Essays On Discrete Choice Models

By

Wei Song

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2017

Date of final oral examination: 05/04/2017

The dissertation is approved by the following members of the Final Oral Committee:
Xiaoxia Shi, Assistant Professor, Economics
Bruce Hansen, Professor, Economics
Jack Porter, Professor, Economics
Joachim Freyberger, Assistant Professor, Economics
Chunming Zhang, Professor, Statistics

Abstract

This dissertation focuses on the identification and estimation of discrete choice models. In practice, if the error term is independent of the covariates and follows some known distribution, the discrete choice model is usually estimated using some parametric estimator, such as Probit and Logit. However, when the distribution of the error is unknown, misspecification would in general cause the estimators inconsistent even if the independence between the covariates and the error still holds. The following two chapters relax the assumptions on the error distribution in the discrete choice models and propose semiparametric estimators.

Chapter 1: "Least Square Estimation of Semiparametric Binary Response Model with Endogeneity"

In this chapter we develop new results on the identification and estimation of semiparametric binary response model with an endogenous explanatory variable. The identification is achieved based on a control variable approach. We also propose a semiparametric estimator, which is $\sqrt{n}$-consistent and asymptotically normal distributed. The estimation is based on a nonlinear least square criterion, which we show is equivalent to an integrated maximum score criterion. In literature there is still no result about whether nonlinear least square estimator would be dominated by other estimators in terms of efficiency regarding semiparametric binary response model with endogeneity. Therefore, we provide a model averaging estimator, which combines the least square estimator and the maximum likelihood estimator of Rothe (2009). Monte Carlo simulation shows the performance of our estimator is consistent with our theory in finite samples. We also apply our estimator to the study of the causal effect of economic conditions on civil conflicts as in Miguel et al. (2004). In their paper, they use two-stage least square to estimate the effect of economic conditions on civil conflicts. We re-estimate their model using our semiparametric least square estimator.

Chapter 2: "A Semiparametric Estimator for Binary Response Models with Endogenous Regressors"

This chapter proposes a new semiparametric estimator for the binary response model with endogenous explanatory variables. We assume a triangular structure and use the control variable approach to account for endogeneity. In order to identify the model, we construct a control variable and assume the error is quantile independent of the covariates given the control variable for a given quantile. This quantile independence assumption compared to the statistical independence is rather flexible in that it admits heteroskedasticity. The semiparametric series estimator in this chapter is an extension of Khan (2013) with control variables. It can estimate both the coefficients and the error distribution, and we prove this estimator is consistent and derive its convergence rate. In the Monte Carlo experiment, our estimator in general has smaller bias and standard deviation in comparison with the parametric two-stage Probit estimator for the binary response model with continuous endogenous regressors.

Acknowledgments

My deep gratitude goes first to Professor Xiaoxia Shi, who expertly guided me through my graduate education and who shared the excitement of six years of discovery. Her personal generosity helped make my time at UW-Madison enjoyable.

My appreciation also extends to Professor Jack Porter, Professor Bruce Hansen and Professor Joachim Freyberger. They have always been generous with their time, support, and honest advice at all stages of my research. Thanks also go to Professor Chao Fu and Professor Chris Taber whose encouragements have been especially valuable.

Above ground, I am indebted to my parents, whose value to me only grows with age. And finally, I acknowledge my dear friends Wanyi Chen, Junjie Guo, Xiao Xiao, Bo Zhang and Xiang Zhang. I have been fortunate to have them along during my graduate study.

To My Parents and To Him.

Contents

List of Tables

List of Figures

Chapter 1

Least Square Estimation of Semiparametric Binary Response Model with Endogeneity

1.1 Introduction

This chapter is concerned with the identification and estimation of a semiparametric binary response model with endogeneity. The binary response model we consider is represented in the form

$$Y = 1[X'\beta - u \geq 0],$$

where X is an observed vector of explanatory variables, and Y is an indicator of the event that the value of $X'\beta - u$ is non-negative. In addition, u is the unobserved error term, and β is the coefficient vector of our interest. If X and u are independent, and u follows a known distribution, this binary response model is usually estimated via some parametric estimation procedures, such as the standard Probit or Logit; see McFadden (1984) for a detailed survey. However, when the distribution of u is unknown, Probit or Logit can be misspecified and lead to inconsistent estimators even if the independence between X and u still holds. More generally, the independence between X and u may also fail.

In this chapter, we develop new results on identification and estimation of semiparametric binary response model with endogeneity, in which the conditional distribution of u given X is not parametrically specified and may depend on X. This chapter contributes to the literature by providing a new set of identification conditions which allows for bounded support of the regressor, and by proposing a semiparametric least square estimator and a model averaging estimator for the binary response model with endogeneity.

First, we extend the identification results under quantile independence in Manski (1975, 1985), to allow for bounded support of the regressors. In order to identify the coefficients in binary response model without endogeneity, there is a trade-off between the support requirement for regressors and the extent to which we specify the distribution of the error term; see Manski (1988). The identification conditions in Manski (1975, 1985) are rather weak compared to previous parametric estimators: the conditional distribution of u given X is median independent, i.e. $Med(u|X) = 0$. However, large (unbounded) support for at least one component of X is required to achieve uniform identification of the coefficients. In the first section of this chapter, we extend Manski's quantile independence identification strategy and assume independence between regressors X and error term u. Although this seems more strict than Manski's quantile independence assumption, we do not require the large support of any regressor.

Second, we introduce a new least square estimator for the semiparametric binary response model. Based on the identification results above, we integrate the maximum score criterion function under τth quantile independence over τ from 0 to 1. We show that the integral is a least square criterion. In the binary response model without endogeneity, this leads to the semiparametric single index estimator of Ichimura (1993).

Third, we extend the identification and estimation results above to the semiparametric binary response model with endogenous explanatory variables. In this chapter, we use the control variable approach of Blundell and Powell (2004). The control variable may be an observable variable or a variable constructed using instrumental variables. Our estimator allows both cases. This estimator extends Ichimura (1993)'s semiparametric single index estimator to incorporate endogenous regressors and is shown to be $\sqrt{n}$-consistent and asymptotic normal.

Fourth, we also propose a model averaging estimator in addition to the semiparametric least square estimator, in order to improve efficiency. The model averaging estimator is a weighted average of estimators from different models. Here we use two models: the preceding least square estimation and the semiparametric maximum likelihood estimation by Rothe (2009). Although in the exogenous case semiparametric maximum likelihood estimator is proved to achieve the semiparametric efficiency bound. There is little evidence that maximum likelihood estimators is more efficient than least square estimators when some of the explanatory variables are endogenous. In this case, we provide a linear combination of these two types of estimators and prove that it has smaller asymptotic variance than both least square and maximum likelihood estimators.

Several semiparametric estimators for the binary response model have been proposed in literature, such as the maximum score estimator (Manski (1975, 1985)), the smoothed maximum score estimator (Horowitz (1992)), the semiparametric maximum likelihood estimator (Klein and Spady (1993)), etc. When some regressors are endogenous, unlike the separable models, the parameters in the binary response model, are not generally identified under the standard independence assumption between the instruments and the error term u, see Blundell and Powell (2003), Chesher (2010), Chesher and Rosen (2013) and references therein. Instead, identification of the coefficients is achieved by the control variable approach. This approach has been used in parametric binary response models. For example, Smith and Blundell (1986) and Rivers and Vuong (1988) introduce a two-stage Probit estimator (2SProbit) for the binary response triangular system with continuous endogenous regressors. As for semiparametric binary response models with endogeneity, the control variable approach is first proposed by Blundell and Powell (2003, 2004). Blundell and Powell (2004) used the control variable to account for endogeneity in the semiparametric binary response model. The estimator they considered was based on the matching estimator proposed by Ahn, Ichimura, and Powell (1996). Rothe (2009) extended Klein and Spady (1993)'s estimator to the endogenous case by forming a triangular system and estimated it using a two-step semiparametric maximum likelihood method. Our model is closely related to Blundell and Powell (2004) and Rothe (2009), but the estimator is different from theirs.

In order to investigate finite sample performance, we provide Monte Carlo simulations and compare all the three estimators: Rothe's maximum likelihood estimator, least square estimator and the model averaging estimator. The results show the finite sample performance of our estimators is consistent with our theory.

We also apply our estimators to the study of the causal effect of economic conditions on civil conflicts. This example is taken from Miguel, Satyanath, and Sergenti (2004). In this application, rainfall variation is used as an instrument of economic conditions. In their chapter, they use two-stage least square to estimate a linear probability model. We re-estimate a semiparametric binary response model using our least square estimator. Contrary to their results, we verify that the effect of economic conditions on civil conflicts is smaller and not statistical significant.

The remainder of this chapter is organized as follows. In the next section, we show our identification conditions for the semiparametric binary response model with endogeneity. In Section 1.3, we describe our least square estimator based the identification result. In Section 1.4, we discuss asymptotic properties of our estimator and propose the model averaging estimator. In Section 1.5, we present the simulation results. In Section 1.6, we apply our estimator to an empirical application. Section 1.7 concludes.

1.2 The model and identification

We formalize a basic linear index threshold-crossing binary response model

$$Y_i = 1(X_i'\beta_0 - u_i \geq 0) \tag{1.1}$$

where Y is the binary dependent variable, $X_i = (X_{1i}, \tilde{X}_i) \in \mathbb{R}^{1+K}$ is the vector of regressors, u_i is an unobserved random error term, and $1(E)$ is the indicator function that equals 1 when E is true and 0 otherwise. Here we assume that X_i does not include the constant term. X_{1i} is a variable with continuous support. [1] The parameter of interest for this type of model is β_0, and the conditional distribution function of the error u given X, denoted as F_0, is a nuisance infinite dimensional parameter.

In this section, we discuss the identification conditions for binary response model without and with endogenous regressors, respectively.

1.2.1 Binary response model without endogenous regressors

In this subsection we introduce the identification result in a simplified setting, where the disturbance u and the regressors X are statistical independent without endogeneity. We do this to highlight the main strategy of our approach. In next subsection we develop identification results when some of the regressors are endogenous.

[1] Other than in special cases, a regressor whose support contains continuous part is necessary to point identify the parameters, and if all the regressors are discrete, β_0 would be set identified, unless a finite-dimensional parametric distribution of u_i is assumed; see Horowitz (2009). Komarova (2013) proposes a consistent estimators of the identified set using linear programming in a binary response model with all discrete regressors.

Conditional median independence with bounded regressors

To fix ideas, it might help to recall the identification results under conditional median independence proposed by Manski (1975, 1985):

$$Median(u \mid X) = c \tag{1.2}$$

where c is an unknown constant. Large support of at least one continuous regressor is required for uniform identification. In practice, though, it is not uncommon to have data sets where all the continuous regressors are not unbounded, such as income, consumption and commute time. In these cases, the large support requirement cannot be satisfied. But as followed, no large support does not necessarily indicate lack of point identification.

We first show a revised set of conditions under which β_0 is point identified in the binary response model with conditional median independence Equation (1.2) and the continuous regressors are allowed to have bounded support. This result resembles Corollary 3.1.1 in Horowitz (2009). We restate it here for completeness, and present a new proof in Appendix.

