach, one possibility is to impose functional form assumptions such as a stochastic polynomial restriction.[3] One can also appeal to the identification-at-infinity argument, as developed by Heckman (1990) and Andrews and Schafgans (1998), to achieve point identification in nonseparable sample selection models by focusing only on observations for which the choice

[2]Recovering the conditional version of parameters of interest rather than the unconditional one appears to be not satisfactory in some contexts. This is because one of the most significant innovations in microeconometrics is the ability to consistently estimate econometric models (for the full population) based on selected samples.

[3]This stochastic polynomial structure has been employed by Florens, Heckman, Meghir, and Vytlacil (2008) in the context of nonseparable models with endogeneity.

probability is sufficiently high and equal to one in the limit. We discuss these possible extensions in Section 3.4. Lastly, we note that the second response to the above nonidentification problem is to pursue partial identification. This is an ongoing research topic in the literature (e.g., Arellano and Bonhomme (2013)).

The chapter is structured as follows. Two economic examples illustrating the importance of allowing for nonadditivity in sample selection settings are given in Section 3.2. Section 3.3 presents a basic nonseparable sample selection model. Section 3.4 presents the identification analysis of the average structural function given selection in a variety of sample selection models without imposing additivity. Section 3.5 discusses possible extensions to identification strategies of parameters for the entire population such as the average and quantile structural functions. Based on identification results in Section 3.4, we suggest a resulting three-step nonparametric series estimator using a control function approach to correct for endogeneity and selectivity jointly in Sections 3.6. In Section 3.7 we derive the convergence rates of the proposed estimator. The estimator is implemented in a Monte Carlo simulation study in Section 3.8. In Section 3.9 we conclude this chapter.

3.2 Motivating Examples

3.2.1 Nonseparable Education Production Function with Endogenous Schooling

A classic economic example borrowed from Imbens (2006) with a slight modification motivates our sample selection model without additivity restrictions. Let X be schooling chosen by the individual agent and Y denote individual earnings determined by $g(X, \varepsilon)$. In this example, it may be important to correct for endogeneity and selectivity jointly and take nonseparability into account. First, selection clearly comes from the fact that earnings is observed only for those who has chosen to work. Sample selection bias may occur when such individual decisions result in a non-random sample not representing the underlying population. Second, we can think of $g(X, \varepsilon)$ as educational production function. Schooling X is endogenous in the sense that the level of education is chosen optimally by the individual

agent to maximize expected earnings $\mathbb{E}(g(X,\varepsilon) \mid \eta, Z)$ minus education cost function $c(X,Z)$ based on his information on, say, the signal of ability η and cost shifters Z. Namely

$$X(Z,\eta) = \arg\max_x \{\mathbb{E}(g(x,\varepsilon) \mid \eta, Z) - c(x,Z)\}.$$

It seems to be the case that a nonseparable educational production function leads to the decision rule of schooling $X(Z,\eta)$ taking the form of nonseparability with disturbances η as well. If, in contrast, we have the educational production function with additive errors: $Y = h(X)+\varepsilon$, then the solution for the decision rule X would be $\arg\max_x\{h(x)+\mathbb{E}(\varepsilon \mid \eta, Z) - c(x,Z)\}$. Consequently, the optimal level of X is the function of Z only and independent of η and thus independent of ε which means X is no longer endogenous. This example illustrates that endogeneity of the choice variable X is likely to be generated by nonseparability of X and ε.

3.2.2 Discrete/Continuous Consumer Demand Models with Endogenous Prices

For many different kinds of economic behavior, a decision-maker makes two choices, which consist of discrete and continuous alternatives, respectively. The mixed choice situation consumers face is the traditional focus of empirical demand analysis, especially in working with micro data on consumer demand.[4] Consider an individual who is assumed to choose a consumer durable out of a finite set of J competing alternatives and its level of consumption jointly as the solution to a utility maximization problem. The formulation of the discrete/continuous choice model begin with the random indirect utility function $V_j = V_j(p,y,x,\varepsilon)$ where V_j is defined as referring to the level of utility associated with the j^{th} alternative of the goods and is a function of price p, income y, the observed attributes x of the jth alternative, and the unobserved utility ε (including the unobserved characteristics of the product or unobserved advertising). Assuming utility maximization behavior, an

[4]For example, a consumer decides which brand of a commodity to buy and how many units to consume; in housing demand analysis, a consumer decides whether to rent or purchase his home and how large a home to live in; in the transportation economics, a household choose how many cars to own and the level of mileage traveled.

individual choosing (discrete) alternative j can be modeled by

$$d_j = \mathbf{1}\{V_j(p, y, x, \varepsilon) > V_i(p, y, x, \varepsilon) \text{ for all } i \neq j\}.$$

It is well known that price is endogenous in the choice model in the sense that it is correlated with, say, the unobserved product attributed or unobserved firms' advertising activities. Suppose that firms set prices that is determined by

$$p_j = W(Z_j, \eta_j),$$

where Z_j are exogenous variables and η_j are unobserved cost shifters that might be affected by unobserved attributes and therefore are possibly correlated with ε_j. By applying the control function approach, one can control for endogeneity of prices by using the above pricing behavior. In addition, the continuous demand for alternative j, q_j, is derived by Roy's identity

$$q_j = \frac{\partial V_j / \partial p_j}{\partial V_j / \partial y}.$$

In such a setting, since the specification of the demand model depends on the specification of the choice model and therefore the disturbances in two models are possibly correlated, sample selection bias arises. On the other hand, as nonseparable models have received increasing attention over the past decade, it has motivated researchers to treat the demand unobservables in a nonseparable manner in preferences (e.g., Gandhi, Kim, and Petrin (2012)). If the nonseparable utility function is considered, then it is likely that the quantity demanded q_j depends on the demand unobservables in a nonseparable manner as well.[5] Furthermore, not only does endogeneity of prices arise in the discrete choice model, it also carry over in the continuous demand model. Essentially, nonseparability of the utility function leads to endogeneity of prices in the demand model. Demand estimation has to account for endogeneity of prices as well as the effect of selection bias.

[5]Newey (2007) considers more general nonparametric formulation and identification of discrete/continuous choice model by allowing for the nonseparable disturbances in both indirect utility and the demand functions.

3.3 Basic Model

We consider a nonseparable model of the form

$$Y^* = g(X, \varepsilon) \tag{3.1}$$

$$Y = D \times Y^* \quad \text{observed}, \tag{3.2}$$

where Y^* is a scalar latent outcome variable affected by a vector of observed variables X and the general disturbance ε with unknown dimension, Y is the observed outcome, and D is a binary selection indicator denoting selection status and is assumed to be determined by a latent-index selection mechanism

$$D = \mathbf{1}\{q(Z) - \nu > 0\}, \tag{3.3}$$

where Z is a vector of variables that affects the probability of selection; X and Z may have common variables but not all the same, i.e., exclusion restrictions; ν is an error term that could be correlated with ε. The sample selection problem arises when Y^* is only observed on a nonrandomly selected sample. That is, D and ε are possibly correlated due to potential correlation between ν and ε. Note that (3.1) allows nonlinear effects of X on the outcome Y to permit random variation due to nonseparability in ε.

3.4 Identification Analysis

3.4.1 Nonidentification Results Without Additivity

To illustrate nonidentification of the parameters for the full population without additive separability, consider a nonparametric sample selection model of the form $Y^* = g(X) + \varepsilon$ and $Y = D \times Y^*$ observed with the following stochastic restriction as a key identifying assumption (implied by, for example, the full independence of (ε, ν) and (X, Z))[6]

$$\mathbb{E}(Y \mid X, Z, D = 1) = g(X) + \mathbb{E}(\varepsilon \mid X, Z, D = 1),$$

[6]A weaker sufficient condition is that the joint distribution of (ε, ν) depends on Z only through the single index $q(Z)$ or the propensity score $P(Z)$.

$$\mathbb{E}(\varepsilon \mid X, Z, D = 1) = \mathbb{E}(\varepsilon \mid X, P(Z)) = \mathbb{E}(\varepsilon \mid P(Z)) \equiv \lambda(P(Z)), \tag{3.4}$$

where the first equality follows from the assumption that ε depends on Z only through the propensity score $P(Z)$ and the second equality means ε is mean independent of X conditional of $P(Z)$ so therefore the propensity score $P(Z)$ serves as a control variable. Then the quantity $\mathbb{E}(Y \mid X, Z, D = 1)$ can be expressed as $g(X) + \mathbb{E}(\varepsilon \mid X, Z, D = 1) = g(X) + \lambda(P) \equiv h(W)$ and $W \equiv (X, P)$.

However, in more general models in which $g(X, \varepsilon)$ is nonadditive in ε, the stochastic restriction (3.4) is in general not sufficient for identification. In other words, without further assumptions, the nonseparable sample selection model doesn't imply the relation between the functional of $h(W)$ and the objects of interest for the full population.

To be precise, in the nonadditive models given by $Y = g(X, \varepsilon)$, it is well known that the structural function $g(X, \varepsilon)$ and the joint distribution $F_{X,\varepsilon}$ cannot be identified simultaneously even when ε is distributed independently of X and g is strictly increasing in ε (see Matzkin (2003)). To deal with this issue, on the one hand, one can impose additional restrictions on g or use a normalization specifying ε to be $U(0, 1)$. On the other hand, without imposing additional restrictions or normalizations, an alternative approach is to shift focus of the object of interest on the *average structural function* suggested by Blundell and Powell (2003), which is an alternative summary version of the structural function g. In particular, the ASF is represented by

$$\text{ASF}(x) \equiv \int g(x, \varepsilon) dF_{\varepsilon}.$$

As a special case, the average structural function in the additive models of the form $Y = h(X) + \varepsilon$ is given by

$$\text{ASF}(x) = \int (h(x) + \varepsilon) dF_{\varepsilon} = h(x) + \mathbb{E}(\varepsilon \mid X).$$

Therefore, the average structural function in an additive model reduces to the usual regression function $h(x) = \mathbb{E}(Y \mid X)$ if $\mathbb{E}(\varepsilon \mid X) = 0$. In fact, the usual goal of the estimation approach to the sample selection models is to draw inferences on the parameter of interest for the full population, even if we can only observe the nonrandomly selected sample. As a

consequence, it seems natural that the (unconditional) ASF is our potential candidate of the parameter of interest. Unfortunately, unlike the separable sample selection models, where β or $h(x)$, the parameter of interest for the full population, can be recovered from the identified object $\mathbb{E}(Y \mid X, Z, D = 1)$, the (unconditional) ASF cannot be identified (and estimated) from the selected sample only (see Table 3.2). In other words, in the context of sample selection models, the object the researcher is able to recover from the data is $\mathbb{E}(Y \mid X, Z, D = 1)$. Imposing separability on sample selection models implies that $\mathbb{E}(Y \mid X, Z)$, the parameter of interest for the full population, can be separated from $\widehat{\mathbb{E}}(Y \mid X, Z, D = 1)$. However, without further restrictions as discussed below, we cannot do so in a more general nonseparable setting.

We start with the full independence assumption.

Assumption FI. (ε, ν) *and* (X, Z) *are independent.*

With the full independence assumption, we can define the propensity score P referring to the conditional probability of selection $D = 1$ given Z. The propensity score plays an important role of controlling for selection bias in the nonparametric selection framework in our identification analysis.

Assumption PS. *Suppose Assumption FI is satisfied and the error term ν is continuously distributed with support on the entire real line with the distribution function $F_\nu(\cdot)$. Then the selection probability (which is often referred to as the "propensity score") is*

$$P = Pr(D = 1 \mid Z) = Pr(\nu < q(Z)) = F_\nu(q(Z)).$$

Under Assumption FI and the definition of the propensity score, Newey (2007) shows that ε and X are independent conditional on P in the selected data, i.e.,

$$(\varepsilon, \nu) \perp\!\!\!\perp (X, Z) \Rightarrow \varepsilon \perp\!\!\!\perp X \mid P(Z), D = 1.$$

On the other hand, under Assumptions FI and PS, we recall a known result that has commonly been used in semi- and non-parametric sample selection models (e.g., Ahn and

Powell (1993), Chen and Khan (2003)): consider the conditional distribution of the outcome errors ε given $D = 1$ and the regressors X and Z

$$\begin{aligned}
\Pr(\varepsilon \le c \mid X, Z, D = 1) &= \Pr(\varepsilon \le c \mid X, Z, \nu < q(Z)) \\
&= \Pr(\varepsilon \le c \mid q(Z), \nu < q(Z)) \\
&= H(c, q(Z)) \\
&= H(c, F_\nu^{-1}(P(Z))) \\
&= G(c, P(Z)),
\end{aligned}$$

where the first and second equalities follow from the selection equation (3.3) and Assumption FI and the fourth equality is due to Assumption PS implying an invertible relation between $q(Z)$ and the propensity score $P(Z)$. This result implies that any functional of the conditional distribution of ε is only a function of the propensity score.

Assumption FI has often been used to be a sufficient condition to identify the additive sample selection models. However, it will still not suffice in general for identification of the ASF in the nonadditive sample selection models. To see this,

$$\begin{aligned}
\mathbb{E}(Y \mid X = x, Z = z, D = 1) &= \mathbb{E}(g(x, \varepsilon) \mid X = x, Z = z, \nu < q(z)) \\
&= \mathbb{E}(g(x, \varepsilon) \mid X = x, q(z), \nu < q(z)) \\
&= \mathbb{E}(g(x, \varepsilon) \mid X = x, q(z), D = 1) \\
&= \mathbb{E}(g(x, \varepsilon) \mid X = x, P(z), D = 1) \\
&\equiv \int g(x, \varepsilon) dF_{\varepsilon \mid X = x, P(z), D = 1} \\
&= \int g(x, \varepsilon) dF_{\varepsilon \mid P(z), D = 1} \\
&\neq \int g(x, \varepsilon) dF_{\varepsilon \mid P(z)}
\end{aligned}$$

unless the outcome errors ε are independent of the selection errors ν conditional on the propensity score. This inspection implies nonidentifibility of $\mathrm{ASF}(x)$ since identification of $\mathrm{ASF}(x)$ requires $\mathbb{E}(g(x, \varepsilon) \mid X, P)$ be identified. Conversely, $\mathrm{ASF}(x)$ is recovered by

integrating the quantity $\int g(x,\varepsilon)dF_{\varepsilon|P}$ over the marginal distribution of the propensity score P, i.e.,

$$\int \int g(x,\varepsilon)dF_{\varepsilon|P}dF_P = \int g(x,\varepsilon)dF_\varepsilon \equiv \text{ASF}(x).^7$$

The loss of identifying power of stochastic restrictions (3.4) in the context of nonseparable sample selection model is not surprising because the mean restrictions typically have to be strengthened to the full independence in general nonseparable models. A more interesting question may be to explore the identifying power of the full independence restriction in nonseparable models with sample selection. We maintain the full independence restriction on the model in the following identification analysis.