Assumption MID. (a) (X_i', Y_i) for $i = 1, \cdots, n$ is an i.i.d. sample and the data generating process follows Equation (1.1) and Equation (1.2);

(b) There exists a subset $\tilde{C}$ of $\mathbb{R}^K$, where the conditional support of X_1 given $\tilde{X} = \tilde{x}$ contains an interval $[a_{\tilde{x}}, b_{\tilde{x}}]$ for all $\tilde{x} \in \tilde{C}$;

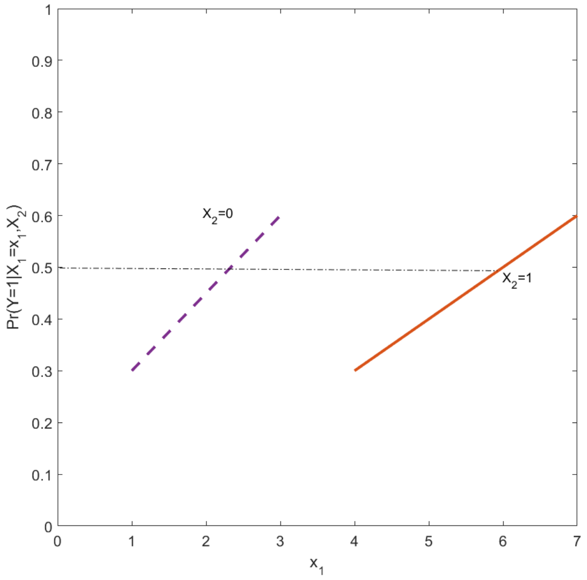
(c) $\sup_{x_1 \in [a_{\tilde{x}}, b_{\tilde{x}}]} \Pr(Y = 1 \mid X_1 = x_1, \tilde{X} = \tilde{x}) > \frac{1}{2}$ and $\inf_{x_1 \in [a_{\tilde{x}}, b_{\tilde{x}}]} \Pr(Y = 1 \mid X_1 = x_1, \tilde{X} = \tilde{x}) < \frac{1}{2}$ for all $\tilde{x} \in \tilde{C}$;

(d) $E[(\begin{matrix} 1 \\ \tilde{X} \end{matrix})(1, \tilde{X}')1(\tilde{X} \in \tilde{C})]$ has full rank;

(e) $\parallel \beta_0 \parallel = 1$.

Assumption MID(b) and Assumption MID(c) together imply that the support of $E(Y \mid X_1, \tilde{X} = \tilde{x})$ contains a neighborhood of 0.5. Assumption MID(d) ensures that there is a

Figure 1.1: Identified under Median Independence



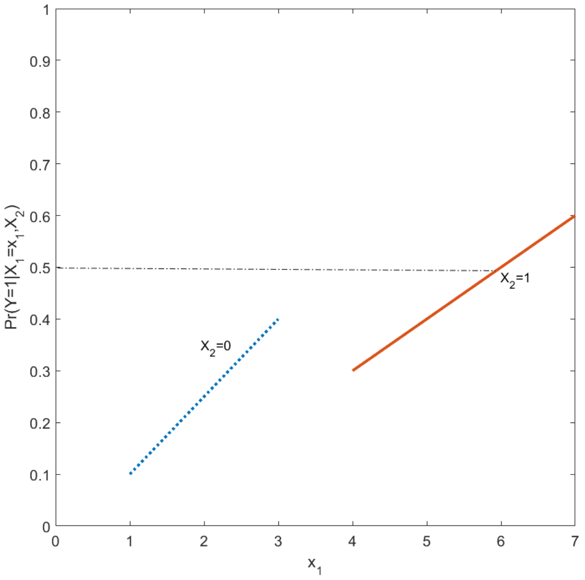
sufficiently rich set of values of $\tilde{X}$ at which Assumption MID(b) and Assumption MID(c) are satisfied. Assumption MID(e) is a standard normalization for binary response models.

Proposition 1.1. *Under Assumption MID, β_0 is identified.*

Figure 1 illustrates Assumptions MID(b) and (c) in a simple setting. In this example, X_1 is a continuous variable and X_2 is a dummy variable. The solid line shows $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 1)$ is between 0.3 and 0.6. If the dashed line represents $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 0)$, both the dashed and the solid lines cross 0.5 and according to Proposition 1, β_1 and β_2 are identified. In Figure 2 the dotted line stands for $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 0)$, however, it doesn't cross 0.5. Then β_1 and β_2 may not be point identified.

Intuitively, the identification under Assumption MID is achieved by a special group of observations. Individuals in this group share the property $E[Y \mid X] = 0.5$. Suppose c in Equation (1.2) is zero for simplicity. In this case, for this group, $X'\beta_0 = 0$, as illustrated by Assumption MID(c). As long as there are sufficiently many different value of X, demonstrated in Assumption MID(d), β_0 is uniquely determined with the normalization.

Figure 1.2: Unidentified under Median Independence



This heuristic example shows that under Manski's median independence assumption, identification is possible when regressors have bounded support, rather than large support. What is required is that $\Pr(Y = 1 \mid X_1, \tilde{X} = \tilde{x})$ has a density around 0.5 for a rich set of $\tilde{x}$ values. If not, point identification may still be impossible unless a more strict assumption about the error term distribution is imposed. In the next subsection, we propose new identification results which assume full independence between X and u while relaxing the support condition above.

Statistical independence with bounded regressors

If we assume that X and u are fully independent with each other, it implies u is quantile independent of X at every quantile. Therefore, even if the conditional choice probabilities given different $\tilde{X}$ values do not overlap around 0.5, we can still achieve identification of β_0 using the same strategy if they overlap around some quantile. Denote the cumulative distribution function of u as $F_0 \in \mathcal{F}$, where $\mathcal{F}$ is some space of distribution functions on $\mathbb{R}$.

Assumption IID. (a) X and u are independent;

(b) The set $\mathcal{F}$ is a family of continuous and strictly increasing distribution functions on $\mathbb{R}$;

(c) $\parallel \beta_0 \parallel = 1$;

(d) There exists a subset $\tilde{\mathcal{X}}$ of the support of $\tilde{X}$ and $0 < \underline{\lambda} < \bar{\lambda} < 1$. For any $\tilde{x} \in \tilde{\mathcal{X}}$, $[\underline{\lambda}, \bar{\lambda}] \subseteq supp(E(Y \mid X_1, \tilde{X} = \tilde{x}))$;

(e) $E[(\begin{matrix} 1 \\ \tilde{X} \end{matrix})(1, \tilde{X}')1(\tilde{X} \in \tilde{\mathcal{X}})]$ has full rank .

Assumption IID(b) is a standard restriction on error term distribution. Assumption IID(c) is the usual scale normalization for binary response models without a parametric error distribution. Assumption IID(d) is the key condition for point identification of β_0 in this chapter. It imposes the main restrictions on the conditional support of X_1 given $\tilde{X}$. It requires that for any quantile $\lambda \in (\underline{\lambda}, \bar{\lambda})$, we can find certain values of $\tilde{X} = \tilde{x}$ conditional on which the support of X_1 contains an interval in $\mathbb{R}$ with the property that $\Pr(Y = 1 \mid X_1, \tilde{X} = \tilde{x})$ contains some neighborhood of λ. Note that the support of X_1 under these assumptions can be bounded. The continuity of $\Pr(Y = 1 \mid X_1, \tilde{X})$ around λ provides information for identification of β_0. Note that $\underline{\lambda}$ and $\bar{\lambda}$ need not be known in our model. The size of the quantile set $(\underline{\lambda}, \bar{\lambda})$ can be very small. Intuitively once one quantile is found satisfying the support condition above, other quantiles close enough will generally possess the same property because of the continuous support of X_1. Assumption IID(e), together with Assumption IID(d), ensures that there are enough distinct values of $\tilde{X}$ such that conditional on each of them the support of the probability of $Y = 1$ contains some neighborhood of the same quantile λ_0. The intuition behind this condition is that if the distribution of the conditional probability of $Y = 1$ is dense over a neighborhood of a certain quantile, there will be sufficient information to identify the coefficients β_0 by arguments similar to the quantile independence analysis by Manski (1975), etc.

Figure 3 and Figure 4 use the same design as Figure 1 and Figure 2. In Figure 3, $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 0)$ (dashed line) and $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 1)$ (solid line) overlap between 0.3 and 0.4. As is illustrated in last subsection, if we impose conditional median

Figure 1.3: Identified under Independence

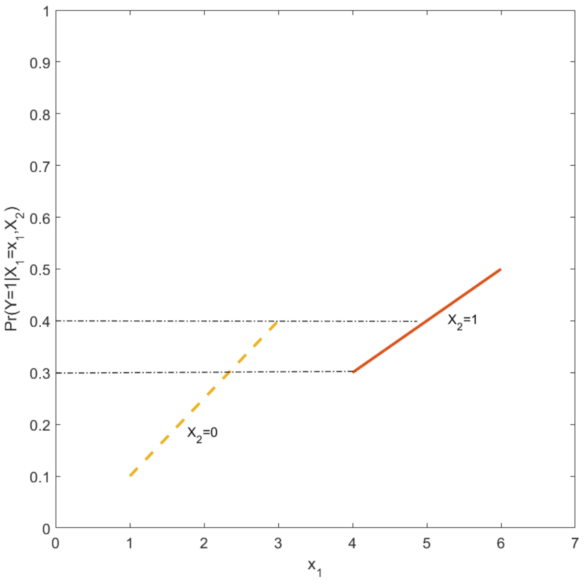
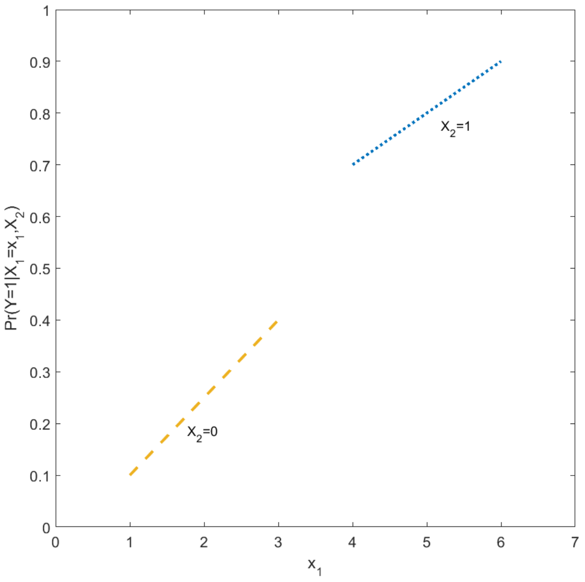


Figure 1.4: Unidentified under Independence



independence here, β_0 may not be point identified because the support of the conditional probabilities of $Y = 1$ is not rich enough to cover any neighborhood of 0.5. On the other hand, if we assume full independence between X and u, we can achieve point identification of β_0 using the rich variation of $E(Y \mid X)$ in $[0.3, 0.4]$. In Figure 4 $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 1)$ is the dotted line. However, identification of β_0 may still be problematic because $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 0)$ (dashed line) and $\Pr(Y = 1 \mid X_1 = x_1, X_2 = 1)$ (dotted line) don't overlap around any quantile.

Intuitively, Assumption IID ensures that there are a group of people whose support of $E(Y \mid X_1, X_2 = \tilde{x})$, for any $\tilde{x} \in \tilde{\mathcal{X}}$, contains an interval $[\underline{\lambda}, \bar{\lambda}]$. Fix $\tilde{x}$ and for any $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, there exists $x_{\tilde{x}}$ such that $E(Y \mid X_1 = x_{\tilde{x}}, \tilde{X} = \tilde{x}) = \lambda$. Therefore, for this group of people, $x_{\tilde{x}}\beta_{0,1} + \tilde{x}'\tilde{\beta}_0 = F_u^{-1}(\lambda)$. As long as there are sufficiently many different values of $\tilde{x}$, β_0 and $F_u^{-1}(\lambda)$ can be uniquely determined.

We now present our identification results based on the assumptions above. Define the population criterion function

$$Q(b, F) = E \int_0^1 \rho_\tau(Y - 1(b_1 X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) \geq 0))d\tau \tag{1.3}$$

where $\rho_\tau(u) = u(1(u > 0) - \tau)$. The check function $\rho_\tau(\cdot)$ is in fact the criterion of Manski's maximum score estimation under τth quantile independence. We will show that $\rho_\tau(Y - 1(b_1 X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) \geq 0)$ achieves its minimum value when $b = \beta_0$ for any $\tau \in (0, 1)$, and furthermore, the minimum is achieved uniquely when $b = \beta_0$ for any $\tau \in (\underline{\lambda}, \bar{\lambda})$. By taking the integral of the check function from 0 to 1 accordingly, $Q(\cdot)$ will achieve its minimum value only when $b = \beta_0$.

Theorem 1.2. *Under Assumption IID, for any $b \neq \beta_0$, we have $Q(b, F) > Q(\beta_0, F_0)$.*

By some algebra, we can simplify the function $Q(b, F)$. We present the equivalence result in the following lemma, which motives the least square estimator.

Lemma 1.3. $\int_0^1 \rho_\tau(Y - 1(b_1 X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) \geq 0))d\tau = \frac{1}{2}[Y - F(b_1 X_1 + \tilde{b}'\tilde{X})]^2$

This lemma shows that the integral of the maximum score criterion function over all the quantiles leads to the least square criterion function. In this least square criterion, there is an unknown function $F(\cdot)$ resembling the criterion function of the semiparametric single index model put forward by Ichimura (1993).

1.2.2 Binary response model with endogenous regressors

In this subsection, we extend the preceding identification analysis to binary response model with endogeneity.

In general, it is difficult to point identify β_0 without any further structural assumption when endogeneity arises. Unlike the linear regression model, in semiparametric nonlinear models, for example in the binary response model here, single-equation IV approach is often not enough to point identify the structural coefficients as pointed out by Chesher (2010) and reference therein. In order to solve this endogeneity problem, we utilize the control variable approach in this chapter. Following Imbens and Newey (2009), we assume that there exists a control variable V such that[2]

Assumption CID. (a) The variables X and u are independent conditional on V, i.e. $X \perp u \mid V$;

(b) The conditional distribution function of u given V, $F_{0,u|V}$, is continuous and strictly increasing on $\mathbb{R}$;

(c) $\parallel \beta_0 \parallel = 1$;

(d) There exists a subset $\mathscr{V}$ of the support of V with positive measure. For each $v \in \mathscr{V}$, there exists quantiles $0 < \underline{\lambda}_v < \bar{\lambda}_v < 1$ and a subset $\tilde{\mathscr{X}}_v$ of the conditional support of $\tilde{X}$ given $V = v$. For any $\tilde{x} \in \tilde{\mathscr{X}}_v$, $[\underline{\lambda}_v, \bar{\lambda}_v] \subseteq supp(E(Y \mid X_1, \tilde{X} = \tilde{x}))$;

(e) $E[(\begin{matrix} 1 \\ \tilde{X} \end{matrix})(1, \tilde{X}')1(\tilde{x} \in \tilde{\mathscr{X}}_v)]$ has full rank for any $v \in \mathscr{V}$.

[2] In this chapter for now, we assume that only one regressor, the continuous one, is endogenous for simplicity. Most of the conclusions in this chapter can be extended to the case in which more than one endogenous regressors arise.

A control variable may come from various sources. In some cases, it is an observable variable in a data set. Thus, we can directly use it in the estimator in the next section. In other cases, we need to construct such a control variable, typically from some instrumental variables Z. Now, we consider two examples of such construction. Suppose X_1 is the endogenous continuous variable and Z is an instrumental variable.

- Following Blundell and Powell (2004), we can assume $X_1 = h(Z) + V$, in which the error term V is additively separable from the instrument Z. In this case, we can use the error term V as control variable. The key assumption is a distributional exclusion restriction: $u \mid X, Z \sim u \mid X, V \sim u \mid V$.

- When the reduced form is not additively separable, $X_1 = h(Z, \eta)$, Imbens and Newey (2009) show that the CDF of η, $V = F_{X_1|Z}(X, Z) = F_\eta(\eta)$, can be used as the control variable under the following assumptions: (u, η) and Z are independent, the CDF of η is strictly increasing, and $h(Z, t)$ is strictly monotonic in t with probability 1.

With Assumption CID(a), we may use the same logic as in the last subsection to show that β_0 is identified with the help of the control variable V. Intuitively, it helps if we suppose V is discrete. Then the population can be divided into subgroups. Individuals within each subgroup share the same value of V, and u and X are independent. u follows some distribution indexed by V. Apply the theorem above to each subgroup, β_0 will be point identified as long as at least one subgroup satisfies the assumptions. Assumption CID(b) is standard and ensures that the τth quantile of the conditional distribution of u given V, $F^{-1}_{0,u|V}(\tau, V)$, is well defined. Assumption CID(c) is a commonly used normalization for discrete choice model. Assumption CID(d) imposes similar restrictions as Assumption IID(d) when endogeneity arises and some control variable is available. Assumption CID(e) is the rank condition similar as Assumption IID(e).

Theorem 1.4. *Under Assumption CID, for any $b \neq \beta_0$,*

$$Q(b, F_{u|V}) > Q(\beta_0, F_{0,u|V})$$

where

$$Q(b, F_{u|V}) = E \int \int_0^1 \rho_\tau(Y - 1(b'X - F^{-1}_{u|V=v}(\tau, v) \geq 0)) d\tau dF_V(v)$$

where the check function $\rho_\tau(u) = u(1(u > 0) - \tau)$.

By Lemma 1, $Q(b, F_{0,u|V}) = E \int \frac{1}{2}[Y - F_{u|V=v}(b'X)]^2 dF_V(v)$. Therefore, the criterion function in the endogenous case is an integral of all the least square criteria evaluated at each value of the control variable weighted by the density of the control variable. Its sample analogue is the sum of the least square criterion weighted by the empirical distribution of the control variable.

1.3 Estimation

In this part, we propose a semiparametric estimator of β_0 based on the preceding identification results.

We start with the case where the V is an observable control variable. If $F_{0,u|V}$ were known, we could use the nonlinear least squares method to $\sqrt{n}$-consistently estimate β_0 by minimizing

$$\tilde{S}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} [y_i - F_{0,u|V}(X_i'\beta, V_i)]^2$$

that is

$$\tilde{\beta} = \arg \min_{\beta \in \mathbb{R}^{1+K}} \tilde{S}_n(\beta)$$

However, the estimator $\tilde{\beta}$ is not feasible because the function $F_{0,u|V}$ is unknown. Thus we first need to estimate $F_{0,u|V}$ and replace it in the criterion function with its estimate $\hat{F}_{u|V}$.

Ichimura (1993) proposes to use the leave-one-out Nadaraya–Watson kernel estimator. In this chapter, we use the same strategy but with more than one index.

Specifically, we use the estimator

$$\hat{F}_{u|V}(u,v) = \frac{1}{nh_{1n}h_{2n}} \frac{\sum_{j\neq i} K_1(\frac{u-X_j'\beta}{h_{1n}})K_2(\frac{v-V_j}{h_{2n}})Y_i}{p(u,v)}$$

where K_1 and K_2 are kernel functions on $\mathbb{R}$, h_{1n} and h_{2n} are two bandwidth sequences that go to zero as n goes to infinity, and

$$p(u,v) = \sum_{j\neq i} K_1(\frac{u-X_j'\beta}{h_{1n}})K_2(\frac{v-V_j}{h_{2n}})$$

Therefore, in the case that the control variable V is observed, our semiparametric estimator for β_0 could be

$$\hat{\beta} = \arg \min_{\beta\in\mathbb{R}^{1+K}} \hat{S}_n(\beta)$$

where

$$\hat{S}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{F}_{0,u|V}(X_i'\beta, V_i)]^2$$

In many cases, V is not directly observed from data. As discussed in last section, we estimate V in a first stage regression. Let $\hat{V}$ be an estimated version of V. We plug $\hat{V}$ into the nonparametric estimator of $F_{0,u|V}$. That is

$$\hat{F}_{u|V}(u,v) = \frac{1}{nh_{1n}h_{2n}} \frac{\sum_{j\neq i} K_1(\frac{u-X_j'\beta}{h_{1n}})K_2(\frac{v-\hat{V}_j}{h_{2n}})Y_i}{p(u,v)}$$

where

$$p(u,v) = \sum_{j\neq i} K_1(\frac{u-X_j'\beta}{h_{1n}})K_2(\frac{v-\hat{V}_j}{h_{2n}}).$$

We finalize our semiparametric estimator of β_0 as

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{1+K}} \hat{S}_n(\beta)$$

where

$$\hat{S}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{F}_{0,u|V}(X_i' \beta, \hat{V}_i)]^2$$

1.3.1 Response Probability Estimation

In addition to the regression coefficients β_0, our semiparametric estimation procedure is also able to estimate other parameter of interest. The key object is the response probability given a set of X and how the probability changes if some X is modified exogenously, for example by policy maker, but keeping other X and the marginal distribution of u untouched. Blundell and Powell (2003, 2004) propose the average structural function (ASF) to measure this type of effect.

When all the regressors are exogenous, the marginal probability distribution of the error term u is F_0. Thus, the ASF of some given X equals $F_0(X'\beta_0)$. With the estimator of $\hat{\beta}$, we can estimate the index as $X'\hat{\beta}$. Then we can run a nonparametric regression of Y_i on $X_i\hat{\beta}$, and obtain the estimate for F_0, $\hat{F}$. The estimator for ASF would be $\hat{F}(X'\hat{\beta})$.

When some of the regressors are endogenous, the probability distribution of u also depends on the control variable V. Denote ASF as $G(X'\beta_0)$. In this case, $G(\cdot)$ can be identified as the partial mean of the distribution function $F_0(X'\beta_0, V_i)$ over the control variable V_i. That is

$$G(X'\beta_0) = \int F_0(X'\beta_0, V_i) dF(V_i)$$

In the same way, the index can be estimated as $X'\hat{\beta}$, and then we can obtain the estimate of the joint distribution function F_0 by a nonparametric regression of Y_i on $X_i\hat{\beta}$ and $\hat{V}_i$, and take a sample average over $\hat{V}_i$.

$$\hat{G}(X'\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}(X'\hat{\beta}, \hat{V}_i)$$

1.4 Asymptotic Properties

In this section, we investigate the asymptotic properties of the semiparametric estimator proposed above. We start with results on consistency and then derive the asymptotic normality. The assumptions needed for each property are listed in the corresponding subsection.

1.4.1 Consistency

Our semiparametric estimator of β_0 is consistent under the following assumptions.