In response to nonidentification results in nonseparable sample selection models, there are at least three approaches to deal with the lack of identification of parameters for the full population due to nonseparability. First, one can shift the focus of attention on the parameter that is identified given the selected data, e.g., the ASF given selection. This approach is the main focus of this chapter. We defer the discussion of this approach to Section 3.4.2. The second approach is to impose additional structures on the model to secure point identification. Following this approach, one possibility is to impose functional form assumptions such as a stochastic polynomial structure.[8] One can also appeal to the identification-at-infinity argument, as developed by Heckman (1990) and Andrews and Schafgans (1998), to achieve point identification in nonseparable sample selection models by focusing only on observations for which the choice probability is sufficiently high and equal to one in the limit. We discuss these possible extensions in Sections 3.5.1 and 3.5.2, respectively. Lastly, we note that the third approach in response to the above nonidentification problem is to pursue partial identification. This is an ongoing research topic in the literature (Arellano and Bonhomme (2013)).

[7]For the outer integral before the first equality to be well-defined, we need to impose the common support assumption that the conditional support of P given X and the marginal support of P conincides (and vise versa).

[8]This stochastic polynomial structure has been employed by Florens, Heckman, Meghir, and Vytlacil (2008) in the context of nonseparable models with endogeneity.

3.4.2 Identification of $\text{ASF}^S(x)$

In this section, we study the identification of the average structural function given selection based on the first approach as mentioned in the previous section. Corresponding to the nonparametric separable sample selection models developed by DNV, we move from a simplest nonseparable sample selection model to progressively more complex models, including a nonseparable sample selection model with a continuous endogenous regressor and a model where the endogenous variable enters the selection equation as well. The parameters of interest, identifying assumptions, and the identification results are also discussed.

3.4.2.1 The Simplest Nonseparable Sample Selection Model

The simplest nonseparable sample selection models (3.1)-(3.3), relaxing the functional forms of the outcome equation and the distribution of the error term and assuming that the observables and the error term are independent, have briefly been considered by Newey (2007). We repeat his results for completeness and provide a detailed discussion in order to view them as basic results on which we can make extensions that might be more appropriate for empirical applications, as discussed below.

The model (3.1)-(3.3) is a general form of sample selection models. Their parametric, semiparametric and nonparametric counterparts in separable settings are special cases of this general model. We use the following two examples to illustrate the relationship among these models.

Example 3.1. Linear Parametric Sample Selection Models with Normality Assumptions
The conventional sample selection models in a fully linear parametric setting are given by

$$Y^* = X'\beta + \varepsilon,$$
$$D = \mathbf{1}\{Z'\alpha - \nu > 0\},$$
$$\text{and } Y = D \times Y^* \text{observed}.$$

Together with assuming that (ε, ν) are jointly normally distributed with full independence of Z, we then have

$$\begin{aligned}\mathbb{E}[\varepsilon \mid Z, D = 1] &= \mathbb{E}[\varepsilon \mid Z, \nu < Z'\alpha] = \rho_{\varepsilon\nu}\mathbb{E}[\nu \mid Z, \nu < Z'\alpha] \\ &= -\sigma_{\varepsilon\nu}\phi(Z'\alpha)/\Phi(Z'\alpha),\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density density and distribution functions of the standard normal distribution.

Alternatively, by introducing the propensity score, $\mathbb{E}[\varepsilon \mid Z, D = 1]$ can also be expressed as the function of the propensity score

$$P(Z'\alpha) \equiv \Pr(D = 1 \mid Z) = \Pr(\nu < Z'\alpha) = F_\nu(Z'\alpha),$$

which implies

$$Z'\alpha = F_\nu^{-1}(P(Z'\alpha))$$

under the assumption of strict monotonicity of F_ν. Therefore

$$\mathbb{E}[\varepsilon \mid Z, D = 1] = \mathbb{E}[\varepsilon \mid \nu < Z'\alpha] = \mathbb{E}[\varepsilon \mid \nu < F_\nu^{-1}(P(Z'\alpha))] = \lambda_0(P).$$

We call $\mathbb{E}[\varepsilon \mid Z, D = 1]$ the control function or the selection correction term because entering it as an additional regressor in the outcome equation corrects for selection bias. In a linear setting with normality assumption of (ε, ν) as shown above, the control function $\lambda_0(P)$ becomes a function of the inverse Mills ratio.

Example 3.2. Semi- and Non-parametric Sample Selection Models

In order to avoid the potential misspecification resulting from the assumption of error distribution, by relaxing the normality assumption, we then have the semiparametric sample selection model

$$\mathbb{E}[Y \mid X, Z, D = 1] = X'\beta + \lambda_1(Z'\alpha),$$

where the control function $\lambda_1(Z'\alpha)$ or $\lambda_0(P)$ is unknown since F_ν is not specified parametrically. In semiparametric settings, the most important task is the estimation of $\lambda_0(P)$ which

is unknown. A number of estimators have previously been developed in the literature to deal with this problem, such as Cosslett's (1991) dummy variable model and Newey's (2009) series approximation etc. In addition, Powell (2001) and Ahn and Powell (1993) develop an alternate approach by "pairwise differencing" observations with the same probability of selection to sidestep the estimation of the unknown $\lambda_0(P)$. More recently, DNV extends the parametric and semiparametric sample selection models to nonparametric settings with separable additivity restrictions. In particular, the model they consider is

$$\mathbb{E}[Y \mid X, Z, D = 1] = h(X) + \lambda_0(P),$$

where $h(\cdot)$ and $\lambda_0(\cdot)$ are unknown functions.

As discussed in the previous section, the object of interest for the full population such as the (unconditional) ASF can not be separated from the estimate of $\mathbb{E}(Y \mid X, Z, D = 1)$. As a consequence, instead of the ASF, one identifiable parameter of primary interest in the nonseparable sample selection models is the ASF given selection:[9]

$$\text{ASF}^S(x) \equiv \int g(x, \varepsilon) dF_\varepsilon^S.$$

For notational simplicity, let the distribution function and expectation with superscript S denote the corresponding objects in the selected sample, i.e., $F_\varepsilon^S(\cdot) \equiv F_{\varepsilon|D=1}(\cdot)$.

We would like to emphasize that whether or not the conditional average structural function on selection (i.e., $\text{ASF}^S(x)$) is of interest depends on the context. We may be more interested in $\text{ASF}^S(x)$ than in $\text{ASF}(x)$ in some economic applications where there are a lot of units never being selected. This is because the former measures the effects on those for whom the program is actually intended and hence is more policy relevant. For example, in a study of empirically assessing the wage effects of the Job Corps program, a job training program in the U.S., Lee (2009) mentions that it is useful to assess the impacts on wages of the job training program *conditional on being employed*. In contrast, researchers may be

[9]Newey (2007) shows that, in addition to the ASF given selection, the derivative of the ASF conditional on selection and the quantile structural function for the selected sample can also be identified in a nonseparable sample selection model.

interested in $\text{ASF}(x)$ in the case where the effects for the public in general is of primary interests, not just for people who actually participated in treatment, e.g. the effects of exercise on blood pressure are concerns for the public, not just for people who exercise. Nevertheless, (point and partial) identification of the unconditional average structural function or other parameters of interest for the full population in a nonseparable sample selection model remains an important direction for future research.

We want to show below that averaging the conditional mean of Y given X, P, and $D = 1$ over the distribution of P in the selected data gives the $\text{ASF}^S(x)$ for the nonseparable sample selection models. By identification we mean that parameters of interest can be recovered from the distributions of the observable random variables. It is assumed that we have i.i.d. data on $\{Y_i, X_i, Z_i, D_i\}$ for $i = 1, \ldots, n$, which ensures that the joint distribution of (Y, X, Z, D) is by definition identified and conditional distributions and moments can also be consistently estimated.

Under Assumption FI and the definition of the propensity score, Newey (2007) shows that ε and X are independent conditional on the propensity score P in the selected data, i.e.,

$$(\varepsilon, \nu) \perp\!\!\!\perp (X, Z) \implies \varepsilon \perp\!\!\!\perp X \mid P(Z), D = 1. \tag{3.5}$$

Based on the result of conditional independence (3.5), Newey (2007) further applies the known results of identification of nonseparable models using the control function approach in Imbens and Newey (2009) to the nonseparable sample selection model and shows that $\text{ASF}^S(x)$ and the quantile structural function given selection are identified. Newey also points out that the average derivative given selection is identified without the common support assumption.

Assumption CS. *For all $X \in \mathcal{X}$, the support of P conditional on X and selection equals the support of P conditional on selection.*

Theorem 3.1. *(Newey (2007)) In the model of equations* (3.1)-(3.3)*, if Assumption FI holds, then X and ε are independent conditional on P in the selected data. If additionally Assumption CS is satisfied, then the average structural function given selection is identified.*

The identification strategy proceeds as follows. First note that for any bounded function $a(X)$, by independence of X and ε,

$$\begin{aligned}\mathbb{E}^S[a(X) \mid \varepsilon, P] &= \int a(X) dF^S_{X|\varepsilon,P}(x \mid \varepsilon, P) \\ &= \int a(X) dF^S_{X|P}(x \mid P) \\ &= \mathbb{E}^S[a(X) \mid P].\end{aligned}$$

Therefore, for any bounded function $b(\varepsilon)$ we have

$$\begin{aligned}\mathbb{E}^S[a(X)b(\varepsilon) \mid P] &= \mathbb{E}^S[\mathbb{E}^S(a(X)b(\varepsilon) \mid P, \varepsilon) \mid P] \\ &= \mathbb{E}^S[b(\varepsilon)\mathbb{E}^S(a(X) \mid P, \varepsilon) \mid P] \\ &= \mathbb{E}^S[b(\varepsilon)\mathbb{E}^S(a(X) \mid P) \mid P] \\ &= \mathbb{E}^S[a(X) \mid P]\mathbb{E}^S[b(\varepsilon) \mid P].\end{aligned}$$

The first equality follows from the law of iterated expectations. Next consider

$$\begin{aligned}\mathbb{E}^S[Y \mid X = x, p] &= \mathbb{E}^S[g(x, \varepsilon) \mid X = x, P] \\ &= \int g(x, \varepsilon) dF^S_{\varepsilon|X=x,P} \\ &= \int g(x, \varepsilon) dF^S_{\varepsilon|P} \\ &\equiv g^*(x, P, D = 1).\end{aligned}$$

Due to the assumption of common support of P, integrating $g^*(x, P, D = 1)$ over the marginal distribution of P conditional on selection is well defined and thus gives $\text{ASF}^S(x)$ as follows

$$\begin{aligned}\int g^*(x, P, D = 1) dF^S_P &= \int \int g(x, \varepsilon) dF^S_{\varepsilon|P} dF^S_P \\ &= \int g(x, \varepsilon) dF^S_\varepsilon \\ &\equiv \text{ASF}^S(x).\end{aligned}$$

We then conclude that $\text{ASF}^S(x)$ is identified since both $\mathbb{E}^S[Y \mid X = x, P]$ and F_P^S are identifiable objects.

We emphasize here that there are two implications from the result of Theorem 3.1. First of all, it is worth noting that $g^*(x, P, D = 1)$ can be regarded as a nonseparable generalization of the control function for the additive models. In an additive model, $g^*(x, P, D = 1)$ turns out to be

$$\begin{aligned} g^*(x, P, D = 1) &\equiv \int (h(x) + \varepsilon) dF^S_{\varepsilon|P} \\ &= \int (h(x) + \varepsilon) dF^S_{\varepsilon|X,P} \\ &= h(x) + \mathbb{E}(\varepsilon \mid X, P, D = 1) \\ &= h(x) + \lambda_0(P), \end{aligned}$$

where $\mathbb{E}(\varepsilon \mid X, P, D = 1) \equiv \lambda_0(P)$ is a control function, as mentioned above. Second, by using the following equality derived from Theorem 3.1, we have

$$\int g^*(x, P, D = 1) dF_P^S = \int g(x, \varepsilon) dF_\varepsilon^S.$$

we can see that identification of the object of interest, $\text{ASF}^S(x)$, can be achieved by replacing the unidentified objects $g(x, \varepsilon)$ and F_ε^S with $g^*(x, P, D = 1)$ and F_P^S respectively.

3.4.2.2 Nonseparable Sample Selection Models with a Continuous Endogenous Regressor

Since it is quite often to consider the econometric models with endogenous variables, we extend the simplest nonseparable sample selection model discussed in the previous section to allow for endogenous regressors. It is the model that we propose the nonparametric estimation approach, derive asymptotic theory, and implement simulations in the following

sections. The nonseparable sample selection model with endogenous regressors is given by

$$Y^* = g(X, \varepsilon), \tag{3.6}$$

$$X_1 = \pi(Z, \eta), \tag{3.7}$$

$$Y = D \times Y^* \quad \text{observed}, \tag{3.8}$$

$$D = \mathbf{1}\{q(Z) - \nu > 0\}, \tag{3.9}$$

where there is a single endogenous variable X_1 included in $X = (X_1, Z_1')'$ and instruments $Z = (Z_1', Z_2')'$; ε can be a vector or a scalar; η is a scalar; there are potential correlations among ε, η, and ν.

Example 3.3. DNV's Nonparametric Sample Selection Models with Endogenous Regressors The model (3.6)-(3.9) is a general form of DNV's nonparametric sample selection model with endogeneity. In DNV's model where $Y^* = g(X_1, Z_1) + \varepsilon$, $X_1 = \pi(Z_1, Z_2) + \eta$, and $Y = d \times Y^*$ observed, the control function $\lambda_2(\cdot)$ takes the following form

$$\mathbb{E}[Y \mid X, Z, \eta, D = 1] = h(X) + \mathbb{E}[\varepsilon \mid Z, \eta, D = 1] = h(X) + \lambda_2(P, \eta).$$

Assumption 3.1. *(i)(Full independence)* (ε, η, ν) *and* Z *are independent;*
(ii)(Common support) For all $X \in \mathcal{X}$*, the support of* (P, η) *conditional on* X *and selection equals the support of* (P, η) *conditional on selection;*
(iii)(Strict monotonicity in η*) If* $\pi(z, \eta) > \pi(z, \eta')$ *for some* (z, η, η')*, then* $\pi(z', \eta) > \pi(z', \eta')$ *for all* z'*.*

The Assumption 3.1 (i) and (ii) are similar to the corresponding assumptions in the previous model. Assumption 3.1 (iii) is needed because imposing the strict monotonicity in η enables us to introduce a control variable $V \equiv F_{X_1|Z}(X_1 \mid Z)$ in place of η. The reason for that will become clear below. However, this is indeed a strong restriction, which implies that the nature of endogenous regressors has to be continuous. This is because for the case with a binary endogenous variable X_1, imposing strict monotonicity in η forces η to take on two distinct values only and therefore η is perfectly correlated with X_1, meaning that there

is no variation in X_1 conditional on η and hence we learn nothing about the causal effect of X_1. Another reason for ruling out the case of binary endogenous variables is that the control variable V, as defined above, is legitimate only for a continuous endogenous variable.

Theorem 3.2. *In the model of equations* (3.6)-(3.9)*, if Assumptions 3.1 (i) are satisfied, then X and ε are independent conditional on p and η given selection.*

Proof. The proof is similar to that of Theorem 1. For any bounded function $a(X)$, by independence of Z and (ε, η)

$$\begin{aligned}
\mathbb{E}^S[a(X) \mid \varepsilon, P, \eta] &= \int a(\pi(Z, \eta)) dF^S_{Z|\varepsilon,P,\eta}(z \mid \varepsilon, P, \eta) \\
&= \int a(\pi(Z, \eta)) dF^S_{Z|P}(z \mid p) \\
&= \mathbb{E}^S[a(X) \mid P, \eta].
\end{aligned}$$

Therefore, for any bounded function $b(\varepsilon)$, by the law of iterated expectations, we have

$$\begin{aligned}
\mathbb{E}^S[a(X)b(\varepsilon) \mid P, \eta] &= \mathbb{E}^S[\mathbb{E}^S(a(X)b(\varepsilon) \mid \varepsilon, P, \eta) \mid P, \eta] \\
&= \mathbb{E}^S[b(\varepsilon)\mathbb{E}^S(a(X) \mid \varepsilon, P, \eta) \mid P, \eta] \\
&= \mathbb{E}^S[b(\varepsilon)\mathbb{E}^S(a(X) \mid P, \eta) \mid P, \eta] \\
&= \mathbb{E}^S[a(X) \mid P, \eta]\mathbb{E}^S[b(\varepsilon) \mid P, \eta].
\end{aligned}$$

□

Following the similar argument to Theorem 3.1, we can show below that averaging the conditional mean of Y given X, P, η, and $D = 1$ over the joint distribution of (P, η) in the selected data gives the ASF given selection for the model (3.1)-(3.3).

$$\begin{aligned}
\mathbb{E}^S[Y \mid X = x, P, \eta] &= \mathbb{E}^S[g(X, \varepsilon) \mid X = x, P, \eta] \\
&= \int g(x, \varepsilon) dF^S_{\varepsilon|X=x,P,\eta} \\
&= \int g(x, \varepsilon) dF^S_{\varepsilon|P,\eta} \\
&\equiv g^*(x, P, \eta, D = 1).
\end{aligned}$$

Integrating $g^*(x, P, \eta, D = 1)$ over the conditional distribution of P given selection and the marginal distribution of η conditional on selection, we obtain $\text{ASF}^S(x)$ as follows

$$\begin{aligned}\int\int g^*(x,P,\eta,D=1)dF^S_{P|\eta}dF^S_\eta &= \int\int\int g(x,\varepsilon)dF^S_{\varepsilon|P,\eta}dF^S_{P|\eta}dF^S_\eta \\ &= \int g(x,\varepsilon)dF^S_\varepsilon \\ &\equiv \text{ASF}^S(x).\end{aligned}$$

Since

$$\int\int g^*(x,P,\eta,D=1)dF^S_{P|\eta}dF^S_\eta = \int \mathbb{E}^S[Y \mid X=x,P,\eta]dF^S_{P,\eta},$$

we obtain the similar results indicating that $\text{ASF}^S(x)$ is the integral over the joint distribution of (P,η) conditional on selection of $\mathbb{E}^S[Y \mid X = x, P, \eta]$. Again, the above integral on the right side is well defined due to common support in Assumption 3.1 (ii).

In an additive model where $X_1 = \pi(Z) + \eta$ with $\mathbb{E}(\eta \mid Z) = 0$, the function π as well as $F_{\eta,Z}$ can be identified (see Matzkin (2007) for proofs). As a consequence, η is identified and consistently estimated since $\eta = X_1 - \pi(Z)$. However, the problem of identification arises due to nonseparability of the reduced form for X_1. Since the reduced form equation for X_1 has an unknown function $\pi(\cdot)$ with an nonadditive disturbance η, it leads to unidentifiability of $\pi(Z,\eta)$ even when π is assumed to be strictly increasingly in η and η is distributed independently of Z (see Matzkin (2003)). Under Assumption 3.1 (iii), there is a trick Imbens and Newey (2009) employ to construct an alternative variable V in place of η as a control variable by using the following argument

$$\begin{aligned}F_\eta(\bar\eta) &= \Pr(\eta \le \bar\eta \mid Z=z) = \Pr(X_1 \le \pi(Z,\bar\eta) \mid Z=z) = \Pr(X_1 \le x_1 \mid Z=z) \\ &= F_{X_1|Z}(x_1 \mid z) = \mathbb{E}[\mathbf{1}\{X_1 \le x_1\} \mid Z=z] \equiv v(x_1,z).\end{aligned}$$

Based on the fact that V is a one-to-one function of η, we then obtain, by replacing η with V and following the same argument above, the immediate results of independence of X and ε conditional on P, V, and $D = 1$ and then identification of the ASF given selection, as stated in Theorem 3.3.

Theorem 3.3. *In the model of equations* (3.6)-(3.9)*, if Assumptions 3.1 are satisfied, then* X *and* ε *are independent conditional on* P*,* V*, and selection and the average structural function given selection is identified.*

Theorem 3.3 states that V in place of η, along with propensity score, can be used as control variables to correct for endogeneity and selection bias jointly. It extends previous results given by Newey (2007) and Imbens and Newey (2009) to the cases where both selection and endogeneity are taken into consideration.

3.4.2.3 Nonseparable Sample Selection Models with an Endogenous Variable Entering the Selection Equation

One possible extension of the models considered above is to the case in which not only does the endogenous variable enters the outcome equation but also enter the selection equation. This model, for example, can be considered as a labor supply model where X_1 are an endogenous schooling or continuous treatment variable that affects wages (Y) and the participation decision at the same time.[10] This extended model is given by

$$Y^* = g(X, \varepsilon), \tag{3.10}$$

$$X_1 = \pi(Z, \eta), \tag{3.11}$$

$$Y = D \times Y^* \quad \text{observed}, \tag{3.12}$$

$$D = \mathbf{1}\{q(X_1, Z) - \nu > 0\}. \tag{3.13}$$

[10] An interesting question is to empirically assess whether any earning gains from participation of employment-related training programs are achieved through raising individual's wage rate by an increase of human capital or simply through increasing the likelihood of employment and hours of work without any increase in wage rates.

Example 3.4. The control function in the parametric linear model corresponding to the above general model can be expressed by

$$\begin{aligned}\mathbb{E}[\varepsilon \mid X_1, Z, D = 1] =& \mathbb{E}[\varepsilon \mid X_1, Z, \nu < (x_1\delta + Z'\alpha)] = \mathbb{E}[\varepsilon \mid X_1, Z, \nu < (Z'\gamma + \eta)\delta + Z'\alpha)] \\ =& \mathbb{E}[\varepsilon \mid X_1, Z, \nu - \eta\delta < Z'(\gamma\delta + \alpha)] \\ =& \sigma\mathbb{E}[\nu - \eta\delta \mid \nu - \eta\delta < Z'\theta] \\ =& -\sigma\phi(Z'\theta)/\Phi(Z'\theta) = \rho\lambda_3(Z'\theta),\end{aligned}$$

where σ is the covariance between ε and $\nu - \eta\delta$; $\theta = \gamma\delta + \alpha$; $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal density and distribution functions for $\nu - \eta\delta$.

Theorem 3.4. *In the model of equations* (3.10)-(3.13)*, if Assumption 3.1 (i) is satisfied, then X_1 and ν are independent conditional on v and X and ε are independent conditional on P, V, and selection.*

Proof. For any bounded function $a(X_1)$, by independence of Z and (ε, η)

$$\mathbb{E}^S[a(X_1) \mid \nu, V] = \int a(\pi(Z, \eta))dF^S_{Z|\nu,V}(z \mid \nu, v) = \int a(\pi(Z, \eta))dF^S_{Z|V}(z \mid v) = \mathbb{E}^S[a(X_1) \mid V].$$

Therefore, for any bounded function $b(\nu)$, by iterated expectations, we have

$$\begin{aligned}\mathbb{E}^S[a(X_1)b(\nu) \mid V] &= \mathbb{E}^S[\mathbb{E}^S(a(X_1)b(\nu) \mid \nu, V) \mid V] = \mathbb{E}^S[b(\nu)\mathbb{E}^S(a(X_1) \mid \nu, V) \mid V] \\ &= \mathbb{E}^S[b(\nu)\mathbb{E}^S(a(X_1) \mid V) \mid V] = \mathbb{E}^S[a(X_1) \mid V]\mathbb{E}^S[b(\nu) \mid V].\end{aligned}$$

The proof of the second part is the same as that of Theorem 3.2. □

Theorem 3.5. *In the model of equations* (3.10)-(3.13)*, if Assumption 3.1 is satisfied, then the average structural function given selection is identified.*

Proof. The proof is the same as that of Theore 3.3. □

Another interesting extension of the nonseparable sample selection model is that not only does the dependent variable in the outcome equation be only observed in the selected

sample but also endogenous variables subject to selection. An economic application of such a model could be a labor supply model where Y is hours worked and X_1 are endogenous wages and both of them are only observed when people choose to work. This extension can work in an separable setting since we can recover $\pi(Z)$ (and thus η) from $\mathbb{E}^S(X_1 \mid Z, P)$ since $\mathbb{E}^S(X_1 \mid Z, P) = \pi(Z) + \lambda(P)$ (see DNV for a detailed discussion). Yet this would be a challenging extension in a setting of nonseparable reduced forms for X_1. The reason for that difficulty is that the object we can identify and estimate is $\mathbb{E}^S[\mathbf{1}\{X_1 \leq x_1\} \mid Z = z]$, which is the conditional expectation of $\mathbf{1}\{X_1 \leq x_1\}$ *on selection*, rather than $\mathbb{E}[\mathbf{1}\{X_1 \leq x_1\} \mid Z = z]$ *for the full population.*

3.5 Extension

3.5.1 Identification of ASF(x) in Partially Separable Models

Nonseparable models are motivated by the fact that additive separability is hard to justify from economic theory or empirical evidence. However, in the context of sample selection, as discussed in the previous section, fully nonseparable sample selection models are in general not able to achieve point identification of $\text{ASF}(x)$ without further restrictions. It is natural to ask if point identification can be achieved by restricting the class of underlying models. Here we provide a special case that can be viewed as the middle ground between fully separable and fully nonseparable models. In particular, we characterize a class of partially separable models for which it is possible to (point) identify the (unconditional) ASF.

Consider a structural nonseparable outcome equation of the following partially separable form

$$\begin{aligned} Y^* &= g(X, \varepsilon) \\ &= g_1(X) + g_2(X, \varepsilon) \\ &= g_1(X) + \sum_{j=0}^{J} g_{2j}(X)\varepsilon_j \end{aligned} \tag{3.14}$$

with $Y = Y^* \times D$ observed, where X is a scalar, $g_{2j}(X) : \mathbb{R} \to \mathbb{R}, j = 0, \ldots, J$ are possibly unknown functions, and J is known. Without loss of generality, we can normalize $\mathbb{E}(\varepsilon_j) = 0$,

$j = 0, 1, \ldots, J$. This specification of the outcome equation can be viewed as a partially separable model with nonseparability in ε maintained. By imposing this particular specification, it is easy to see that $\text{ASF}(x)$ can be expressed as

$$\text{ASF}(x) \equiv \int g(x, \varepsilon) dF_\varepsilon = g_1(x) + \sum_{j=0}^{J} g_{2j}(x) \mathbb{E}(\varepsilon_j) = g_1(x).$$

We begin the identification analysis of $\text{ASF}(x)$ by Assumption FI. Assumption FI ensures that $(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_J)$ are independent of X conditional on the propensity score in the selected data, which implies that $\mathbb{E}(\varepsilon_j \mid X, Z, D = 1)$ for $j = 0, 1, \ldots, J$ is only a function of the propensity score.

As a consequence, the partially separable specification enables us to express the identifiable object $\mathbb{E}(Y \mid X = x, Z = z, D = 1)$ as an additive function consisting of the parameter of interest, $\text{ASF}(x)$, and the selection correction term, i.e.,

$$\begin{aligned}
\mathbb{E}(Y \mid X = x, Z = z, D = 1) &= \text{ASF}(x) + \sum_{j=0}^{J} g_{2j}(x) \mathbb{E}(\varepsilon_j \mid X = x, Z = z, D = 1) \\
&= \text{ASF}(x) + \sum_{j=0}^{J} g_{2j}(x) \mathbb{E}(\varepsilon_j \mid X = x, q(z), \nu < q(z)) \\
&= \text{ASF}(x) + \sum_{j=0}^{J} g_{2j}(x) \mathbb{E}(\varepsilon_j \mid X = x, P(z), \nu < F_\nu^{-1}(P(q(z)))) \\
&= \text{ASF}(x) + \sum_{j=0}^{J} g_{2j}(x) \mathbb{E}(\varepsilon_j \mid P(z), \nu < F_\nu^{-1}(P(q(z)))) \\
&\equiv \text{ASF}(x) + \sum_{j=0}^{J} g_{2j}(x) \lambda_j(P(z)), \qquad (3.15)
\end{aligned}$$

where $\lambda_j(\cdot)$ for $j = 0, 1, \ldots, J$ are unknown functions of the propensity score. Note that the selection correction is now a function of the covariate X and the propensity score. The selection correction involving the covariate X, without further restrictions, is not able to be expressed as a function of the propensity score alone and therefore fails identification of $\text{ASF}(x)$. However, the stochastic polynomial specification and the smoothness assumption, as discussed below, play a crucial role for the identification.

Assumption SP. *(i) (Stochastic Polynomial Specification) The additive functions $g_{2j}(X), j = 0, 1, \ldots, J$ take a stochastic polynomial form, i.e., $g_{2j}(X) = X^j$.*

(ii) (Smoothness) $g_1(X)$ *is* J *times differentiable in* X*.*
(iii) (Measurable Separability) Any function of X *almost surely equal to a function of the propensity score* $P(Z)$ *must be almost surely equal to a constant.*

The stochastic polynomial specification, along with the smoothness assumption, enable us to eliminate the covariate involved in the selection correction by taking derivative enough times. Similar to the identifying strategy suggested by Florens, Heckman, Meghir, and Vytlacil (2008), the following theorem shows that $\text{ASF}(x)$ is identified.

Theorem 3.6. *Suppose Assumptions FI and SP hold in models* (3.1)-(3.3) *and* (3.14)*. The average structural function is identified.*

Proof. The proof is analogous to that of Theorem 1 in Florens, Heckman, Meghir, and Vytlacil (2008). We omit the proof for brevity. □

Remark 3.1. This argument extends the identification results by Florens, Heckman, Meghir, and Vytlacil (2008) to the sample selection models. Note that for nonseparable models with endogenous regressors using the control function approach, imposing a stochastic polynomial assumption serves as an alternative method to obtain identification without the need to impose the common support assumption employed by Imbens and Newey (2009). Theorems 3.6 and 3.1 below show that these two approaches identify conditional and unconditional ASFs respectively in the nonseparable models in the presence of sample selection.

The stochastic polynomial specification as equation (3.14) encompasses heteroskedastic sample selection models as a special case where the outcome equation is of the form

$$Y^* = g_1(X) + g_{20}(X)\varepsilon_0. \tag{3.16}$$

It is well known that conventional estimation procedures for sample selection models yield inconsistent estimators in the presence of conditional heteroskedasticity. It is not difficult to show that the function g_1 is identified by employing the strategy adapted from Chen and Khan (2003), as stated in the following theorem.

Theorem 3.7. *Suppose a heteroskedastic sample selection model 3.1-3.3 and 3.16. Under Assumptions FI and SP(iii), the average structural function $g_1(x)$ for all $x \in \mathcal{X}$ is identified.*

Proof. First note that

$$\begin{aligned}\mathbb{E}(Y^* \mid X = x, Z = z, D = 1) &= g_1(x) + g_{20}(x)\mathbb{E}(\varepsilon_0 \mid X, Z, D = 1) \\ &= g_1(x) + g_{20}(x)\lambda_0(P).\end{aligned}$$

Then by picking two different quantiles τ_1 and τ_2 of the observed outcome Y, we have

$$\begin{aligned}Q_{\tau_1}(Y^* \mid X = x, Z = z, D = 1) &= g_1(x) + g_{20}(x)Q_{\tau_1}(\varepsilon_0 \mid X = x, Z = z, D = 1) \\ &= g_1(x) + g_{20}(x)\lambda_{\tau_1}(P(z))\end{aligned}$$

and similarly for τ_2.