Assumption CON. (a) The parameter space $\mathcal{B}$ of β_0 is a compact subset of $\mathbb{R}^{1+K}$ and β_0 is in its interior;
(b) Both kernel functions, $K_1(\cdot)$ and $K_2(\cdot)$, satisfy: $\int K(x)dx = 1$; $\int x^s K(x)dx = 0$, for $s = 1, \cdots, r-1$, for some $r \in \mathbb{N}$; $\int x^r K(x)dx < \infty$; $K(x)$ is r times continuously differentiable;
(c) The estimator $\hat{V}_i$ of V_i satisfies $\max_i |\hat{V}_i - V_i| = o_p(1)$.

Assumption CON(a) is a regularity condition and is standard in the literature. Assumption CON(b) if stronger than what we need to prove consistency, but will be needed for asymptotic normality. The consistency proof only requires that both $K_1(\cdot)$ and $K_2(\cdot)$ are twice differentiable and have bounded second derivatives. Assumption CON(c) requires that the estimator of the control variable converges to the true control variable uniformly in probability. This assumption can be easily satisfied by many estimators. For example, for the separable first stage case in Blundell and Powell (2004), a valid control variable is the residual of the conditional mean regression. As long as the conditional mean estimator is uniformly consistent, this control variable estimator satisfies our assumption. Later in the proof of asymptotic distribution, this assumption will get strengthened, where a convergence rate is assumed.

Theorem 1.5. *Under Assumption CID and Assumption CON,*

$$\hat{\beta} = \beta_0 + o_p(1)$$

as $n \to \infty$.

We now present the limiting distribution of our semiparametric estimator of β_0.

1.4.2 Asymptotic Normality

The limiting distribution of $\hat{\beta}$ is normal under additional assumptions. Because the control variable may be estimated if not directly observed, this generated regressor is included in our criterion function. The generated regressor constitutes the main difficulty for the proof of asymptotic normality. For that we use the general results on a class of semiparametric optimization estimators provided by Chen, Linton, and Van Keilegom (2003).

When V is constructed via a first stage regression, the construction defines a mapping $V = \phi(X, Z)$. Let $h \equiv (F, \phi)$ and $\hat{h} \equiv (\hat{F}, \hat{\phi})$. Denote the space of h as $\mathcal{H} = \mathcal{F} \times \Phi$, where $\mathcal{F}$ and Φ are the spaces for F and ϕ, respectively. Define the norm of space $\mathcal{H}$ as $\| h \|_{\mathcal{H}} = \max\{\| F \|_\infty, \| \phi \|_\infty\}$.

Assumption NOR. (a) The bandwidth satisfies $h_j = c_j n^{-\delta_j}$, $j = 1, 2$, for $1/(2r+1) \leq \delta_i < 1/4$;

(b) Let $p(\cdot, \cdot)$ be the density of $X'\beta$ and V. This density function is bounded below by a positive constant. Both F_0 and p are r times differentiable. The r-th derivatives are Lipschitz continuous uniformly over $\mathcal{B}$;

(c) The matrix $E[\frac{\partial F_0(X_i'\beta_0, V_i)}{\partial \beta} \frac{\partial F_0(X_i'\beta_0, V_i)}{\partial \beta'}]$ has full rank;

(d) The estimator $\hat{V}_i$ of V_i satisfies $\hat{V}_i - V_i = \frac{1}{n}\sum_{j=1}^{n} g_n(Z_i, Z_j)\Psi_j + r_{in}$with $max_i|r_{in}| = o_p(n^{-1/2})$, where Ψ_j is an influence function with $E(\Psi_j \mid Z_j) = 0$, $Var(\Psi_j^2 \mid Z_j) < \infty$ and $E(g_n(Z_i, Z_j)^2) = o(n)$;

(e) $\| \hat{\phi} - \phi \|_\infty = o_p(n^{-1/4})$;

(f) $\Pr(\hat{h} \in \mathcal{H}) \to 1$;

(g) The functional space $\mathcal{H}$ satisfies $\int_0^\infty \sqrt{logN(\lambda, \mathcal{H}, \| \cdot \|_{\mathcal{H}})} d\lambda < \infty$, where $N(\lambda, \mathcal{H}, \| \cdot \|_{\mathcal{H}})$ is the covering number with respect to the norm $\| \cdot \|_{\mathcal{H}}$ of the class $\mathcal{H}$. It is the minimal number of balls of $\| \cdot \|_{\mathcal{H}}$-radius λ needed to cover $\mathcal{H}$.

Assumption NOR(a) is used for reducing asymptotic bias of the estimate of F_0 and its derivatives using under smoothing technique. Assumption NOR(c) rules out the singularity of the asymptotic variance of our estimator. Assumption NOR(d), used also by Rothe (2009) in the maximum likelihood setting, requires the estimator of V_i to follow a certain asymptotic expansion. This condition is not restrictive. Many parametric and nonparametric estimators of the control variable fulfill it. For example, in a linear first stage, $X_i = Z_i'\gamma + V_i$, with $E(V \mid Z) = 0$. $\hat{V}_i$ is the residual of the linear regression, $\hat{V}_i = X_i - Z_i'\hat{\gamma}$. In this case, one can find $\Psi_i = -V_i$ and $g_n(Z_i, Z_j) = Z_i'(\frac{1}{n}\sum_{k=1}^{n} Z_k' Z_k)^{-1} Z_j$.

Assumption NOR(f) requires the nonparametric/parametric estimators belong to some well-behaved functional spaces with probability approaching 1. Assumption NOR(g) imposes entropy restrictions which are needed as one of the primitive conditions for stochastic equicontinuity in Chen, Linton, and Van Keilegom (2003). This type of assumption is widely used in semiparametric estimation studies; see, e.g, Linton, Sperlich, and Van Keilegom (2008). Many commonly used functional spaces can satisfy this requirement; for example, the Hölder ball defined in Van der Vaart and Wellner (1996)(p. 154).

Definition. Let $\underline{\alpha}$ be the largest integer smaller than α. Define for any vector $k = (k_1, \cdots, k_d)$ of d integers the differential operator $D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$, where $|k| = \sum k_i$. Then for a function $f : \mathcal{X} \mapsto \mathbb{R}$, let

$$\| f \|_{\alpha} = \max_{|k| \leq \underline{\alpha}} \sup_{x} |D^k f(x)| + \max_{|k| = \underline{\alpha}} \sup_{x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\| x - y \|^{\alpha - \underline{\alpha}}}$$

where the suprema are taken over all x, y in the interior of $\mathcal{X}$. Let $C^{\alpha}_{M}(\mathcal{X})$ be the set of all continuous function $f : \mathcal{X} \mapsto \mathbb{R}$ with $\| f \|_{\alpha} \leq M$.

If we assume that for some bounded $M > 0$ and $\alpha > \frac{1+dv}{2}$, $\mathcal{F} \subseteq C^{\alpha}_{M}(\mathbb{R}^{1+dv})$ and $\Pr(\hat{F} \in \mathcal{F}) \to 1$. This assumption restricts the space of conditional distribution for the error term given the control variable. It is not as strict as it seems and allows for many nonparametric estimators of F_0. By Theorem 2.7.1 in Van der Vaart and Wellner (1996), $\log N(\epsilon, C^{\alpha}_{1}(\mathcal{X}), \| \cdot \|_{\infty}) \leq const. \times (\frac{1}{\epsilon})^{dx/\alpha}$. Therefore, $\int_0^{\infty} \sqrt{\log N(\epsilon, \mathcal{F}, \| \cdot \|_{\infty})} d\epsilon < \infty$. As for the complexity of the space Φ, we take the linear first stage for example: $X_i = Z_i'\gamma + V_i$, with $E(V \mid Z) = 0$. In this case, $V_i = \phi(X_i, Z_i) = X_i - Z_i'\gamma$, the space of ϕ is indexed by γ. This space Φ is the "type I class" of Andrews (1994), which is manageable. Therefore, Φ satisfies Pollard's entropy condition, which leads to Assumption NOR(g).

The main results concerning the asymptotic distribution of our semiparametric estimator are given by the following theorem.

Theorem 1.6. *Under Assumption CID, Assumption CON and Assumption NOR,*

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightsquigarrow N(0, \Omega)$$

as $n \to \infty$, where

$$\Omega = \Sigma^{-1}(V_1 + V_2)\Sigma^{-1}$$

and

$$\Sigma = E[\frac{\partial F_0(x_i'\beta_0, v_{0i})}{\partial \beta} \frac{\partial F_0(x_i'\beta_0, v_{0i})}{\partial \beta'}]$$

$$V_1 = E[\frac{\partial F_0(x_i'\beta_0, v_{0i})}{\partial \beta} \frac{\partial F_0(x_i'\beta_0, v_{0i})}{\partial \beta'} \cdot F_0(x_i'\beta_0, v_{0i})[1 - F_0(x_i'\beta_0, v_{0i})]]$$

$$V_2 = E[\zeta\psi_i\psi_i'\zeta']$$

where $\zeta = E[g_n(Z, Z_i)\frac{\partial F_0(x'\beta_0,v_0)}{\partial\beta}\frac{\partial F_0(x'\beta_0,v_0)}{\partial v} \mid Z_i]$.

1.4.3 Estimation of the Covariance Matrix

In order to perform inference on the parameter, we need consistent estimates of the covariance matrix Ω. The first estimator is the plug-in estimator.

$$\hat{\Omega} = \hat{\Sigma}^{-1}(\hat{V}_1 + \hat{V}_2)\hat{\Sigma}^{-1}$$

where

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{F}(X_i'\hat{\beta},\hat{V}_i)}{\partial\beta}\frac{\partial\hat{F}(X_i'\hat{\beta},\hat{V}_i)}{\partial\beta'}$$

$$\hat{V}_1 = \frac{1}{n}\sum_{i=1}^{n}[\frac{\partial\hat{F}(X_i'\hat{\beta},\hat{V}_i)}{\partial\beta}\frac{\partial\hat{F}(X_i'\hat{\beta},\hat{V}_i)}{\partial\beta'}\hat{F}(X_i'\hat{\beta},\hat{V}_i)(1-\hat{F}(X_i'\hat{\beta},\hat{V}_i))]$$

$$\hat{V}_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\zeta}_i\hat{\psi}_i\hat{\psi}_i'\hat{\zeta}_i'$$

Here $\hat{\zeta}_i$ is a consistent estimator of $E[g_n(Z, Z_i)\frac{\partial F_0(x'\beta_0,v_0)}{\partial\beta}\frac{\partial F_0(x'\beta_0,v_0)}{\partial v} \mid Z_i]$. We are able to use several methods to construct $\hat{\zeta}_i$, for example, the fitted value of nonparametric kernel (or series) regression of $\hat{g}_n(Z, Z_i)\frac{\partial\hat{F}(X_i'\hat{\beta},\hat{V}_i)}{\partial\beta}\frac{\partial\hat{F}(X_i'\hat{\beta},\hat{V}_i)}{\partial v}$ on Z_i.