Taking the difference of these two quantiles of the observed outcome Y gives

$$\Delta Q_{\tau_{12}} = g_{20}(x)\Delta\lambda_{\tau_{12}}(P),$$

where $\Delta Q_{\tau_{12}} \equiv Q_{\tau_1}(Y^* \mid X = x, Z = z, D = 1) - Q_{\tau_2}(Y^* \mid X = x, Z = z, D = 1)$ and $\Delta\lambda_{\tau_{12}}(P) \equiv \lambda_{\tau_1}(P(z)) - \lambda_{\tau_2}(P(z))$.

Define the transformed variables

$$\widetilde{Y} = \frac{Y}{\Delta Q_{\tau_{12}}}, \quad \widetilde{g}_1(x) = \frac{g_1(x)}{\Delta Q_{\tau_{12}}}, \text{ and } \widetilde{\lambda}(P) = \frac{\lambda(P)}{\Delta\lambda_{\tau_{12}}(P)}.$$

Our analysis leads to the following relationship

$$\mathbb{E}(\widetilde{Y} \mid X, Z, D = 1) = \widetilde{g}_1(X) + \widetilde{\lambda}(P).$$

Under the measurable separability assumption (i.e., Assumption SP (iii)), the function $\widetilde{g}_1$ is identified and therefore $g_1(x) = \widetilde{g}_1(x)\Delta Q_{\tau_{12}}$ is identified. □

3.5.2 Identification of Structural Function $g(x, \varepsilon)$

3.5.2.1 Nonparametric Quantile Regression Representation

The following independence and monotonicity assumptions are standard in nonseparable models. The independence assumption is weaker than it would be in separable models due

to the fact that ε enters the outcome equation in a nonadditive way so that the general interaction between X and ε is allowed. The model with endogeneity by relaxing the independence assumption will be considered later. Monotonicity in ε for every x guarantees invertibility of the structural function in the second argument. Matzkin (2003) discusses alternative normalization strategies.

Assumption IMN. *(Independence) ε is a scalar error term that is distributed, with a distribution function F_ε , independently of X in the population;*
(Monotonicity) For every $x \in \mathcal{X}$, $g(x,\varepsilon)$ is strictly increasing in ε;
(Normalization) The error term ε is uniformly distributed on [0,1].

Matzkin (2003) studies identification and estimation of a nonadditive model $Y = g(X,\varepsilon)$. Matzkin's Lemma 1 showed that the structural function g is only identified up to a strict transformation [11] even when g is assumed to be strictly increasing in ε and X is independent of ε. On the other hand, once g is identified, so is F_ε, and vice versa. This is because

$$\begin{aligned} F_\varepsilon(e) &= \Pr(\varepsilon \le e) \\ &= \Pr(\varepsilon \le e \mid X = x) \\ &= \Pr(g(X,\varepsilon) \le g(x,e) \mid X = x) \\ &= F_{Y^*|X=x}(g(x,e)) \end{aligned} \tag{3.17}$$

and therefore

$$g(x,e) = F^{-1}_{Y^*|X=x}(F_\varepsilon(e)) \tag{3.18}$$

provided that the conditional distribution function, $F_{Y^*|X=x}$, of Y^* given $X = x$ is strictly increasing, i.e. the density of ε is everywhere positive.

Clearly, $g(x,e)$ is not identified because the conditional distribution function of ε , $F_\varepsilon(e)$, is unobserved. Normalization of the structural function or of the distribution of ε is required for $g(x,e)$ to be identified.

[11] That is, $g(x,\varepsilon)$ and $g'(x,s(\varepsilon))$ are observationally equivalent if and only if $g(x,\varepsilon) = g'(x,s(\varepsilon))$ where s is some continuous and strictly increasing function. See Matzkin (2003, 2007) for details.

Normalization of F_ε: $F_\varepsilon \sim U(0,1)$. If $F_\varepsilon(\bar{\varepsilon})$ is continuous then $\eta = F_\varepsilon \sim U(0,1)$. So we can write

$$g(X,\varepsilon) = g(X, F_\varepsilon^{-1}(\eta)) \equiv g'(X,\eta).$$

This normalization strategy was employed by e.g. Imbens and Newey (2009) and Blundell and Powell (2003). Under this normalization, equation (3.18) becomes

$$g(x,e) = F^{-1}_{Y^*|X=x}(e). \tag{3.19}$$

Equation (3.19) implies that $g(x,e)$ is the e^{th} quantile of the conditional distribution of Y^* given $X = x$. Therefore, one can estimate the conditional quantile function $m(x,e)$ by inverting the estimate of conditional distribution function that has been developed in the literature. One could also equivalently use the check function approach to compute $g(x,e)$, see e.g. Chaudhuri (1991) and Chaudhuri, Doksum, and Samarov (1997). Clearly, once we can identify the object $F_{Y^*|X=x}(\cdot)$ from the selected sample, identification of the structural function g follows immediately. However, this is not usually the case in the presence of sample selection. We can identify $F_{Y|X=x,D=1}(\cdot)$ rather than $F_{Y^*|X=x}(\cdot)$ from the observed data without further restrictions.

3.5.2.2 Nonparametric Quantile Regression with Sample Selection

Under Assumption IMD and the normalization, the *nonseparable* outcome equation $Y^* = g(X,\varepsilon)$ can be equivalently represented as the usual nonparametric quantile regression

$$Y^* = g_\tau(X) + \varepsilon_\tau, \tag{3.20}$$

where $g_\tau(X) \equiv g(X,\tau)$ and $\varepsilon_\tau \equiv g(X,\varepsilon) - g(X,\tau)$. Note that the τ^{th} quantile of ε_τ (denoted by $Q_\tau(\varepsilon_\tau)$) is equal to zero by construction. To see this,

$$\begin{aligned} Q_\tau(\varepsilon_\tau \mid X = x) &= Q_\tau\left(g(X,\varepsilon) - g(X,\tau) \mid X = x\right) \\ &= g(x, Q_\tau(\varepsilon)) - g(x,\tau) \\ &= g(x,\tau) - g(x,\tau) \\ &= 0, \end{aligned}$$

where the second equality follows by the monotonicity assumption and the third equality is due to normalization of $F_\varepsilon(\cdot)$ to be uniformly distributed on [0,1].

Equation (3.20) provides a basis for the identification analysis of the structural function g that is convenient to work with in the context of sample selection. Identification and estimation of the quantile structural function g_τ are interesting tasks in empirical work. For example, in the empirical analysis of wages and participation, estimating g_τ allows the researcher to correct wage inequality for nonrandom selection to work. This correction for selection bias is particularly important in the case in which employment rates vary over time or by groups (for details see Arellano and Bonhomme (2013)). In the absence of sample selection, the structural function g_τ can be estimated by the standard nonparametric regression models (e.g., local polynomial estimation by Chaudhuri (1991)). In the presence of sample selection, the model reduces to the nonparametric quantile regression with sample selection. In particular, the conditional quantile of the observed outcome is given by

$$Q_\tau(Y^* \mid X, Z, D = 1) = g_\tau(X) + Q_\tau(\varepsilon_\tau \mid X, Z, D = 1), \tag{3.21}$$

where $Q_\tau(\varepsilon_\tau \mid X, Z, D = 1)$ is the quantile version of the so-called selection correction term and in general $Q_\tau(\varepsilon_\tau \mid X, Z, D = 1) \neq 0$ due to the fact that the error term ε_τ (or ε) is generally not independent of the selection indicator variable D.

However, unlike the conventional sample selection model, the main difficulty to achieve identification of the structural function g based on (3.21) is that the selection correction term,

$Q_\tau(\varepsilon_\tau \mid X, Z, D = 1)$, is in general no longer separable between X and $P(Z)$,[12] indicating that conditioning on the propensity score is not sufficient to control for selection bias. To see this, the conditional distribution function of ε_τ is given by

$$\begin{aligned}
\Pr(\varepsilon_\tau \le c \mid X, Z, D = 1) &= \Pr(g(X, \varepsilon) \le c + g(X, \tau) \mid X, Z, D = 1) \\
&= \Pr(\varepsilon \le g^{-1}(X, c + g(X, \tau)) \mid X, Z, D = 1) \\
&= G(g^{-1}(X, c + g(X, \tau)), P(Z)) \\
&\equiv f(c, \tau, X, P(Z)), \qquad (3.22)
\end{aligned}$$

where $g^{-1}(x, \cdot)$ denotes the inverse of $g(x, \cdot)$ and G is the conditional distribution function of $\varepsilon \mid Z, D = 1$. Hence, $Q_\tau(\varepsilon_\tau \mid X, Z, D = 1)$ can generally be expressed as $f(\tau, X, P(Z))$. In this case, it is clear that $g_\tau(x)$ is not distinguishable from $g^*_\tau(x)$ since

$$\begin{aligned}
Y &= g_\tau(X) + f(\tau, X, P(Z)) + U_\tau \\
&= g^*_\tau(X) + f^*(\tau, X, P(Z)) + U_\tau, \qquad (3.23)
\end{aligned}$$

where $g^*_\tau(X) = g_\tau(X) - a(X)$ and $f^*(\tau, X, P(Z)) = f(\tau, X, P(Z)) + a(X)$ satisfy the same restriction $\Pr(U_\tau \le 0 \mid X, Z, D = 1) = \tau$.

We note that the fact that the selection correction term in (3.21) involves the covariates X complicates the point identification of the quantile function g_τ. This implies that, as discussed in Angrist (1997), conditioning on the propensity score P is not sufficient to control for selection bias.[13]

One possibility is to achieve point identification of g_τ by appealing to identification-at-infinity, as discussed in the next section.

[12] Arellano and Bonhomme (2013) show that $Q_\tau(Y^* \mid X, Z, D = 1)$ is non-additive in X and $P(Z)$, unless in the linear specification $q_\tau(X) = X'\beta_\tau$ all coefficients of β_τ but the constant are independent of τ, or ε and ν are statistically independent.

[13] Buchinsky (1998, 2001) considers linear quantile regression models with sample selection where the outcome equation is $Y^* = X'\beta_0 + \varepsilon = X'\beta_\tau + \varepsilon_\tau$ and $\varepsilon_\tau \equiv X'(\beta_0 - \beta_\tau) + \varepsilon$ satisfying $Q_\tau(\varepsilon_\tau \mid X) = 0$. Under Buchinsky's (2001) Assumption E, he shows that the propensity score as well as the conditional distribution function of ε_τ given Z are only a function of $q(Z)$. It then follows that the selection correction $Q_\tau(\varepsilon_\tau \mid X, Z, D = 1)$ depends only on $q(Z)$. However, this assumption implies independence between ε and X conditional on selection probability and therefore all quantile curves are parallel. This limitation was also pointed out by Melly and Huber (2012).

3.5.2.3 Identification at Infinity

Based on equation (3.21), one approach is to appeal to the so-called "identification at infinity" argument. That is, the quantile function g_τ is point identified by focusing only on the observations for which the probability of being selected is high, i.e. the group of observations with the propensity score close to one, i.e.,

$$P > 1 - \gamma_n^{-1},$$

where the parameter γ_n is the bandwidth satisfying $\gamma_n \to \infty$ as the sample size n goes to infinity.

For some value $X = x$ suppose we can find a limit set $\mathbb{Z}$ so that for $z \in \mathbb{Z}$ the regression function $g(z) \to \infty$ and therefore $P = P(D = 1 \mid Z = z) = P(\nu \leq g(z) \mid Z = z) \to 1$, i.e.,

$$\lim_{z \to \mathbb{Z}} P(D \mid Z = z) = 1.$$

As a consequence,

$$\begin{aligned}
\lim_{P(z)\to 1} Q_\tau(Y \mid X = x, Z = z, D = 1) &= g_\tau(x) + \lim_{P(z)\to 1} Q_\tau(\varepsilon_\tau \mid X = x, Z = z, \nu \leq q(z)) \\
&= g_\tau(x) + Q_\tau(\varepsilon_\tau \mid X = x, Z = z) \\
&= g_\tau(x).
\end{aligned}$$

Thus g_τ is (point) identified. In words, identification at infinity allows us to solve the selection bias problem by essentially identifying a group of individuals for whom there is no selection problem.

We briefly discuss the estimation strategy using the identification at infinity argument. To facilitate the proof of the distribution theory, one can follow Andrews and Schafgans (1998) and Klein, Shen, and Vella (2011) by replacing the indicator function $\mathbf{1}\{\cdot\}$ with a smooth function $S(\cdot)$. Then the smoothed versions of the local constant and local linear estimators of the quantile function are given by

$$\widehat{g}_\tau^{lc}(x) = \arg\min_\alpha \sum_{i=1}^n \rho_\tau(Y_i - \alpha) K\left(\frac{X_i - x}{h_n}\right) S(t_i(\widehat{P}_i - 1 + \gamma_n^{-1}))$$

and

$$(\widehat{g}_\tau^{ll}, \widehat{\beta}) = \arg\min_{q,\beta} \sum_{i=1}^{n} \rho_\tau(Y_i - q - (X_i - x)'\beta) K\left(\frac{X_i - x}{h_n}\right) S(t_i(\widehat{P}_i - 1 + \gamma_n^{-1}))$$

respectively where $S(\cdot)$ is a non-decreasing [0,1]-valued function for which $S(x) = 0$ for $x \leq 0$ and $S(x) = 1$ for $x \geq b$ for some $0 < b < \infty$, $\widehat{P}$ is the estimator for the propensity score $P(D \mid Z)$, and

$$t_i \equiv 1(\widehat{f}_Z(z_i) \geq O_p(n^{-\delta}))$$

is a trimming function restricting the estimate of the density of Z to be away from 0 and δ is a small positive number. We leave deriving asymptotic properties of the estimators $\widehat{g}_\tau^{lc}(x)$ and $\widehat{g}_\tau^{ll}$ to future research.

3.6 Nonparametric Estimation

In this section, we adopt a control function approach as the estimation strategy to estimate the nonseparable sample selection models with continuous endogenous regressors as discussed in Section 3.4.2.2. Recall that the identification and estimation of $\text{ASF}^S(x)$ rely on introducing the CDF of X_1 conditional on Z, $V \equiv F_{X_1|Z}(X_1, Z)$, along with the propensity score P, as control variables to correct for endogeneity and selection jointly.