It is potentially difficult to directly calculate $\hat{\Sigma}, \hat{V}_1$ and $\hat{V}_2$ because the formulas have derivatives with respect to β and V involved. Chen, Linton, and Van Keilegom (2003) propose conditions under which the ordinary nonparametric bootstrap can consistently estimate the asymptotic distribution of $\sqrt{n}(\hat{\beta}-\beta_0)$. Let $\{X_i^*, Y_i^*, Z_i^*\}_{i=1}^n$ be drawn randomly with replacement from $\{X_i, Y_i, Z_i\}_{i=1}^n$, and let $\hat{h}^* = (\hat{F}^*, \hat{V}^*)$ be the same estimator as $\hat{h} = (\hat{F}, \hat{V})$ but based on the bootstrap data. Then the bootstrap estimator

$$\hat{\beta}^* = \arg\min_{\beta\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^{n}[Y_i^* - \hat{F}(X_i^{*'}\beta, \hat{V}^*)]^2$$

Theorem 1.7. *Under Assumption CID, Assumption CON and Assumption NOR,*

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \rightsquigarrow N(0, \Omega)$$

in P^*-probability.

This theorem indicates that the re-centered bootstrap estimator $\hat{\beta}^*$ has the same asymptotic distribution as $\hat{\beta}$, and therefore the ordinary nonparametric bootstrap procedure is valid to calculate the confidence regions for the unknown parameter β_0.

1.4.4 Model Averaging

In this part, we propose a model averaging estimator for the semiparametric binary response model. The model averaging estimator is a weighted average of estimators from different models. We use model averaging here in order to reduce asymptotic variance.

The two models we use here are the preceding semiparametric least square estimator, called Model 1, and the semiparametric maximum likelihood estimator introduced by Rothe (2009), called Model 2. Here we restate Rothe's estimator for completeness.

$$\hat{\beta}^2 = \arg\max_{\beta \in \mathcal{B}} L_n(\beta)$$

where

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i \log \hat{F}(X_i'\beta, \hat{V}_i)) + (1 - Y_i)\log(1 - \hat{F}(X_i'\beta, \hat{V}_i)).$$

Our model averaging estimator follows the form $\hat{\beta}^{MA} = \lambda\hat{\beta}^1 + (1-\lambda)\hat{\beta}^2$, where $\hat{\beta}^i$ is the estimator of β_0 based on Model i. The influence functions for $\hat{\beta}^1$ and $\hat{\beta}^2$ are

$$infl_i^1 = \Sigma^{-1}[(Y_i - F_0) \cdot \frac{\partial F_0}{\partial \beta} - E[\frac{\partial F_0}{\partial V}\frac{\partial F_0}{\partial \beta} g_n(Z, Z_i) \mid Z_i]]\Psi_i$$

$$infl_i^2 = \Sigma^{-1}[\frac{Y_i - F_0}{F_0 \cdot (1 - F_0)} \cdot \frac{\partial F_0}{\partial \beta} - E[\frac{\partial F_0}{\partial V}\frac{\partial F_0}{\partial \beta} \cdot \frac{g_n(Z, Z_i)}{F_0 \cdot (1 - F_0)} \mid Z_i]]\Psi_i$$

where $F_0 \equiv F_0(X_i'\beta_0, V_{i0})$.

Theorem 1.8. *Let $\sqrt{n}(\hat{\beta}^i - \beta_0) \rightsquigarrow N(0, \Omega_i)$, we have*

$$\sqrt{n}(\hat{\beta}^{MA} - \beta_0) \rightsquigarrow N(0, \Omega),$$

where $\Omega = \lambda^2\Omega_1 + (1-\lambda)^2\Omega_2 + \lambda(1-\lambda)(A + A')$, and $A = E(infl_i^1 \times infl_i^{2'})$;

To minimize the asymptotic variance of each component of $\hat{\beta}^{MA}$, We need to solve the minimization problem on the diagonal of Ω. Let $\hat{\beta}_k^{MA}$ be the kth dimension of $\hat{\beta}^{MA}$, $AVar(\hat{\beta}_k^i)^3$ kth element on the diagonal of Ω_i, and $Cov(\hat{\beta}_k^1, \hat{\beta}_k^2)$ the kth element on the diagonal of $A + A'$. For each k, we solve the following minimization problem

$$\min_{\lambda_k \in [0,1]} AVar(\hat{\beta}_k^{MA}) = \lambda_k^2 AVar(\hat{\beta}_k^1) + (1-\lambda_k)^2 AVar(\hat{\beta}_k^2) + \lambda_k(1-\lambda_k)Cov(\hat{\beta}_k^1, \hat{\beta}_k^2)$$

By simple algebra, we can get the optimal λ_k is $\lambda_k^* = \frac{AVar(\hat{\beta}_k^2) - Cov(\hat{\beta}_k^1, \hat{\beta}_k^2)}{AVar(\hat{\beta}_k^1) + AVar(\hat{\beta}_k^2) - 2Cov(\hat{\beta}_k^1, \hat{\beta}_k^2)}$.

Corollary 1.9. *When $\lambda_k = \lambda_k^*$, $AVar(\hat{\beta}_k^{MA}) \leq \min\{AVar(\hat{\beta}_k^1), AVar(\hat{\beta}_k^2)\}$.*

An estimator of λ_k^* can be formed by the variances of these two estimators and their covariance.

Estimation of the covariance between $\hat{\beta}^1$ and $\hat{\beta}^2$ may pose some obstacle here. One way to calculate this covariance, including the two variances, is bootstrapping. Let $\{(Y_i^*, X_i^*, Z_i^*)\}_{i=1}^n$ be drawn randomly with replacement from $\{(Y_i, X_i, Z_i)\}_{i=1}^n$. Let $\hat{\beta}^{1*}$ and $\hat{\beta}^{2*}$ be the minimizers of the criterion function based on the bootstrap data using least square and maximum likelihood, respectively. Then resample hundreds of times and we can get an estimator of the covariance between $\hat{\beta}^1$ and $\hat{\beta}^2$.

But bootstrapping is computationally intensive. if we have large data set or a complicated estimator this process can be extremely expensive. Another way to calculate the covariance is using influence function of each estimator. We calculate the estimate of the influence

[3] "$AVar$" means "asymptotic variance" in this chapter.

function for each estimator and for each observation, and stack them. Then we can calculate a consistent estimate of both the variance and the covariance.

1.5 Monte Carlo Simulation

In order to investigate the finite sample performance of our estimator, we provide small-scale Monte Carlo simulation in this part. In each experiment, we compare the estimation results of our semiparametric nonlinear least square estimator and the model averaging estimator with the semiparametric maximum likelihood estimator proposed by Rothe (2009). To play it fair, we adopt the same experiment setup as in Rothe (2009).

$$Y = I(X_1 + Z_1\beta_1 > u)$$

$$X_1 = \gamma_0 + Z_1\gamma_1 + Z_2\gamma_2 + Z_3\gamma_3 + V$$

where $\beta_1 = 1$[4] and $\gamma = (1, \frac{2}{3}, \frac{2}{3}, \frac{1}{3})'$. Z_1 follows exponential distribution truncated from above at 3 and have mean zero and variance two. Z_2 , Z_3 and V follow standard normal distribution. $u = \eta + V$. In different designs, we change the joint distribution of the error terms (u, V).

For this experiment, it is natural to use V as the control variable, because conditional on V, other components of X_1 are the instruments Z_1, Z_2 and Z_3, which are all independent of η. Therefore, $X_1 \perp u \mid V$. In this case, to correspond with Assumption NOR(d), $\hat{\Psi} = -\hat{V}$, and $\hat{g}_n(Z_i, Z_j) = Z_i'(\frac{1}{n}\sum_k Z_k Z_k')Z_j$.

We consider three data-generating designs by changing the distribution of η:

1. $\eta \sim N(0, 5)$

2. $\eta \sim 0.8N(-1, 0.6) + 0.2N(4, 2)$

[4] In order to facility comparison of simulation results, we use this normalization which is different from the one in Section 2. These two ways of normalization are equivalent as long as the sign of β_1 is positive.

3. $\eta \sim N(0, \exp(0.1 + 0.5 \times (X_1 + Z_1)))$

Three sample sizes are used for each design: N=250, 500 and 1000. And for each experiment we carry out 5000 replications. The bandwidths are chosen, following Hardle, Hall, Ichimura, et al. (1993) and Delecroix, Hristache, and Patilea (2006), jointly with the other parameters in the maximization. They are treated as additional parameters which is also the same as Rothe (2009).

The results of experiments for all three designs are summarized in Table 1-3. The true value of β is 1. "KS" corresponds to Rothe's estimator, "LS" our least square estimator, and "MA" our model averaging estimator. For each estimator, we report the mean bias, standard deviation, mean squared error, median bias and the coverage rate of the asymptotic and bootstrap confidence intervals with nominal level of 90%.

Table 1 illustrates the performance of our estimators under Design 1. In general, least square estimator and model averaging estimator have smaller biases and standard deviations than Rothe's semiparametric maximum likelihood estimator. In this design, among all the estimators, model averaging estimator performs best. The asymptotic confidence intervals are all below the nominal level 90%, but the bootstrap confidence intervals are all above the nominal level with proper lengths.

Table 2 shows the simulation results under Design 2. Least square estimator and model averaging estimator also have smaller biases than Rothe's. Meanwhile, the model averaging estimator has the smallest standard deviations across all the sample sizes. As for the confidence intervals, as sample size increases, the coverage rates of asymptotic intervals for least square and model averaging estimators are getting close the nominal level. The bootstrap confidence intervals all have coverage level above the nominal level.

Table 3 demonstrates the performance of our estimators under Design 3, the most complex one. Least square and model averaging estimators still have smaller mean biases and standard deviations than semiparametric maximum likelihood estimator. For n=250, model averaging estimator has smaller standard deviation than both estimators it takes averages. This is

Table 1.1: Design 1 $\eta \sim N(0,5)$

		Mean Bias	SD	MSE	MAD	Asymptotic CR	CI Length	Bootstrap CR	CI Length
n=250	KS	0.6683	1.1432	1.7531	0.3791	56.6%	0.94	99.6%	4.64
	LS	0.6444	1.1260	1.6829	0.3854	64.2%	0.98	99.9%	3.70
	MA	0.6142	1.1041	1.5960	0.3487	53.2%	0.70	100%	3.79
n=500	KS	0.5502	0.9308	1.1689	0.2952	59.7%	1.08	100%	4.01
	LS	0.5171	0.8988	1.0751	0.2960	70.3%	1.12	99.0%	3.37
	MA	0.4831	0.8667	0.9844	0.2656	59.4%	0.85	99.2%	3.38
n=1000	KS	0.3107	0.6151	0.4748	0.1856	63.9%	1.16	100%	3.06
	LS	0.2867	0.5756	0.4135	0.1709	78.9%	1.19	98.6%	2.88
	MA	0.2707	0.5657	0.3932	0.1577	68.3%	0.97	98.7%	2.72

Table 1.2: Design 2 $\eta \sim 0.8N(-1, 0.6) + 0.2N(4, 2)$

		Mean Bias	SD	MSE	MAD	Asymptotic CR	CI Length	Bootstrap CR	CI Length
n=250	KS	0.2583	0.5112	0.3280	0.1684	73.2%	0.81	99.8%	2.46
	LS	0.2219	0.5110	0.3104	0.1294	85.2%	0.95	98.6%	2.65
	MA	0.2244	0.4930	0.2933	0.1374	74.1%	0.64	98.2%	2.34
n=500	KS	0.1431	0.2904	0.1048	0.1085	77.1%	0.94	99.3%	1.74
	LS	0.1206	0.2960	0.1021	0.0826	91.7%	1.19	98.9%	2.26
	MA	0.1271	0.2889	0.0996	0.0921	83.1%	0.80	97.1%	1.74
n=1000	KS	0.0740	0.1780	0.0372	0.0622	79.2%	1.08	99.9%	1.32
	LS	0.0564	0.1778	0.0348	0.0490	97.2%	1.44	99.9%	1.94
	MA	0.0641	0.1761	0.0351	0.0546	90.1%	0.99	99.3%	1.37

Table 1.3: Design 3 $\eta \sim N(0, \exp(0.1 + 0.5 \times (X_1 + Z_1)))$

		Mean Bias	SD	MSE	MAD	Asymptotic CR	CI Length	Bootstrap CR	CI Length
n=250	KS	0.5152	0.9046	1.0836	0.3052	51.0%	0.73	99.3%	2.60
	LS	0.4372	0.8311	0.8817	0.2469	75.3%	0.91	99.7%	2.30
	MA	0.4364	0.8272	0.8746	0.2503	53.3%	0.51	99.3%	2.17
n=500	KS	0.3207	0.5236	0.3770	0.2334	54.7%	0.89	99.8%	2.05
	LS	0.2254	0.4481	0.2515	0.1513	87.3%	1.07	99.8%	1.94
	MA	0.2609	0.4717	0.2905	0.1839	65.1%	0.68	99.2%	1.69
n=1000	KS	0.1994	0.2879	0.1226	0.1692	57.4%	1.03	99.8%	1.47
	LS	0.1280	0.2572	0.0825	0.1038	93.5%	1.26	99.8%	1.62
	MA	0.1626	0.2713	0.1000	0.1351	72.6%	0.85	98.5%	1.24

consistent with our theory. For larger sample n=500 and n=1000, the “ordinary” least square estimator performs best.