First step: the first step is to nonparametrically estimate control variables P_i and V_i, $i = 1, \ldots, n$. Since $P_i \equiv \mathbb{E}(D_i \mid Z_i)$ and $V_i \equiv F_{X_1|Z}(X_{1i}, Z_i) = \mathbb{E}[\mathbf{1}\{X_1 \leq X_{1i}\} \mid Z_i]$, we can apply nonparametric regression methods directly. We use series methods to obtain the first-step estimators of P_i and V_i. Let the number of approximating functions for P and V be L_1 and L_2 respectively and $r^{pL_1}(\cdot)$ denote the $L_1 \times 1$ vector of the first L_1 approximating functions for P and the similar notation applies to V,

$$r^{pL_1}(z) = (r^p_{1L_1}(z), \ldots, r^p_{L_1L_1}(z))',$$

$$r^{vL_2}(z) = (r^v_{1L_2}(z), \ldots, r^v_{L_2L_2}(z))',$$

and let R^p and R^v denote the $n \times L_1$ and $n \times L_2$ matrices whose i^{th} row are given by $r^{pL_1}(Z_i)'$ and $r^{vL_2}(Z_i)'$ respectively, then

$$R^p = \left(r^{pL_1}(Z_1), \ldots, r^{pL_1}(Z_n)\right)',$$

$$R^v = \left(r^{vL_2}(Z_1), \ldots, r^{vL_2}(Z_n)\right)'.$$

Let $\widehat{P}(z)$ and $\widehat{V}(x_1, z)$ be predicted values from regression of D_i on $r^{pL_1}(Z_i)$ and of $\mathbf{1}\{X_{1i} \leq x_1\}$ on $r^{vL_2}(Z_i)$, respectively. We then form the first-step estimators $\widehat{P}_i$ and $\widehat{V}_i$ by

$$\widehat{P}_i \equiv \widehat{\mathbb{E}}(D_i \mid Z_i) = r^{pL_1}(Z_i)'\widehat{\beta}_p$$

and

$$\widehat{V}_i \equiv \widehat{\mathbb{E}}(\mathbf{1}\{X_1 \leq X_{1i}\} \mid Z_i) = r^{vL_2}(Z_i)'\widehat{\beta}_v(X_{1i}),$$

where $\widehat{\beta}_p = (R^{p'}R^p)^- R^{p'}\mathbf{D}$, $\widehat{\beta}_v(X_{1i}) = (R^{v'}R^v)^- R^{v'}\mathbf{1}\{\mathbf{X}_1 \leq X_{1i}\}$, $\mathbf{D} = (D_1, \ldots, D_n)'$, $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n})'$, and $(\cdot)^-$ denotes the generalized inverse of $(\cdot)$.

The first step estimates can then be used to construct an estimator $\widehat{\mathbb{E}}^S(Y|X = x, \widehat{P}_i, \widehat{V}_i)$ of $\mathbb{E}^S(Y|X = x, \widehat{P}_i, \widehat{V}_i)$.

Second Step: recall that averaging the conditional mean of Y given $X = x$, P, V, and $D = 1$ over the joint distribution of (P, V) in the selected data gives $\mathrm{ASF}^S(x)$, the parameter of interest. As a result, we obtain nonparametric estimates $\widehat{\mathbb{E}}^S(Y_i \mid x, \widehat{P}_i, \widehat{V}_i)$ by regressing Y_i on $\widehat{W}_i \equiv (X_i, \widehat{P}_i, \widehat{V}_i)$ in the selected data. To describe a series estimator of $h(w) \equiv \mathbb{E}^S(Y \mid W = w)$ where $w = (x, p, v)$, let $W_i = (X_i, P_i, V_i)$ as the sample base function transformations. Setting $Y_i = h(W_i) + e_i$, $F_i = \sum_{k=K+1}^{\infty} \beta_k r_{kK}(W_i)$, and $h_K^*(W_i) = \sum_{k=0}^{K} \beta_k r_{kK}(W_i)$, we have

$$\begin{aligned} Y_i &= \beta_0 + \sum_{k=1}^{\infty} \beta_k r_{kK}(W_i) + e_i \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k r_{kK}(W_i) + F_i + e_i \\ &= h_K^*(W_i) + F_i + e_i. \end{aligned}$$

The second-step series estimator for $h(w)$ given in the selected data can be formed by $\widehat{h}(w) = r^K(w)'\widehat{\beta}$, where

$$\widehat{\beta} = (\widehat{R}'\widehat{R})^{-1}\widehat{R}'Y, \qquad \widehat{R} = (r^K(\widehat{W}_1), \ldots, r^K(\widehat{W}_n)),$$

$$r^K(\widehat{W}_i) = (r_{1K}(\widehat{W}_i), \ldots, r_{KK}(\widehat{W}_i)), i = 1, \ldots, n, \qquad Y = (Y_1, \ldots, Y_n)',$$

where $\widehat{R}$ denote $R = (r^K(W_1)', \ldots, r^K(W_n)')'$ with the $W_i = (X_i, P_i, V_i)$ replaced by $\widehat{W}_i = (X_i, \widehat{P}_i, \widehat{V}_i)$.

Define $\widehat{w}_i = (x, \widehat{P}_i, \widehat{V}_i)$. The estimate of $h(\widehat{w}_i)$ is $\widehat{h}(\widehat{w}_i) = r^K(\widehat{w}_i)'\widehat{\beta}$ and therefore the estimate of the vector $h = (h(\widehat{w}_1), \ldots, h(\widehat{w}_n))'$ can be formed as

$$\widehat{h} = \widetilde{R}\widehat{\beta} = \widetilde{R}(\widehat{R}'\widehat{R})^{-1}\widehat{R}'Y,$$

where $\widetilde{R} = (r^K(\widehat{w}_1)', \ldots, r^K(\widehat{w}_n)')'$.

Third step: an estimator for $\mathrm{ASF}^S(x)$, $\widehat{\mathrm{ASF}}^S(x)$, can be obtained by plugging in $\widehat{\mathbb{E}}^S(Y_i \mid x, \widehat{P}_i, \widehat{V}_i)$ and by replacing integrals with sample average

$$\widehat{\mathrm{ASF}}^S(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbb{E}}^S(Y_i \mid x, \widehat{P}_i, \widehat{V}_i).$$

3.7 Convergence Rates

We impose the following regularity conditions.

Assumption 3.2. *The data, $\{Y_i, X_i, Z_i, D_i\}$ for $i = 1, \ldots, n$ are i.i.d., and $Var(Y \mid w)$ and $Var(X_1 \mid Z)$ are bounded.*

Assumption 3.3. *The control variables $F_{X_1|Z}(X_1 \mid Z)$ and $P(Z)$ are continuously differentiable of order d_v and d_p on the support of (X_1, Z) and Z respectively. And $h(W)$ is continuously differentiable of order d_h on the support of W.*

As known in standard series estimation, the smoothness conditions in Assumption 4.2 control the bias of the series estimator. To be precise the derivative orders d_p, d_v and d_h determine the rates of approximation of $P(Z)$, $F_{X_1|Z}(X_1 \mid Z)$ and $h(W)$ by the first and second steps regressors respectively. Let r_w be the dimension of w and r_z the dimension of Z. Then for splines and power series the rate of approximation for $h(W)$ will be $O(K^{-d_h/r_w})$. The following conditions restrict the rate of growth of the number of terms K, L_1, and L_2 as sample size goes to infinity.

Assumption 3.4. *For splines* $(K^2 + KL_i^{1/2})[(L_i/n)^{1/2} + L_i^{-s/r_z}] \to 0$ *and for power series,* $(K^3 + K^2 L_i)[(L_i/n)^{1/2} + L_i^{-s/r_z}] \to 0$, $i = 1, 2$ *and* $s = d_p$ *when* $i = 1$ *and* $s = d_v$ *when* $i = 2$.

The following Lemma gives the results of the mean square convergence rates of the first step estimators $\widehat{P}_i$ and $\widehat{V}_i$.

Lemma 3.1. *If Assumptions 3.2-3.4 are satisfied then*

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{P}_i - P_i)^2 = O_p(L_1/n + L_1^{-2d_p/r_z})$$

$$\textit{and} \quad \frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_i - V_i)^2 = O_p(L_2/n + L_2^{1-2d_v/r_z}).$$

Proof. The first part of this Lemma is proven by Theorem 1 of Newey (1997). The proof of the second part is provided in Appendix. □

The mean square convergence rate for $P(Z)$ is the standard result of series estimators: the term L_1/n corresponds to the variance term and the term $L_1^{-2d_p/r_z}$ to the biased term. As for the mean square convergence rate for $v(x_1, z)$, there is an extra term L_2 corresponding to the biased term relative to the standard result. The reason for that is the fact that $\widehat{V}_i = r^{vL_2}(Z_i)'\widehat{\beta}_v(X_{1i})$ where the series coefficients $\widehat{\beta}_v(X_{1i})$ depends on X_{1i} and thus $\widehat{V}_i$ depends on X_{1i} as well, which lowers the convergence rate.

The following result gives mean square error and uniform convergence rates for $h(w)$.

Theorem 3.8. *Let* $F_w(w)$ *denote the cumulative distribution function of* w_i. *Suppose that there exists a sequence of constant* $\zeta_0(K)$ *satisfying* $\sup_w \|r^K(w)\| \leq \zeta_0(K)$ *and that Assumptions 3.2-3.4 hold. Then*

$$\int (\widehat{h}(w) - h(w))^2 dF_w(w) = O_p\left(K/n + K^{-2d_h/r_w} + L_1/n + L_1^{-2d_p/r_z} + L_2/n + L_2^{1-2d_v/r_z}\right)$$

and

$$\sup_w |\widehat{h}(w) - h(w)| = O_p\left(\zeta_0(K)\left(K/n + K^{-2d_h/r_w} + L_1/n + L_1^{-2d_p/r_z} + L_2/n + L_2^{1-2d_v/r_z}\right)\right)^{1/2}.$$

Proof. See Appendix. □

This result says that the mean square convergence rate of $\widehat{h}(w)$ depends on the numbers L_1 and L_2 of the approximating functions in the first step and the number K in the second step. If L_1 is chosen proportional to $n^{r_z/(2d_p+r_z)}$, L_2 proportional to $n^{r_z/2d_v}$, and K proportional to $n^{r_w/(2d_h+r_w)}$, then convergence rates would achieve Stone's (1982) optimal rates, i.e.

$$\int (\widehat{h}(w) - h(w))^2 dF_w(w) = O_p\left(\max\left\{n^{-2d_p/(2d_p+r_z)}, n^{(r_z-2d_v)/2d_v}, n^{-2d_h/(r_w+2d_h)}\right\}\right).$$

Similar to other two-step series estimators by Newey (1997), NPV, and DNV, this result shows that the mean square convergence rate of $\widehat{h}(w)$ is determined by the optimal rates in the first and that in the second steps if p and v did not have to be estimated. Therefore, the mean square convergence rate of $\widehat{h}(w)$ would achieve the optimal rate if the optimal rates in the first step are faster than the optimal rate in the second step if the propensity score P and the conditional CDF V were known. The above result can also be compared to additive models. In nonparametric sample selection models with additivity, $\mathbb{E}^S(Y \mid X, P, V)$ is additively separable in subvectors X and (P, V) of W, i.e.

$$\mathbb{E}^S(Y \mid X, P, V) = g(X) + \lambda(P, V).$$

Suppose that P and V are known. A series estimator for $h(W)$ could be constructed by imposing additivity restrictions, i.e. by including in $r^K(W)$ functions that depend either on X or (P, V), but not on both. If both $g(X)$ and $\lambda(P, V)$ are continuously differentiable of order s_1, the exclusion of the interaction terms will increase the approximation rate to $K^{-s/\chi}$, where s is the derivative order of $g(X)$ and $\lambda(P, V)$; χ is the maximum of the dimension of X and of the dimension of (P, V).

3.8 Simulations

3.8.1 Setup

To illustrate how our methodology may perform in practice, some monte carlo results of empirical properties of the three-step nonparametric estimator we propose in Section 3.6 are

presented. We consider a variety of data generating processes which can be used to evaluate the performance of our estimator in the correctly specified model as well as the cost our estimator may incur under possible misspecification.

Apart from our three-step series estimator in nonseparable sample selection models with an endogenous regressor (labelled 2SNSE), we also consider DNV's two-step estimator in a separable setting (labelled 2SSE), and Heckman's two-step parametric estimator (labelled 2SHK), which is widely used in applied work.

We consider two different sample sizes, n=300 and n=500. The number of replications is 500. As for nonparametric estimation, we consider polynomial series estimation in the first and second stages and let $L_1 = L_2$ for convenience. Regarding the expansion terms entering as regressors in both two stages, in the case of $L_1 = L_2 = 6$ and $K = 8$ in an nonseparable setting for example, the first stage uses regressors (or approximating functions) $(1, Z, Z^2, Z^3, Z^4, Z^5)$. For the estimation of $\mathbb{E}^S[Y \mid X = x, P, V]$ in the second stage, the regressors are $(1, x, \widehat{P}, \widehat{V}, x^2, x\widehat{P}, \widehat{P}^2, \widehat{P}\widehat{V})$. As for the 2SSE estimator, under the assumption $\mathbb{E}^S[\varepsilon \mid X, P, \eta] = \lambda(P, \eta)$, the conditional expectation of Y conditional on $X = x, P, \eta$, and selection can be expressed as

$$\mathbb{E}^S[Y \mid X = x, P, \eta] = g(x) + \lambda(P, \eta),$$

where $\eta = Y - \mathbb{E}(X \mid Z)$.

It is worth noting that there are two differences in the estimation procedures between the 2SNSE and 2SSE estimators. First, to implement the 2SSE estimator, we estimate η_i rather than V_i by $\widehat{\eta}_i = X_i - \widehat{\pi}(Z_i)$. Second, additive restrictions are needed to be imposed on the 2SSE estimator, i.e. the approximating functions depend either on x or on (P, η), but not on both. That is, the regressors used in the first stage is the same as that of 2SNSE. The second stage for $K = 8$ uses regressors $(1, x, \widehat{P}, \widehat{\eta}, x^2, \widehat{P}^2, \widehat{P}\widehat{\eta}, \widehat{\eta}^2)$.

Regarding the designs considered below, we change the specification of the outcome equation and the reduced-form equation for the endogenous variable, including the cases in which both are nonseparable (Design 1), both are nonlinear but separable (Design 2), and

both are linear equations (Design 3). On the other hand, in order to simplify the calculation of the truncated mean of ε, $\mathbb{E}(\varepsilon \mid \nu \leq z'\alpha)$, which is an ingredient to compute the true value of the average structural function given selection, we consider the design in which the error term in the outcome equation, ε, is the combination of two independent uniformly distributed variables and the error term in the selection equation, ν, is expressed as $\rho\varepsilon + u_2$ where u_2 is uniformly distributed (Design 1). For Designs 2 and 3, we use the same specification of the joint distribution of the error terms (ε, ν), i.e. the bivariate normal distribution.

The average sample mean squared error is given by

$$\frac{1}{R \times n_e} \sum_{r=1}^{R} \sum_{k=1}^{n_e} \left[\widehat{\mathrm{ASF}}^{S}_{(r)}(x_k) - \mathrm{ASF}^{S}(x_k) \right]^2,$$

where R is the number of replications, n_e is the number of the evaluation points, x_k is the evaluation points and $\widehat{\mathrm{ASF}}^{S}_{(r)}$ is an estimator of ASF^{S} based on $(x_{r1}, y_{r1}), \ldots, (x_{rn}, y_{rn})$, the rth sample drawn.

3.8.1.1 Design 1

The first design we consider is specified as follows:

$$\begin{aligned}
Y_i &= 1 - 0.5X_i + X_i^3\varepsilon_i + \varepsilon_i, \\
X_i &= \eta_i Z_{1i}^{1-\eta_i}, \quad Z_{1i} \sim U(0,1), \eta_i \sim U(0,1) \\
D_i &= \mathbf{1}\{\alpha_0 + \alpha_1 Z_{2i} + \nu_i < 0\}, \quad Z_{2i} \sim U(0,1) \\
\varepsilon_i &= \theta\eta_i + u_{1i}, \quad u_{1i} \sim U(0,1) \\
\nu_i &= \rho\varepsilon_i + u_{2i}, \quad u_{2i} \sim U(0,1).
\end{aligned}$$

We set the true values of other parameters as follows

$$\alpha_0 = -1,\ \alpha_1 = -1.2,\ \theta = 0.9,\ \text{and } \rho = 0.9.$$

The coefficient ρ reflects the correlation between ε_i and ν_i. Similarly, as ρ increases, the correlation between ε_i and η_i increases and the problem of endogeneity bias magnified.

The true value of the average structural function given selection is given by

$$\begin{aligned}\text{ASF}^S(x) &\equiv \int g(x,\varepsilon)dF_\varepsilon^S \\ &= \mathbb{E}[1 - 0.5X + X^3\varepsilon + \varepsilon \mid X = x, Z, D = 1] \\ &= 1 - 0.5x + (x^3 + 1)\mathbb{E}[\mathbb{E}(\varepsilon \mid Z_2, D = 1)],\end{aligned}$$

where

$$\mathbb{E}[\mathbb{E}(\varepsilon \mid Z_2, D = 1)] = \frac{-1}{2\rho}(2\alpha_0 + \alpha_1 + 1) - \frac{\theta + 1}{4}.$$

3.8.1.2 Design 2

The second design considered here is the nonlinear outcome equation with an nonlinear reduced form for x in a separable setting

$$\begin{aligned}Y_i &= 1 - 0.5\exp(X_i) + \varepsilon_i, \\ X_i &= \log(Z_{1i}) + \eta_i, \quad Z_{1i} \sim U(0,1) \\ D_i &= \mathbf{1}\{\alpha_0 + \alpha_1 Z_{2i} + \nu_i < 0\}, \quad Z_{2i} \sim U(0,1).\end{aligned}$$

Assume that (ε, ν) are jointly normally distributed

$$\begin{pmatrix} \varepsilon_i \\ \nu_i \end{pmatrix} = N\left(0, \begin{pmatrix} 1 & \sigma_{\varepsilon\nu} \\ \sigma_{\varepsilon\nu} & \sigma_\nu \end{pmatrix}\right)$$

and that η_i are determined by

$$\eta_i = r\varepsilon_i + u_i, \ u_i \sim N(0,1).$$

The true values of other parameters are as follows

$$\alpha_0 = -0.3, \ \alpha_1 = -0.8, \ \sigma_\nu = 1, \ \sigma_{\varepsilon\nu} = 0.8, \ r = 0.8.$$

Using the formula of the average structural function given selection, the true value of $\text{ASF}^S(x)$ is of the form

$$\text{ASF}^S(x) = 1 - 0.5\exp(x) + \text{E}_{Z_2}[\mathbb{E}(\varepsilon \mid Z_2, D = 1)]$$

where

$$\mathbb{E}_{Z_2}[\mathbb{E}(\varepsilon \mid Z_2, D = 1)] = -\rho_{\varepsilon\nu}\mathbb{E}_{Z_2}\left[\frac{\phi(\alpha_0 + Z_2\alpha_1)}{\Phi(\alpha_0 + Z_2\alpha_1)}\right].$$

3.8.1.3 Design 3

The last design we consider fully exploits the parametric assumptions consisting of the specifications of the familiar linear outcome equation with normality assumption of errors ε and ν. We also consider exclusion restrictions to avoid the multicollinearity problem that would lead Heckman's estimator to perform poorly. The model in Design 3 is given by

$$\begin{aligned}
Y_i &= 1 - 0.5X_i + \varepsilon_i, \\
X_i &= 1 + 2Z_{1i} + \eta_i, \quad Z_{1i} \sim U(0,1) \\
D_i &= \mathbf{1}\{-0.3 - 0.8Z_{2i} + \nu_i < 0\}, \quad Z_{2i} \sim U(0,1)
\end{aligned}$$

The true value of $\text{ASF}^S(x)$ is given by

$$\text{ASF}^S(x) = 1 - 0.5x + \mathbb{E}_{Z_2}[\mathbb{E}(\varepsilon \mid Z_2, D = 1)].$$

3.8.2 Results

The average sample mean square errors for $\widehat{\text{ASF}}^S(x)$ in different designs are shown in Tables 3.3-3.5. The average of $\widehat{\text{ASF}}^S(x)$ (top panel) and the median, together with the upper and lower 0.05 quantiles, of $\widehat{\text{ASF}}^S(x)$ (bottom panel) across replications for each x are presented in Figures 3.1-3.4. The numbers of the approximating functions, L_1, L_2, and K, are set exogenously and their different combinations are also considered.

When the model is correctly specified for the 2SNSE estimator, some conclusions can be drawn. First, the 2SNSE estimator is less biased than the 2SSE and 2SHK estimators to estimate $\widehat{\text{ASF}}^S(x)$ but the bias tends to be larger as x is close to the boundaries of its support. In terms of average sample mean square error, the 2SNSE estimator performs well uniformly over the other two estimators we consider. Also, the average sample mean square error of $\widehat{\text{ASF}}^S(x)$ from the 2SNSE procedure falls as the sample size increases. Second, for a given sample size, the 2SNSE estimator appears to require a lower level of smoothing

in the first stage and has the smaller average sample mean square error around $K = 7$ in the second stage. Third, the true value of $\text{ASF}^S(x)$ lies inside the quantile range for the 2SNSE estimator. Overall, the 2SNSE estimator performs well when the nonseparable model is correctly specified. As for Design 2 where the model is correctly specified for the 2SSE estimator, it is not surprising that the 2SSE estimator performs best in this setting. The 2SSE estimator actually does a good job of fitting $\text{ASF}^S(x)$. On the other hand, the bias of the 2SNSE estimator is not very large and is relatively smaller than that of the 2SHK estimator. Roughly speaking, the 2SNSE estimator still tracks the true value of $\text{ASF}^S(x)$ (see Figure 3.2). From a theoretical point of view, the control variable $V_i = F_{X_1|Z}(X_{1i} \mid Z_i)$ of the 2SNSE estimator is the one-to-one function of the reduced form error, which is $\eta_i = X_{1i} - \pi(Z_i)$ in the model of Design 2. Therefore, when the model is correctly specified for the 2SSE estimator, the theory would predict that the 2SNSE and 2SSE estimators would not perform differently. In fact, when we change the order of the series terms entering the regression in the second step, i.e. $(1, x, x^2, x^3, \widehat{P}, \widehat{V}, x\widehat{P}, \widehat{P}^2)$ for 2SNSE and $(1, x, x^2, x^3, \widehat{P}, \widehat{\eta}, \widehat{P}^2, \widehat{P}\widehat{\eta})$ for 2SSE, these two estimators have almost the same performance and are close to the true value of $\text{ASF}^S(x)$ (see Figure 3.3), although the average sample mean square errors are not reported here. In addition, even under misspecification, the true value of $\text{ASF}^S(x)$ almost lies inside the quantile range for the 2SNSE estimator. As for Desigm 3 that is misspecified for the 2SNSE estimator, the 2SNSE estimator turns out to have larger bias, especially in the lower range of x. The average sample mean square error of the 2SNSE estimator (0.10643) exceeds that of the correctly specified Heckman's two-step estimator (0.02108) by about 400%. That may be regarded as the cost our estimator incurs under misspecification.

3.9 Conclusion

This chapter has considered a variety of nonseparable sample selection models. In order for the average structural function given selection to be identified in a nonseparable sample selection setting with endogeneity, we use the conditional cumulative distribution function of the endogenous variables given the instruments, as employed by Imbens and Newey (2009),

along with the propensity score, as control variables. We present a simple three-step control function approach to identifying and estimating the average structural function conditional on selection nonparametrically. The convergence rates of the estimator for the average structural function given selection depend on the convergence rate of the estimator for the propensity score and that for the conditional CDF of the endogenous variables given the instruments. The simulation results show that the proposed estimator performs well in comparison with other existing estimators in a correctly specified model. We further explore the possibility of identifying the unconditional parameters of interest such as the average and quantile structural functions in nonseparable sample selection models. This remains an important extension for future research.

3.10 Appendix

Proof of Lemma 3.1

Recall that

$$\widehat{F}_{X_1|Z}(x_1 \mid z) = r^{vL_2}(z)' \left(\frac{1}{n} \sum_{i=1}^{n} r^{vL_2}(Z_i) r^{vL_2}(Z_i)' \right)^{-} \left(\frac{1}{n} \sum_{j=1}^{n} r^{vL_2}(Z_j) \mathbf{1}\{X_{1j} \le x_1\} \right).$$

Let $q_i = r^{vL_2}(Z_i)$ and $\widehat{Q} = 1/n \sum_{i=1}^{n} q_i q_i'$.

The estimate of V_i is given by

$$\begin{aligned}
\widehat{V}_i &= \widehat{F}_{X_1|Z}(X_{1i} \mid Z_i) \\
&= q_i' \widehat{Q}^{-} \frac{1}{n} \sum_{j=1}^{n} [q_j \mathbf{1}\{X_{1j} \le X_{1i}\}] \\
&= q_i' \widehat{Q}^{-} \frac{1}{n} \sum_{j=1}^{n} \left[q_j \left[\mathbf{1}\{X_{1j} \le X_{1i}\} - F_{X_1|\mathbf{z}}(X_{1i} \mid Z_j) + F_{X_1|Z}(X_{1i} \mid Z_j) - q_j' \beta_v(X_{1i}) + q_j' \beta_v(X_{1i}) \right] \right] \\
&= q_i' \widehat{Q}^{-} \frac{1}{n} \sum_{j=1}^{n} \left[q_j \left(\mathbf{1}\{X_{1j} \le X_{1i}\} - F_{X_1|Z}(X_{1i} \mid Z_j) + F_{X_1|Z}(X_{1i} \mid Z_j) - q_j' \beta_v(X_{1i}) \right) \right] + q_i' \beta_v(X_{1i}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widehat{V}_i - V_i = &\left\{ q_i' \widehat{Q}^{-} \frac{1}{n} \sum_{j=1}^{n} q_j \left(\mathbf{1}\{X_{1j} \le X_{1i}\} - F_{X_1|Z}(X_{1i} \mid Z_j) \right) \right\} \\
&+ \left\{ q_i' \hat{Q}^{-} \frac{1}{n} \sum_{j=1}^{n} q_j \left(F_{X_1|Z}(X_{1i} \mid Z_j) - q_j' \beta_v(X_{1i}) \right) \right\} + \left\{ q_i' \beta_v(X_{1i}) - F_{X_1|Z}(X_{1i} \mid Z_i) \right\}.
\end{aligned}$$

The first term above is of the order $O_p(L_2/n)$. For the second term, we first have

$$\begin{aligned}
\sum_{i=1}^{n} q_i' \widehat{Q}^{-} q_i = \operatorname{tr}\left(\sum_{i=1}^{n} q_i' \widehat{Q}^{-} q_i \right) &= \sum_{i=1}^{n} \operatorname{tr}(q_i' \widehat{Q}^{-} q_i) = \sum_{i=1}^{n} \operatorname{tr}(q_i q_i' \widehat{Q}^{-}) = \operatorname{tr}\left(\sum_{i=1}^{n} q_i q_i' \widehat{Q}^{-} \right) \\
&= n \times \operatorname{tr}(\widehat{Q} \widehat{Q}^{-}) = n L_2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \left[q_i' \widehat{Q}^{-} \frac{1}{n} \sum_{j=1}^{n} q_j \left(F_{X_1|Z}(X_{1i} \mid Z_j) - q_j' \beta_v(X_{1i}) \right) \right]^2 &\le \frac{1}{n} \sum_{i=1}^{n} \left[q_i' \widehat{Q}^{-} q_i \frac{1}{n} \sum_{j=1}^{n} \left(F_{X_1|Z}(X_{1i} \mid Z_j) - q_j' \beta_v(X_{1i}) \right)^2 \right] \\
&\le \operatorname{tr}(\widehat{Q} \widehat{Q}^{-}) O_p(L_2^{-2d_v/r_z}) \\
&= O_p(L_2^{1-2d_v/r_z}).
\end{aligned}$$

For the third term, we have

$$\frac{1}{n}\sum_{i=1}^{n}[q_i'\beta_v(X_{1i}) - F_{X_1|Z}(X_{1i} \mid Z_i)]^2 = O_p(L_2^{-2d_v/r_z}),$$

by Assumption 3.4.

Combining the above results, we can conclude

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{V}_i - V_i)^2 = O_p(L_2/n + L_2^{1-2d_v/r_z} + L_2^{-2d_v/r_z}) = O_p(L_2/n + L_2^{1-2d_v/r_z}). \qquad \square$$

Proof of Theorem 3.8

The integrated squared error of $\widehat{h}(w)$ for $h(w)$ is approximately the same as the empirical average

$$\int (\widehat{h}(w) - h(w))^2 dF(w) \cong \frac{1}{n}\sum_{i=1}^{n}(\widehat{h}(W_i) - h(W_i))^2 = \frac{1}{n}(\widehat{h} - h)'(\widehat{h} - h) = \frac{1}{n}\|\widehat{h} - h\|^2.$$

Recall that $\widehat{P}_r = \widehat{R}\left(\widehat{R}'\widehat{R}\right)^{-1}\widehat{R}'$. Define the projection matrix $\widehat{M} = I - \widehat{P}_r$.

$$\begin{aligned}
\|\widehat{h} - h\|^2 &= \|\widehat{P}_r Y - h\|^2 \\
&= \|\widehat{P}_r(h + e) - h\|^2 \\
&= \|\widehat{P}_r e + (\widehat{P}_r - I)h\|^2 \\
&= \|\widehat{P}_r e + (\widehat{P}_r - I)(R\beta + F)\|^2 \\
&= \|\widehat{P}_r e - \widehat{M}F - \widehat{M}R\beta\|^2 \\
&= \|\widehat{P}_r e - \widehat{M}F - \widehat{M}(R - \widehat{R})\beta\|^2 \\
&\leq \{\|\widehat{P}_r e\|^2 + \|\widehat{M}F\|^2 + \|\widehat{M}(R - \widehat{R})\beta\|^2\}
\end{aligned}$$

The sixth equality follows since $\widehat{M}\widehat{R}\beta = (I - \widehat{P}_r)\widehat{R}\beta = \widehat{R}\beta - \widehat{R}(\widehat{R}'\widehat{R})^{-1}\widehat{R}'\widehat{R}\beta = 0$. The last inequality is due to triangle inequality. Let $\mathbb{E}(ee' \mid W) = \Omega$ and note that $\widehat{P}_r$ only depends on X and Z.

$$\mathbb{E}(\|\widehat{P}_r e\|^2 \mid W) = \mathbb{E}[(\widehat{P}_r e)'(\widehat{P}_r e) \mid W] = \mathbb{E}(e'\widehat{P}e \mid W) = \mathrm{tr}(\widehat{P}_r \mathbb{E}(ee' \mid W)) = \mathrm{tr}(\widehat{P}_r \Omega).$$

It can be shown that

$$\|\widehat{P}_r e\|^2 = O_p(K).$$

In addition, $\|F\|^2$ is the squared bias in estimation of h. And the integrated squared bias is assumed to be

$$\mathbb{E}(\|F\|^2) = \int (h_K^*(w) - h(w))^2 dF(w) \leq O(K^{-2d_h/r_w}).$$

Hence we have $\|F\|^2 = O_p(K^{-2d_h/r_w})$.