In general, our least square and model averaging estimators perform better than the maximum likelihood estimator in these finite-sample experiments. In most cases, the standard deviations of model averaging estimators are the smallest amongst all the three estimators, which is consistent with our preceding theory about derivation of the model averaging estimators. We showed the coverage rates and interval lengths for both asymptotic and bootstrap confidence intervals at nominal level 90%. From all the three designs, we can conclude that compared to bootstrap, the asymptotic confidence intervals in general have coverage rates lower than the nominal rate. Therefore, in finite sample, bootstrap confidence interval is more reliable.

1.6 An Empirical Example: Economic Conditions and Civil Conflicts

In this section we apply our semiparametric binary response estimator to investigate the question whether economic conditions impact the likelihood of civil conflict. This example is taken from Miguel, Satyanath, and Sergenti (2004). Despite the fact that a large body of studies have shown the association between economic conditions and civil conflict, they argue that the existing literature does not adequately address the endogeneity of economic variables to civil war and therefore the causal relationship has not been established. One of the sources of the endogeneity can be omitted variable bias: fast-growing countries may be different from slow-growing counties in many government institutional qualities, which could drive both economic growth and civil conflicts, producing estimate bias.

In their chapter, Miguel et al. use variation in rainfall as an instrumental variable for income growth in order to solve the endogeneity issue and estimate the impact of economic growth on civil conflict. Rainfall variation is a plausible instrument for economic growth in economies that mostly rely on rain-fed agriculture.

Their key finding is GDP growth is significantly negatively related to the incidence of civil conflict. The estimation method they use with the instrument is two-stage least square (IV-2SLS), which is a linear probability model. The reason they explain why they do not use Rivers and Vuong (1988) method is that strong specification assumptions are required to justify IV-Probit. It is widely acknowledged that although parametric specification of the error terms in a binary response model can largely facilitate estimation, misspecification can make the estimators inconsistent. Therefore, we use our semiparametric estimation method to reanalyze the data.

Miguel et al. use data of civil conflict, rainfall and economic growth from 41 African countries during 1981-1999. These countries are all in the Sub-Saharan African region, and mostly non-industrialized countries without extensive irrigation systems, which is ideal for rainfall variation to be a valid instrument. The civil conflict data are from the Armed Conflict Data database developed by the International Peace Research Institute of Oslo, Norway, and the University of Uppsala, Sweden. This database records all conflicts with a threshold of 25 battle deaths per year. The information about rainfall variation is based on the Global Precipitation Climatology Project. This database includes both gauge and satellite data, avoiding systematic errors in gauge measures. Data on the other country characteristics are drawn mainly from Fearon and Laitin (2003).

The dependent variable is binary, the incidence of civil war in country i in year t ($conflict_{it}$). The countries with a civil conflict with at least 25 battle deaths per year are denoted ones, and other observations zeros. Economic growth is measured by GDP per capita growth rate of country i in year t ($growth_{it}$). Rainfall growth rate (ΔR_{it}) is the proportional change in rainfall from previous year. Other country characteristics (X_{it}) include logarithm of GDP per capita in year 1979, democracy in previous year, ethnolinguistic fractionalization, religious fractionalization, oil-exporting country, logarithm of mountainous, and logarithm of national population in previous year.

The model we want to estimate is

$$conflict_{it} = 1[X_{it}'\beta + \gamma_0 growth_{it} + \gamma_1 growth_{i,t-1} + \delta year_t \geq \epsilon_{it}]$$

In this model, $growth_{it}$ is considered endogenous, and rainfall variation is used as instrument to correct. As in Miguel, Satyanath, and Sergenti (2004), we assume the first stage equation is

$$growth_{it} = a + X_{it}'b + c_0 \Delta R_{it} + c_1 \Delta R_{i,t-1} + dyear_t + e_{it}$$

We estimate this first stage equation using ordinary least square, and use its residual e_{it} as a control variable. The variable e_{it} can be a valid control variable because it represents all the factors that are associated with economic growth except X_{it} and rainfall variation. Conditional on e_{it}, economic growth should be independent of ϵ_{it}.

We apply Probit, IV-Probit, and our semiparametric nonlinear least square estimators to estimate parameters in the equations above. Note that these estimation methods use different normalizations. In order to compare them, we report the average marginal effects in Table 4.

Column 1 shows the result of first stage estimate. It is simply OLS of the endogenous variable $growth_{it}$ on the instruments including those two rainfall variation variables ΔR_{it} and $\Delta R_{i,t-1}$. It shows that the relationship between economic growth and rainfall variation is positive, and this relationship is significant at over 95% confidence. The residual from this regression will be used as a control variable later in our semiparametric least square regression. Column 2 is the two-stage least square estimate.

Column 3 and 4 show the estimates using Probit without instrument and IV-Probit. When we do not consider endogenous economic growth, Probit regression shows a negative relationship between economic growth and the likelihood of civil conflict. This relationship

Table 1.4: Average Marginal Effects of Economic Growth on Civil Conflicts

Variables	First Stage	2SLS	Probit	IV-Probit	SLS	IV-SLS
	growth_t					
Economic Growth Rate, t		-0.528	-0.350*	-0.725	-0.332	-0.316
		(1.434)	(0.201)	(1.404)	(0.368)	(0.397)
Economic Growth Rate, t-1		-2.076**	-0.127	-0.126	-0.042	-0.069
		(1.030)	(0.198)	(0.198)	(0.110)	(0.103)
Log(GDP per capita), 1979	-0.002	-0.043	-0.063***	-0.063***	-0.011	-0.028**
	(0.002)	(0.049)	(0.023)	(0.023)	(0.014)	(0.014)
Democracy, t-1	0.000	0.003	0.001	0.001	-0.002	0.001
	(0.001)	(0.004)	(0.003)	(0.003)	(0.005)	(0.004)
Ethnolinguistic Fractionalization	0.001	0.226	0.220***	0.220***	0.105	0.099
	(0.013)	(0.277)	(0.081)	(0.087)	(0.069)	(0.062)
Religious Fractionalization	0.002	-0.236	-0.271***	-0.272***	-0.010	-0.098*
	(0.012)	(0.241)	(0.098)	(0.098)	(0.068)	(0.060)
Oil-Exporting Country	-0.007	0.044	0.015	0.013	0.056**	-0.006
	(0.004)	(0.214)	(0.050)	(0.051)	(0.028)	(0.027)
Log(mountainous)	0.000	0.077*	0.072***	0.072***	0.058**	0.111***
	(0.001)	(0.039)	(0.015)	(0.013)	(0.028)	(0.031)
Log(national population), t-1	-0.001	0.068*	0.074***	0.074***	0.056**	0.016
	(0.002)	(0..050)	(0.016)	(0.016)	(0.026)	(0.024)
Growth in Rainfall, t	0.055***					
	(0.016)					
Growth in Rainfall, t-1	0.034**					
	(0.014)					

Note: Standard errors are in parentheses. "SLS" is the semiparametric least square estimator without control variables. "IV-SLS" is the semiparametric least square estimator with control variable to correct endogeneity. * Significantly different from zero at 90 percent confidence; ** Significantly different from zero at 95 percent confidence; *** Significantly different from zero at 99 percent confidence.

is significant at 90% confidence with the current growth, but not the lagged one. When we consider economic growth may be endogenous, using IV-Probit is one option; see Smith and Blundell (1986) and Rivers and Vuong (1988). The result shows that although the average marginal effect of current economic growth doubles compared to Probit, it is not significant any more. The lagged growth is also not significant.

Column 5 and 6 show the estimates of our semiparametric least square estimator. Column 5 considers the current economic growth as an exogenous variable. The average marginal effect of current growth is not much different from the one of Probit, but because of its large standard error, it is not significant. And so is the lagged economic growth. The last column shows our semiparametric least square estimator for binary response model with endogeneity. Compared to the exogenous least square, the average marginal effects of both current economic growth and lagged economic growth do not change much. The standard errors of these two estimates are too large to conclude there is significantly negative relationship between economic growth and the likelihood of civil conflict. The estimates of other average marginal effects in these models are with the expected sign.

Miguel, Satyanath, and Sergenti (2004) use linear two-stage least square to estimate the above model and find growth is strongly negatively related to civil conflict. The reason why they choose the linear probability model instead of other available estimation methods is that they think specification assumptions are strong to justify IV-Probit. IV-Probit assumes the distribution of error terms is jointly normal distribution. That my be true, but linear probability models have their own flaws. For example, the fitted value using linear probability model may be outside the unit interval, and the marginal effect is restricted to be the same regardless of the initial value of the regressors. The semiparametric least square estimation proposed in this chapter fits this study better, because on one hand it is a nonlinear model, on the other hand it does not specify error term distribution parametrically. The results we found are that although economic growth is negatively related to the likelihood of civil conflict, the effect of growth is not significant.

1.7 Conclusions

This chapter proposes a semiparametric nonlinear least square estimator for the binary response model with an endogenous regressor. The estimator is an extension of semiparametric least square estimator for the single index model by Ichimura (1993) that accounts for regressor endogeneity. The identification is achieved based on a control variable approach. We also extend Manski's identification conditions for quantile independence and present a set of conditions that allow regressors with bounded support. Consistency and asymptotic normality are established for the estimator, and we prove the ordinary nonparametric bootstrap works for estimating the covariance matrix.

To the best of our knowledge the current available estimators for semiparametric binary response model with endogeneity via the control variable approach include the extension of matching estimator by Blundell and Powell (2004) and maximum likelihood estimator by Rothe (2009). There is little evidence of which estimator would dominate the others regarding efficiency. Therefore, we propose a model averaging estimator, which combines our semiparametric least square estimator and Rothe's maximum likelihood estimator, in order to improve efficiency. The Monte Carlo simulation demonstrates that in general our least square estimator and model averaging estimator behave better than Rothe's semiparametric maximum likelihood estimator. We apply the proposed least square estimator to study the effect of economic conditions on civil conflicts as in Miguel, Satyanath, and Sergenti (2004). The rainfall variation serves as the instrumental variable for economic growth here. The findings suggest smaller and insignificant effect of economic growth on the likelihood of civil conflicts in comparison with the two-stage least square estimates in Miguel, Satyanath, and Sergenti (2004).

References

Ahn, H., H. Ichimura, and J. L. Powell (1996): “Simple estimators for monotone index models,” *manuscript, Department of Economics, UC Berkeley.*

Andrews, D. W. (1994): “Empirical process methods in econometrics,” *Handbook of econometrics*, 4, 2247–2294.

Blundell, R., and J. L. Powell (2003): “Endogeneity in nonparametric and semiparametric regression models,” *ECONOMETRIC SOCIETY MONOGRAPHS*, 36, 312–357.

Blundell, R. W., and J. L. Powell (2004): “Endogeneity in semiparametric binary response models,” *The Review of Economic Studies*, 71(3), 655–679.

Chen, X., O. Linton, and I. Van Keilegom (2003): “Estimation of semiparametric models when the criterion function is not smooth,” *Econometrica*, 71(5), 1591–1608.

Chesher, A. (2010): “Instrumental variable models for discrete outcomes,” *Econometrica*, 78(2), 575–601.

Chesher, A., and A. M. Rosen (2013): “What Do Instrumental Variable Models Deliver with Discrete Dependent Variables?,” *American Economic Review*, 103(3), 557–62.

Delecroix, M., M. Hristache, and V. Patilea (2006): “On semiparametric M-estimation in single-index regression,” *Journal of Statistical Planning and Inference*, 136(3), 730–769.

Fearon, J. D., and D. D. Laitin (2003): “Ethnicity, insurgency, and civil war,” *American political science review*, 97(01), 75–90.

Hardle, W., P. Hall, H. Ichimura, et al. (1993): “Optimal smoothing in single-index models,” *The annals of Statistics*, 21(1), 157–178.

Horowitz, J. (2009): *Semiparametric and nonparametric methods in econometrics*, vol. 692. Springer.

Horowitz, J. L. (1992): “A smoothed maximum score estimator for the binary response model,” *Econometrica: Journal of the Econometric Society*, pp. 505–531.

Ichimura, H. (1993): “Semiparametric least squares (SLS) and weighted SLS estimation of single-index models,” *Journal of Econometrics*, 58(1), 71–120.

Imbens, G. W., and W. K. Newey (2009): “Identification and estimation of triangular simultaneous equations models without additivity,” *Econometrica*, 77(5), 1481–1512.

Klein, R. W., and R. H. Spady (1993): “An efficient semiparametric estimator for binary response models,” *Econometrica: Journal of the Econometric Society*, pp. 387–421.

Komarova, T. (2013): “Binary choice models with discrete regressors: Identification and misspecification,” *Journal of Econometrics*, 177(1), 14–33.

Linton, O., S. Sperlich, and I. Van Keilegom (2008): “Estimation of a semiparametric transformation model,” *The Annals of Statistics*, pp. 686–718.

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

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

——— (1988): “Identification of binary response models,” *Journal of the American Statistical Association*, 83(403), 729–738.

Masry, E. (1996): “Multivariate regression estimation local polynomial fitting for time series,” *Stochastic Processes and their Applications*, 65(1), 81–101.

McFadden, D. L. (1984): "Econometric analysis of qualitative response models," *Handbook of econometrics*, 2, 1395–1457.

Miguel, E., S. Satyanath, and E. Sergenti (2004): "Economic shocks and civil conflict: An instrumental variables approach," *Journal of political Economy*, 112(4), 725–753.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

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

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

Smith, R. J., and R. W. Blundell (1986): "An exogeneity test for a simultaneous equation Tobit model with an application to labor supply," *Econometrica: Journal of the Econometric Society*, pp. 679–685.

Van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes.* Springer, New York.

APPENDIX:

In this appendix, we provide the proofs of Theorem 1-6, Proposition 1, Corollary 1 and several lemmas used to prove the theorems. Proposition 1 restates similar results in Horowitz (2009), but present a different proof. Theorem 2 and 3 are the identification results, which serve as necessary conditions for consistency and asymptotic normality. Lemma 1 establishes an equivalence between integrated maximum score criterion and least square criterion and is the foundation of our estimation method. Theorem 3 proves consistency of our estimator via verifying the conditions in Newey and McFadden handbook chapter. Lemma 1-7 are needed for the proof of Theorem 4, the asymptotic distribution. Theorem 5 shows that bootstrap standard error is valid. Theorem 6 and Corollary 1 illustrate that model averaging estimator is asymptotically normal and there exists an optimal weight.

Proof of Proposition 1:

Proof. Let $Q(b,c) = E[|Y - 1(b_1X + \tilde{X}'\tilde{b} > c)|]$. By simple algebra,

$$Q(b,c) = E|Y - 1(b_1X + \tilde{X}'\tilde{b} > c)| = E[Y - (2Y-1)\cdot 1(b_1X + \tilde{X}'\tilde{b} > c)]$$

Define $Q(b,c \mid x,\tilde{x}) = E\{[Y - (2Y-1)\cdot 1(b_1X + \tilde{X}'\tilde{b} > c)] \mid X = x, \tilde{X} = \tilde{x}\}$. Therefore

$$\begin{aligned} Q(b,c) - Q(\beta_0,c_0) &= E[Q(b,c \mid X,\tilde{X}) - Q(\beta_0,c_0 \mid X,\tilde{X})] \\ &= E\{(2P(Y=1|X,\tilde{X}) - 1)[1(\beta_{0,1}X + \tilde{X}'\tilde{\beta}_0 > c_0) - 1(b_1X + \tilde{X}'\tilde{b} > c)]\} \end{aligned}$$

By assumption, $P(Y=1|X,\tilde{X}) > \frac{1}{2}$ if $\beta_{0,1}X + \tilde{X}'\tilde{\beta}_0 > c_0$ and $P(Y=1|X,\tilde{X}) \leq \frac{1}{2}$ if $\beta_{0,1}X + \tilde{X}'\tilde{\beta}_0 \leq c_0$. Therefore $Q(b,c \mid X,\tilde{X}) - Q(\beta_0,c_0 \mid X,\tilde{X}) \geq 0$ everywhere. Then it follows that $Q(b,c) - Q(\beta_0,c_0)$ is always non-negative, and $Q(b,c) - Q(\beta_0,c_0) = 0$ if $b = \beta_0$and $c = c_0$.

Next, we show that for any $b \neq \beta_0$, $\forall c, c_0$, such that $\parallel b \parallel = 1$, $Q(b,c) > Q(\beta_0,c_0)$, i.e. β_0 is the unique minimizer of $Q(b)$. Suppose it does not hold, i.e. $\exists b^* \neq \beta_0$ such that $\parallel b^* \parallel = 1$ and $Q(b^*,c^*) = Q(\beta_0,c_0)$, for some c^*. Since $Q(b,c|X,\tilde{X}) - Q(\beta_0,c_0|X,\tilde{X}) \geq 0$ and $Q(b,c) - Q(\beta_0,c_0) = E\{E[Q(b,c|X,\tilde{X}) - Q(\beta_0,c_0|X,\tilde{X}) \mid \tilde{X}]\}$, it follows that $E[Q(b^*,c^*|X,\tilde{X}) - Q(\beta_0,c_0|X,\tilde{X}) \mid \tilde{X}] = 0$

almost everywhere on the support of $\tilde{X}$. Define $Q(b,c|\tilde{X}) = E[Q(b,c|X,\tilde{X}) \mid \tilde{X}]$. Then it is necessary that $Q(b^*,c^*|\tilde{X}) = Q(\beta_0,c_0|\tilde{X})$ almost everywhere on $\tilde{C}$.

For any $\tilde{x} \in \tilde{C}$, by assumptions that the support of X conditional on $\tilde{X} = \tilde{x}$ contains an interval $[a_{\tilde{x}}, b_{\tilde{x}}]$ and that

$$\sup_{x \in [a_{\tilde{x}}, b_{\tilde{x}}]} \Pr(Y = 1 \mid X = x, \tilde{X} = \tilde{x}) > \frac{1}{2}$$

and

$$\inf_{x \in [a_{\tilde{x}}, b_{\tilde{x}}]} \Pr(Y = 1 \mid X = x, \tilde{X} = \tilde{x}) < \frac{1}{2},$$

there exists $x_{\tilde{x}}^0 \in [a_{\tilde{x}}, b_{\tilde{x}}]$ such that $\beta_{0,1} x_{\tilde{x}}^0 + \tilde{\beta_0}' \tilde{x} = c_0$. We claim that $x_{\tilde{x}}^0$ also satisfies $b_1^* x_{\tilde{x}}^0 + \tilde{b^*}' \tilde{x} = c^*$ almost everywhere on $\tilde{C}$. If not, without loss of generality, assume that $b_1^* x_{\tilde{x}}^0 + \tilde{b^*}' \tilde{x} > c^*$ and $\beta_{0,1} > 0$. Then there exits $\epsilon > 0$ such that for any $x_{\tilde{x}} \in (x_{\tilde{x}}^0 - \epsilon, x_{\tilde{x}}^0)$, $\beta_{0,1} x_{\tilde{x}} + \tilde{\beta}_0' \tilde{x} < c_0$ and $b_1^* x_{\tilde{x}} + \tilde{b^*}' \tilde{x} > c^*$, because of the continuity of X. In this case, $Q(b^*, c^*|\tilde{X} = \tilde{x}) \neq Q(\beta_0, c_0|\tilde{X} = \tilde{x})$, which can only happen for finite points on $\tilde{C}$.

Therefore, we have[5]

$$x_{\tilde{x}}^0 + \tilde{\beta}_0' \tilde{x} = c_0 \Rightarrow x_{\tilde{x}}^0 + \tilde{b^*}' \tilde{x} = c^* \Rightarrow (\tilde{\beta}_0 - \tilde{b}^*)\tilde{x} = c_0 - c^*$$

almost everywhere on $\tilde{C}$. So when $E[(\begin{array}{c} 1 \\ \tilde{x} \end{array})(1, \tilde{x}')1(\tilde{x} \in \tilde{C})]$ has full rank. β_0 is identified.

□

Proof of Theorem 1:

Proof. Let $b^\tau \equiv F^{-1}(\tau)$ and $\beta_0^\tau \equiv F_0^{-1}(\tau)$. Define

$$q(\tau, b, b^\tau) = E\rho_\tau(Y - 1(b_1 X_1 + \tilde{b}' \tilde{X} - b^\tau \geq 0)).$$

[5] Without loss of generality, we use $\beta_{0,1} = 1$ and $b_1^* = 1$ as normalization instead. This would simplify our proof and meanwhile this is equivalent to the unit Euclidean norm.