The last step is to show the convergence rate of $\|(R - \widehat{R})\beta\|^2$.

$$\|(R - \widehat{R})\beta\|^2 = \sum_{j=1}^{n} \left[\sum_{i=1}^{K} \left(r_{iK}(W_j) - r_{iK}(\widehat{W}_j) \right) \beta_i \right]^2,$$

which is bounded by $C^2 \sum_{j=1}^{n} (W_j - \widehat{W}_j)^2 \left(\sum_{i=1}^{K} |\beta_i i| \right)^2$ and C is a constant.

The stochastic order of $\sum_{j=1}^{n} (W_j - \widehat{W}_j)^2$ is determined by the slower of the rates of convergence of $\sum_{j=1}^{n} (P_j - \widehat{P}_j)^2$ and $\sum_{j=1}^{n} (V_j - \widehat{V}_j)^2$. Therefore, $1/n \sum_{j=1}^{n} (W_j - \widehat{W}_j)^2$ can be expressed as

$$O_p\left(L_1/n + L_1^{-2d_p/r_z} + L_2/n + L_2^{1-2d_v/r_z} \right)$$

according to Lemma 3.1.

For the second part,

$$\begin{aligned}
\sup_w |\widehat{h}(w) - h(w)| &= \sup_w |r^K(w)'\beta - h(w) + r^K(w)'(\widehat{\beta} - \beta)| \\
&\leq \sup_w |r^K(w)'\beta - h(w)| + \sup_w |r^K(w)'(\hat{\beta} - \beta)| \\
&\leq O(K^{-d_w/r_w}) + \zeta_0(K)\|\widehat{\beta} - \beta\| \\
&= O_p\left(\zeta_0(K) \left(K^{-2d_w/r_w} + K/n + L_1/n + L_1^{-2d_p/r_z} + L_2/n + L_2^{1-2d_v/r_z} \right) \right)^{1/2}.
\end{aligned}$$

This completes the proof. □

Table 3.1 Relation to the literature

	Endogeneity	Selection	Endogeneity & Selection
Linear models	$Y = X\beta + \varepsilon$ $\mathbb{E}(X\varepsilon) \neq 0$ IA: $\mathbb{E}(Zu) = 0$	$Y^* = X_1\beta + \varepsilon$ $d = \mathbf{1}\{X_2\alpha + \nu > 0\}$ $Y = Y^* \times d$ observed $\mathbb{E}(\varepsilon \mid X_2, D = 1) = \lambda(x_2'\alpha)$ where $\lambda(x) = \rho\phi(x)/\Phi(x)$	-
Semiparametric models	-	$\lambda(x_2'\alpha)$ unknown since $\Pr(D = 1 \mid X_2) = F(x_2'\alpha)$ is unspecified	-
Nonparametric models			
(i)Separable models	NPV (1999) $Y = g(X) + \varepsilon$ $X_1 = \Pi(Z) + \eta$ IA/CF: $\mathbb{E}[\varepsilon \mid X, Z] = \mathbb{E}(\varepsilon \mid X, \eta)$ $= \mathbb{E}(\varepsilon \mid \eta)$	DNV (2003) $Y^* = g(X) + \varepsilon$ $Y = Y^* \times d$ observed IA/CF: $\mathbb{E}[\varepsilon \mid X, Z, D = 1]$ $= \lambda(p)$	DNV (2003) $Y^* = g(X) + \varepsilon$ $X_1 = \pi(Z) + \eta$ $Y = Y^* \times d$ observed IA/CF: $\mathbb{E}[\varepsilon \mid X, Z, \eta, D = 1]$ $= \lambda(p, \eta)$
(ii)Nonseparable models	Imbens & Newey (2009) $Y = g(X, \varepsilon)$ $X_1 = h(Z, \eta)$ v=$F_{X_1\|z} = F_\eta$ IA/CF: $\varepsilon \mid X, v \sim \varepsilon \mid v$	Newey (2007) $Y^* = g(X, \varepsilon)$ $Y = Y^* \times d$ observed IA/CF: $\varepsilon \mid X, Z, p, D = 1$ $\sim \varepsilon \mid p, D = 1$	Our model $Y^* = g(X, \varepsilon)$ $X_1 = \Pi(Z, \eta)$ $y = y^* \times d$ observed IA/CF: $\varepsilon \mid X, Z, p, v, D = 1$ $\sim \varepsilon \mid p, v, D = 1$

Note: (i) IA: Identifying assumptions; CF: Control function approach; $\sim$ denotes equality of conditional distributions; (ii) In the linear models, the orthogonality conditions $\mathbb{E}(Z\varepsilon) = 0$ (along with the rank condition) are sufficient for consistency. The conditional mean restriction is imposed for identification and estimation in separable models. For nonseparable models, the stronger conditional independence assumptions are needed; (iii) We can also use IV approach to identification and estimation for separable and nonseparable models with endogenous regressors, see Blundell and Powell (2003) for a detailed discussion; (iv) Newey (2007) considers identification of the nonseparable sample selection model only.

Table 3.2 Parameters of interest in sample selection models

Sample Selection Models	Parameters of Interest	Outcome equations	Estimating objects
Parametric models	β	$Y = X\beta + \varepsilon$	$\mathbb{E}(Y \mid X, Z, D = 1) = x'\beta + \gamma\text{IMR}$
Semiparametric models	β	$Y = X\beta + \varepsilon$	$\mathbb{E}(Y \mid X, Z, p, D = 1) = x'\beta + \lambda(p)$
Nonparametric models			
(i) Separable models	$h(X)$	$Y = h(X) + \varepsilon$	$\mathbb{E}(Y \mid X, Z, p, D = 1) = h(x) + \lambda(p)$
(ii) Nonseparable models	$\text{ASF}^S(x)$	$Y = g(X, \varepsilon)$	$\mathbb{E}(Y \mid X, Z, D = 1)$

Note: IMR stands for the inverse Mills ratio.

Table 3.3 Design 1: average sample mean square error for $\widehat{\mathrm{ASF}}^{S}(x)$

		n = 300			n = 500		
$L_1 = L_2$	K	2SNSE	2SSE	2SHK	2SNSE	2SSE	2SHK
3	5	0.054	2.211	0.111	0.042	0.123	0.129
3	6	0.044	14.479	0.111	0.033	94.254	0.129
3	7	0.029	7.665	0.111	0.022	81.490	0.129
3	8	0.045	49.625	0.113	0.039	31.335	0.129
3	9	0.082	47.841	0.113	0.075	25.868	0.129
4	5	0.074	0.199	0.111	0.061	0.205	0.129
4	6	0.064	0.240	0.111	0.051	0.070	0.129
4	7	0.048	0.251	0.111	0.035	0.064	0.129
4	8	0.066	0.116	0.111	0.060	0.026	0.129
4	9	0.100	0.140	0.111	0.099	0.040	0.129
5	5	0.095	0.205	0.111	0.074	0.211	0.129
5	6	0.084	0.224	0.111	0.064	0.207	0.129
5	7	0.071	0.215	0.111	0.051	0.192	0.129
5	8	0.092	0.199	0.111	0.082	0.175	0.129
5	9	0.126	0.210	0.111	0.121	0.180	0.129
6	5	0.105	0.202	0.111	0.083	0.204	0.129
6	6	0.094	0.204	0.111	0.073	0.191	0.129
6	7	0.082	0.196	0.111	0.062	0.180	0.129
6	8	0.099	0.196	0.111	0.089	0.183	0.129
6	9	0.129	0.205	0.111	0.124	0.193	0.129
7	5	0.112	0.189	0.111	0.093	0.205	0.129
7	6	0.102	0.189	0.111	0.084	0.195	0.129
7	7	0.092	0.181	0.111	0.073	0.185	0.129
7	8	0.103	0.200	0.111	0.099	0.198	0.129
7	9	0.134	0.204	0.111	0.134	0.203	0.129
8	5	0.118	0.184	0.111	0.102	0.207	0.129
8	6	0.108	0.183	0.111	0.092	0.204	0.129
8	7	0.099	0.175	0.111	0.083	0.195	0.129
8	8	0.109	0.183	0.111	0.106	0.204	0.129
8	9	0.138	0.191	0.111	0.139	0.210	0.129

Table 3.4 Design 2: average sample mean square error for $\widehat{\mathrm{ASF}}^{S}(x)$

		n = 300			n = 500		
$L_1 = L_2$	K	2SNSE	2SSE	2SHK	2SNSE	2SSE	2SHK
4	6	0.412	0.664	1.737	0.363	0.659	1.741
4	7	0.382	0.613	1.737	0.324	0.580	1.741
4	8	0.358	1.148	1.737	0.296	1.060	1.741
4	9	0.517	0.137	1.737	0.394	0.113	1.741
5	6	0.415	0.558	1.737	0.406	0.578	1.741
5	7	0.401	0.515	1.737	0.378	0.526	1.741
5	8	0.385	0.920	1.737	0.357	0.925	1.741
5	9	0.534	0.097	1.737	0.494	0.103	1.741
6	6	0.419	0.516	1.737	0.406	0.532	1.741
6	7	0.403	0.494	1.737	0.389	0.489	1.741
6	8	0.392	0.903	1.737	0.375	0.861	1.741
6	9	0.538	0.097	1.737	0.518	0.094	1.741
7	6	0.446	0.522	1.737	0.417	0.512	1.741
7	7	0.435	0.504	1.737	0.406	0.485	1.741
7	8	0.422	0.874	1.737	0.391	0.837	1.741
7	9	0.574	0.105	1.737	0.542	0.095	1.741
8	6	0.459	0.520	1.737	0.415	0.499	1.741
8	7	0.452	0.513	1.737	0.405	0.478	1.741
8	8	0.439	0.871	1.737	0.391	0.814	1.741
8	9	0.593	0.114	1.737	0.536	0.094	1.741
9	6	0.449	0.496	1.737	0.422	0.495	1.741
9	7	0.440	0.493	1.737	0.414	0.481	1.741
9	8	0.432	0.834	1.737	0.398	0.812	1.741
9	9	0.578	0.104	1.737	0.542	0.096	1.741

Table 3.5 Design 3: average sample mean square error for $\widehat{\text{ASF}}^S(x)$

		n = 300			n = 500		
$L_1 = L_2$	K	2SNSE	2SSE	2SHK	2SNSE	2SSE	2SHK
3	5	0.209	0.208	0.021	0.234	0.183	0.025
3	6	0.223	2.277	0.021	0.250	39.618	0.025
3	7	0.244	2.278	0.021	0.281	28.057	0.025
3	8	0.272	7.302	0.021	0.311	10.512	0.025
4	5	0.146	0.008	0.021	0.139	0.003	0.025
4	6	0.154	0.004	0.021	0.153	0.005	0.025
4	7	0.166	0.005	0.021	0.173	0.007	0.025
4	8	0.173	0.001	0.021	0.191	0.0001	0.025
5	5	0.135	0.010	0.021	0.119	0.004	0.024
5	6	0.141	0.009	0.021	0.127	0.003	0.024
5	7	0.155	0.010	0.021	0.141	0.003	0.024
5	8	0.162	0.002	0.021	0.155	0.0003	0.024
6	5	0.135	0.011	0.021	0.111	0.005	0.024
6	6	0.137	0.011	0.021	0.117	0.004	0.024
6	7	0.149	0.012	0.021	0.128	0.005	0.024
6	8	0.154	0.002	0.021	0.140	0.00002	0.024
7	5	0.139	0.014	0.021	0.106	0.005	0.022
7	6	0.142	0.013	0.021	0.111	0.005	0.022
7	7	0.153	0.014	0.021	0.122	0.006	0.022
7	8	0.157	0.003	0.021	0.129	0.0003	0.022
8	5	0.143	0.016	0.021	0.110	0.007	0.022
8	6	0.145	0.015	0.021	0.114	0.006	0.022
8	7	0.153	0.015	0.021	0.122	0.007	0.022
8	8	0.158	0.004	0.021	0.128	0.001	0.022

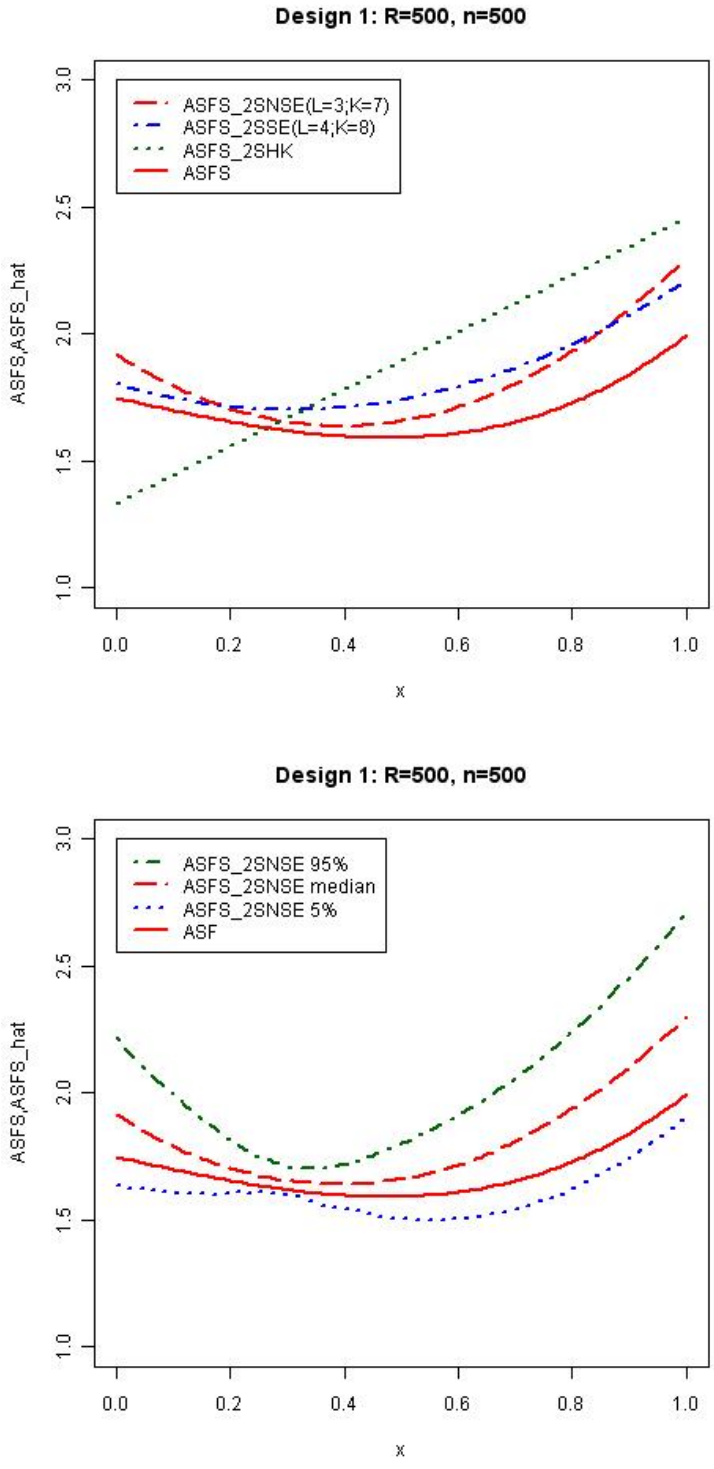

Figure 3.1 Estimates of $\mathrm{ASF}^S(x)$ in Design 1

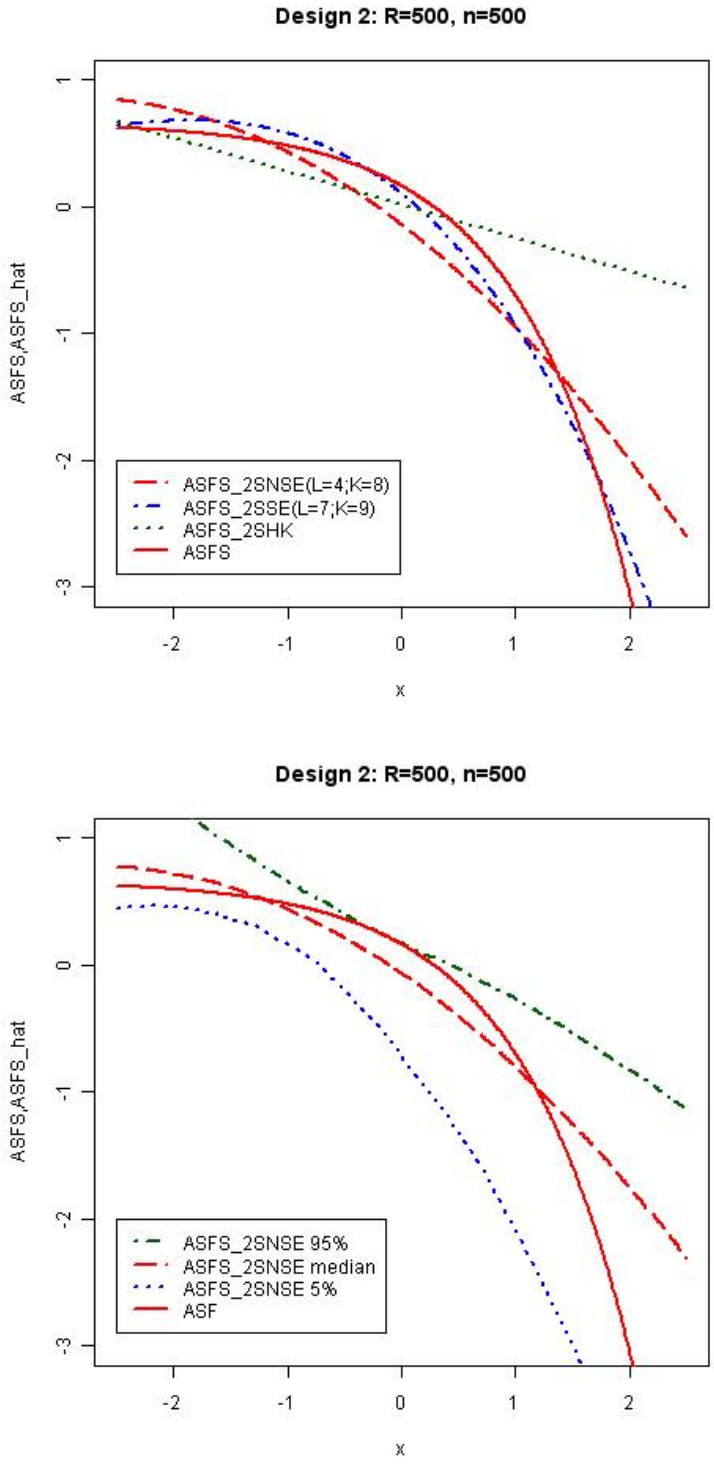

Figure 3.2 Estimates of $\text{ASF}^S(x)$ in Design 2 (with $(1, x, \widehat{P}, \widehat{V}, x^2, x\widehat{P}, \widehat{P}^2, \widehat{P}\widehat{V})$ for 2SNSE and $(1, x, \widehat{P}, \widehat{\eta}, x^2, \widehat{P}^2, \widehat{P}\widehat{\eta}, \widehat{\eta}^2, x^3)$ for 2SSE in the second stage)

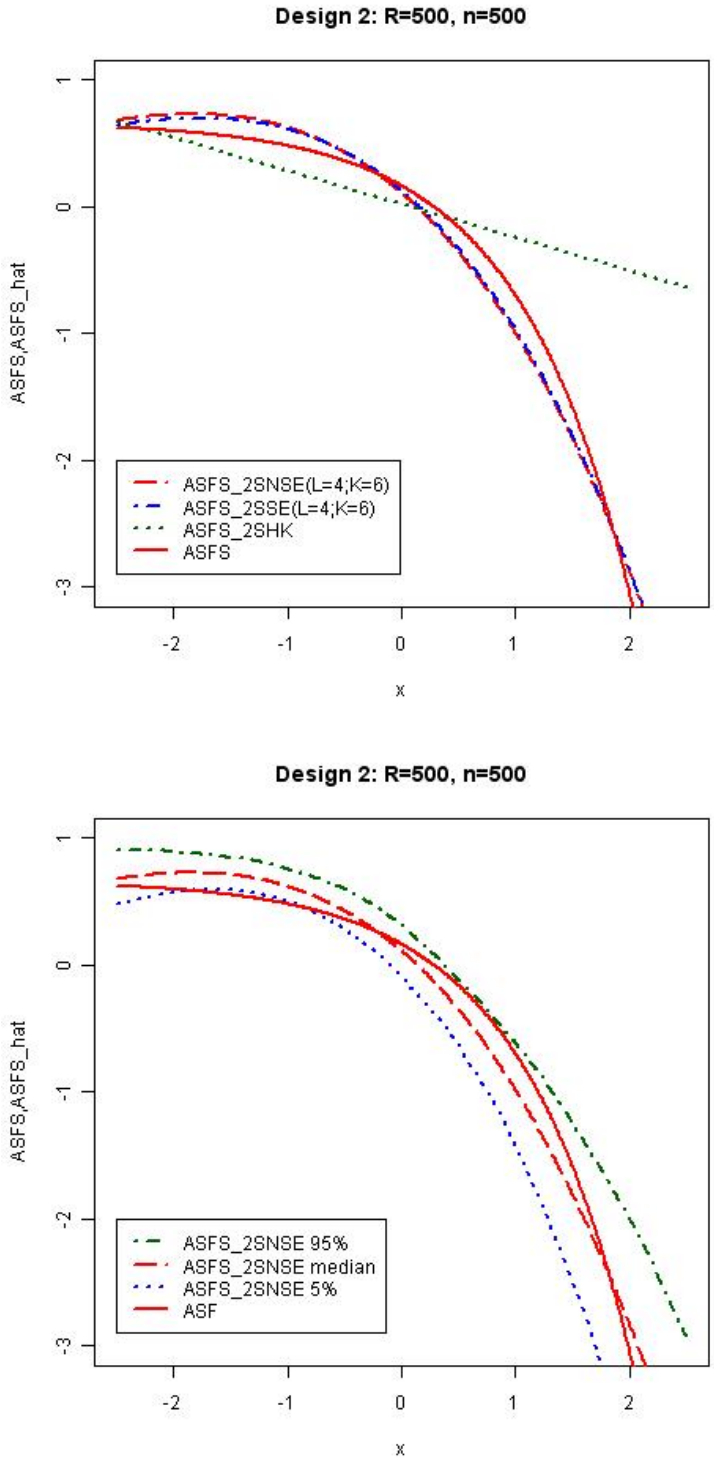

Figure 3.3 Estimates of $\text{ASF}^S(x)$ in Design 2 (with $(1, x, x^2, x^3, \widehat{P}, \widehat{V})$ for 2SNSE and $(1, x, x^2, x^3, \widehat{P}, \widehat{\eta})$ for 2SSE in the second stage)

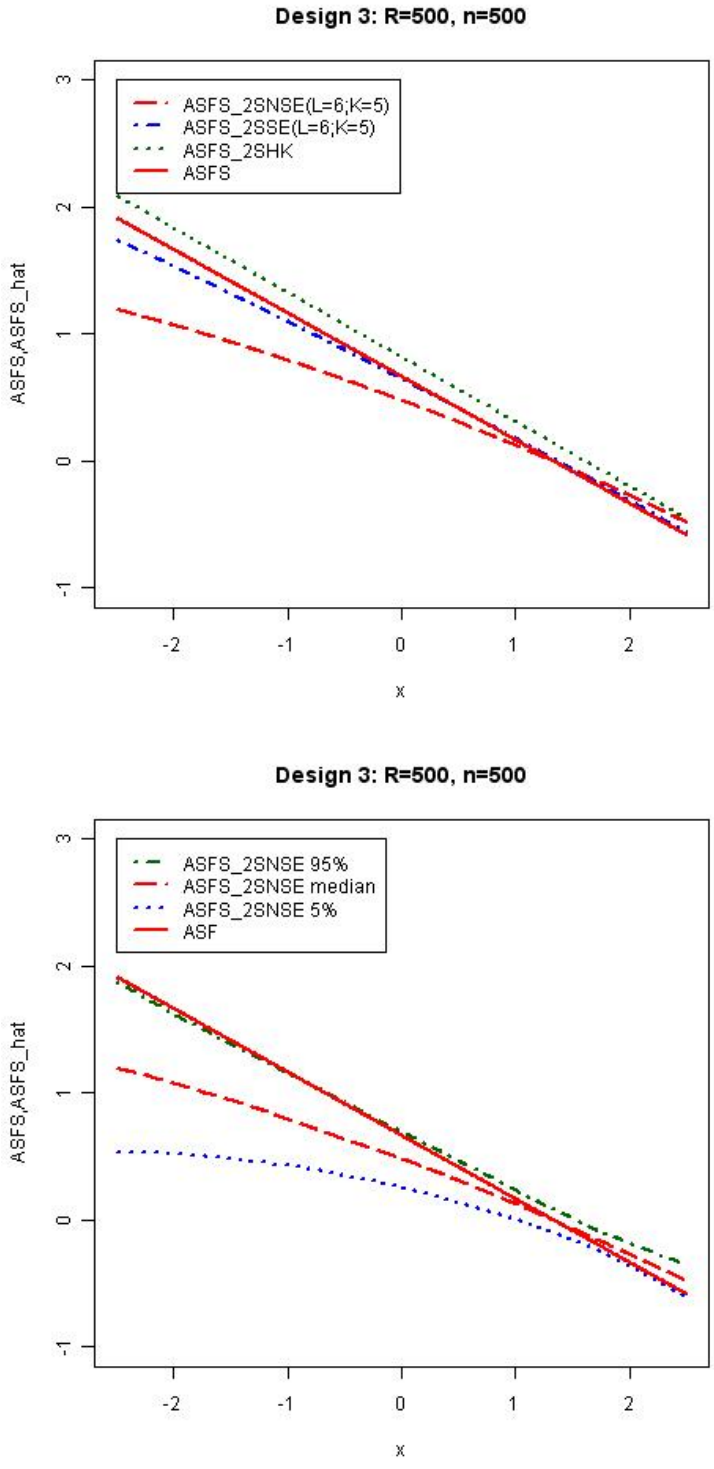

Figure 3.4 Estimates of $\mathrm{ASF}^S(x)$ in Design 3

Bibliography

Ahn, H., and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58(1-2), pp. 3–29.

Altonji, J. G., H. Ichimura, and T. Otsu (2012): "Estimating Derivatives in Nonseparable Models With Limited Dependent Variables," *Econometrica*, 80(4), 1701–1719.

Andrews, D. W. K., and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 65(3), pp. 497–517.

Angrist, J. D. (1997): "Conditional independence in sample selection models," *Economics Letters*, 54(2), pp. 103–112.

Arellano, M., and S. Bonhomme (2013): "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality," Working paper.

Blundell, R., and J. L. Horowitz (2007): "A Non-Parametric Test of Exogeneity," *The Review of Economic Studies*, 74(4), pp. 1035–1058.

Blundell, R., and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. by L. P. H. Mathias Dewatripont, and S. J. Turnovsky, pp. 312–357. Cambridge University Press.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, pp. 655–679.

Buchinsky, M. (1998): "The dynamics of changes in the female wage distribution in the USA: a quantile regression approach," *Journal of Applied Econometrics*, 13(1), pp. 1–30.

——— (2001): "Quantile regression with sample selection: Estimating women's return to education in the U.S," *Empirical Economics*, 26(1), pp. 87–113.

Chaudhuri, P. (1991): "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," *The Annals of Statistics*, 19(2), pp. 760–777.

Chaudhuri, P., K. Doksum, and A. Samarov (1997): "On Average Derivative Quantile Regression," *The Annals of Statistics*, 25(2), pp. 715–744.

Chen, S., and S. Khan (2003): "Semiparametric Estimation of a Heteroskedastic Sample Selection Model," *Econometric Theory*, 19(6), pp. 1040–1064.

Chernozhukov, V., I. Fernndez-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), pp. 245–261.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139(1), pp. 4–14.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), pp. 1405–1441.

——— (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73(5), pp. 1525–1550.

Cosslett, S. R. (1991): "Semiparametric Estimation of a Regression with Sample Selectivity," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by G. T. W.A. Barnett, James L. Powell, pp. 175–197. Cambridge University Press.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70(1), pp. 33–58.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76(5), pp. 1191–1206.

Gandhi, A., K. Kim, and A. Petrin (2012): "Identification and Estimation in Discrete Choice Demand Models when Endogenous Variables Interact with the Error," Working Paper 16894, National Bureau of Economic Research.

Heckman, J. J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80(2), pp. 313–18.

Hoderlein, S., and E. Mammen (2007): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," *Econometrica*, 75(5), pp. 1513–1518.

Ichimura, H., and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 74. Elsevier.

Imbens, G. W. (2006): "Nonadditive Models with Endogenous Regressors," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by N. Blundell, and Persson, pp. 17–46. Cambridge University Press.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), pp. 1481–1512.

Klein, R., C. Shen, and F. Vella (2011): "Semiparametric Selection Models with Binary Outcomes," IZA Discussion Papers 6008, Institute for the Study of Labor (IZA).

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76(3), pp. 1071–1102.

Lee, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141(2), pp. 1131–1158.

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71(5), pp. 1339–1375.

——— (2007): "Nonparametric identification," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 73. Elsevier.

Melly, B., and M. Huber (2012): "A Test of the Conditional Independence Assumption in Sample Selection Models," Working paper.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), pp. 147–168.

——— (2007): "Nonparametric Continuous/Discrete Choice Models," *International Economic Review*, 48(4), pp. 1429–1439.

——— (2009): "Two-step series estimation of sample selection models," *Econometrics Journal*, 12, S217–S229.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), pp. 565–603.

Pinkse, J. (2000): "Nonparametric Two-Step Regression Estimation When Regressors and Error Are Dependent," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28(2), pp. 289–300.

Powell, J. L. (2001): "Semiparametric Estimation of Censored Selection Models," in *Nonlinear statistical modeling*, ed. by J. L. P. Cheng Hsiao, Kimio Morimune, pp. 165–196. Cambridge University Press.

Rothe, C. (2009): "Semiparametric estimation of binary response models with endogenous regressors," *Journal of Econometrics*, 153(1), pp. 51–64.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10(4), pp. 1040–1053.