By dominated convergence theorem, the criterion function

$$\begin{aligned} Q(b,F) &= E\int_0^1 \rho_\tau(Y - 1(b'X - F^{-1}(\tau) \geq 0))d\tau \\ &= \int_0^1 E\rho_\tau(Y - 1(b'X - F^{-1}(\tau) \geq 0))d\tau \\ &= \int_0^1 q(\tau, b, b^\tau)d\tau \end{aligned}$$

It is sufficient to show that for any $(b, b^\tau) \neq (\beta, \beta_0^\tau)$, $q(\tau, b, b^\tau) \geq q(\tau, \beta, \beta_0^\tau)$ for $\tau \in (0,1)$ and $q(\tau, b, b^\tau) > q(\tau, \beta, \beta_0^\tau)$ for $\tau \in (\underline{\lambda}, \bar{\lambda})$.

$$q(\tau, b, b^\tau) = E\{E[\rho_\tau(Y - 1(b_1X_1 + \tilde{b}'\tilde{X} - b^\tau \geq 0)) \mid X_1, \tilde{X}]\}$$

From simple algebra, we can get

$$E[\rho_\tau(Y - 1(b_1X_1 + \tilde{b}'\tilde{X} - b^\tau \geq 0)) \mid X_1, \tilde{X}] = (1-\tau)P + 1(b_1X_1 + \tilde{b}'\tilde{X} - b^\tau \geq 0) \cdot (\tau - P)$$

where $P \equiv \Pr(Y = 1 \mid X_1, \tilde{X}) = E(Y = 1 \mid X_1, \tilde{X})$. By the independence between ϵ and $(X_1, \tilde{X})$, $P = F_0(\beta_1X_1 + \tilde{\beta}'\tilde{X})$. Since F_0 is continuous and strictly increasing, we have for any $\tau \in (0,1)$,

$$\beta_1X_1 + \tilde{\beta}'\tilde{X} \geq \beta_0^\tau \iff P \geq \tau \tag{1.4}$$

$$\beta_1X_1 + \tilde{\beta}'\tilde{X} < \beta_0^\tau \iff P < \tau \tag{1.5}$$

Therefore, $F^{-1}_{Y|X_1,\tilde{X}}(1 - \tau \mid X_1 = x_1, \tilde{X} = \tilde{x}) = 1(\beta_1x_1 + \tilde{\beta}'\tilde{x} - \beta_0^\tau \geq 0)$. For any $(b, b^\tau) \neq (\beta, \beta_0^\tau)$, $\tau \in (0,1)$,

$$\begin{aligned} E[\rho_\tau(Y - 1(b_1X_1 + \tilde{b}'\tilde{X} - b^\tau \geq 0)) \mid X_1, \tilde{X}] - E[\rho_\tau(Y - 1(\beta_1X_1 + \tilde{\beta}'\tilde{X} - \beta_0^\tau \geq 0)) \mid X_1, \tilde{X}] \\ = (\tau - P)[1(b_1X_1 + \tilde{b}'\tilde{X} - b^\tau \geq 0) - 1(\beta_1X_1 + \tilde{\beta}'\tilde{X} - \beta_0^\tau \geq 0)] \end{aligned}$$

which is always nonnegative and equals zero if $(b, b^\tau) = (\beta, \beta_0^\tau)$. So for any $\tau \in (0,1)$, $q(\tau, b, b^\tau) \geq q(\tau, \beta, \beta_0^\tau)$.

Consider some $\lambda \in (\underline{\lambda}, \bar{\lambda})$ in assumption, for any $\tilde{x} \in \tilde{\mathcal{X}}$, there exists $\bar{x}_1, \underline{x}_1 \in (a_{\tilde{x}}, b_{\tilde{x}})$ such that $\beta_1 \bar{x}_1 + \tilde{\beta}'\tilde{x} > \beta_0^\lambda$ and $\beta_1 \underline{x}_1 + \tilde{\beta}'\tilde{x} < \beta_0^\lambda$ because F_0 is continuous and strictly increasing by (1.4) and (1.5). Thus, there exits some $x_1^* \in (a_{\tilde{x}}^\lambda, b_{\tilde{x}}^\lambda)$ such that $\beta_1 x_1^* + \tilde{\beta}'\tilde{x} = \beta_0^\lambda$ by Assumption IID(d).

Suppose there exists $(b, b^\lambda) \neq (\beta, \beta_0^\lambda)$ such that $q(\lambda, b, b^\lambda) = q(\lambda, \beta, \beta_0^\lambda)$. Then we must have $1(b_1 X_1 + \tilde{b}'\tilde{X} - b^\lambda \geq 0) = 1(\beta_1 X_1 + \tilde{\beta}'\tilde{X} - \beta_0^\lambda \geq 0)$ almost surely on $(a_{\tilde{x}}, b_{\tilde{x}})$ given $\tilde{X} = \tilde{x}$. We claim that $b_1 x_1^* + \tilde{b}'\tilde{x} = b^\lambda$ must also hold. If not, by the continuity of the conditional support of X_1 given $\tilde{X} = \tilde{x}$, it is easy to find either a left or a right neighborhood of x_1^*, $N_\delta(x_1^*)$, such that$1(b_1 X_1 + \tilde{b}'\tilde{x} - b^\lambda \geq 0) \neq 1(\beta_1 X_1 + \tilde{\beta}'\tilde{x} - \beta_0^\lambda \geq 0)$ for any $X_1 \in N_\delta(x_1^*) \cap (a_{\tilde{x}}, b_{\tilde{x}})$. Therefore, $q(\tau, b, b^\tau) \neq q(\tau, \beta, \beta_0^\tau)$. For example, suppose $b_1 x_1^* + \tilde{b}'\tilde{x} > b^\lambda$, without loss of generality assume $\beta_1 > 0$ and $b_1 > 0$. Let $\delta = \frac{b^\lambda - \tilde{b}'\tilde{x}}{b_1}$ and it is obvious that for any $X_1 \in (x_1^* - \delta, x_1^*) \cap (a_{\tilde{x}}, b_{\tilde{x}})$, we have $\beta_1 X_1 + \tilde{\beta}'\tilde{x} - \beta_0^\lambda < 0$ and $b_1 X_1 + \tilde{b}'\tilde{x} - b^\lambda > 0$ simultaneously.

Now we have both $\beta_1 x_1^* + \tilde{\beta}'\tilde{x} = \beta_0^\lambda$ and $b_1 x_1^* + \tilde{b}'\tilde{x} = b^\lambda$. Since $\tilde{x}$ is arbitrary and by Assumption IID(e), $(b, b^\lambda) = (\beta, \beta_0^\lambda)$.

□

Proof of Theorem 2:

Proof. For any $(b_1, \tilde{b}) \neq (\beta_1, \tilde{\beta})$, we have proved in Theorem 2

$$E\rho_\tau(Y - 1(b'X - F^{-1}(\tau, v) \geq 0)) \geq E\rho_\tau(Y - 1(\beta'X - F^{-1}(\tau, v) \geq 0))$$

for any $\tau \in (0, 1)$ and $v \in supp(V)$. Therefore, for any $v \in supp(V)$,

$$E\int_0^1 \rho_\tau(Y - 1(b'X - F^{-1}(\tau, v) \geq 0))d\tau \geq E\int_0^1 \rho_\tau(Y - 1(\beta'X - F^{-1}(\tau, v) \geq 0))d\tau$$

For $v' \in \mathcal{V}$, with Assumption CID(d) and Assumption CID(e), use similar argument in the proof of Theorem 2 and we get

$$E\int_0^1 \rho_\tau(Y - 1(b'X - F^{-1}(\tau, v') \geq 0))d\tau > E\int_0^1 \rho_\tau(Y - 1(\beta'X - F^{-1}(\tau, v') \geq 0))d\tau$$

Thus, because $\mathcal{V}$ has a positive measure,

$$E\int_{\mathcal{V}}\int_0^1 \rho_\tau(Y-1(b'X-F^{-1}(\tau,v')\geq 0))d\tau dF_V(v)$$
$$>E\int_{\mathcal{V}}\int_0^1 \rho_\tau(Y-1(\beta'X-F^{-1}(\tau,v')\geq 0))d\tau dF_V(v)$$

In addition,

$$E\int\int_0^1 \rho_\tau(Y-1(b'X-F^{-1}(\tau,v')\geq 0))d\tau dF_V(v)$$
$$>E\int\int_0^1 \rho_\tau(Y-1(\beta'X-F^{-1}(\tau,v')\geq 0))d\tau dF_V(v)$$

□

Proof of Lemma 1:

Proof. Left hand side

$$\begin{aligned}\rho_\tau(Y-1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)) = &[Y-1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)]\\ &\cdot[1\{Y-1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)\}-\tau]\\ =&[Y-1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)]\\ &\cdot[1\{Y=1\}1\{1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)=0\}-\tau]\\ =&[Y-1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)]\\ &\cdot[1\{Y=1\}1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)<0)-\tau]\end{aligned}$$

When $Y=1$, left hand side equals

$$\int_0^1[1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)<0)-\tau+\tau\cdot 1(b_1X_1+\tilde{b}'\tilde{X}-F^{-1}(\tau)\geq 0)]d\tau$$

where $\int_0^1 1(b_1X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) < 0)d\tau = \int_{F(b_1X_1+\tilde{b}'\tilde{X})}^1 1d\tau = 1 - F(b_1X_1 + \tilde{b}'\tilde{X})$. Then left hand equals

$$\begin{aligned}
&1 - F(b_1X_1 + \tilde{b}'\tilde{X}) + \int_0^1 [-\tau + \tau \cdot 1(b_1X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) \geq 0)]d\tau \\
=&\frac{1}{2} - F(b_1X_1 + \tilde{b}'\tilde{X}) + \int_0^1 [\tau \cdot 1(b_1X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) \geq 0]d\tau \\
=&\frac{1}{2} - F(b_1X_1 + \tilde{b}'\tilde{X}) + \frac{1}{2}F(b_1X_1 + \tilde{b}'\tilde{X})^2
\end{aligned}$$

which equals right hand side.

When $Y = 0$, left hand side equals

$$\int_0^1 [1(b_1X_1 + \tilde{b}'\tilde{X} - F^{-1}(\tau) \geq 0) \cdot \tau]d\tau = \frac{1}{2}F(b_1X_1 + \tilde{b}'\tilde{X})^2$$

which equals right hand side.

□

Proof of Theorem 3:

Proof. We follow Therorem 2.1 in Newey and McFadden (1994) handbook chapter to prove the consistency of $\hat{\beta}$. It requires that

(i) $S_0(\beta)$ is uniquely maximized at β_0. This part can be proved by the identification results above.

(ii) The parameter space is compact. This is satisfied by the assumption $\mathcal{B}$ is compact.

(iii) $S_0(\beta)$ is continuous. $S_0(\beta) = E[Y_i - F_0(X_i'\beta, v_{i0})]^2$. The continuity of S_0 regarding β holds because we assume F_0 is a continuous CDF.

(iv)$\sup_{\beta \in \mathcal{B}} |\hat{S}_n(\beta) - S_0(\beta)| \to 0$. Uniform convergence of the criterion function. First, define an infeasible criterion function

$$\tilde{S}_n(\beta) = \frac{1}{n}\sum_{i=1}^n [Y_i - F_0(X_i'\beta, V_{i0})]^2$$

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}|\hat{S}_n(\beta)-S_0(\beta)| &= \sup_{\beta\in\mathcal{B}}|\hat{S}_n(\beta)-\tilde{S}_n(\beta)+\tilde{S}_n(\beta)-S_0(\beta)| \\ &\leq \sup_{\beta\in\mathcal{B}}|\hat{S}_n(\beta)-\tilde{S}_n(\beta)|+\sup_{\beta\in\mathcal{B}}|\tilde{S}_n(\beta)-S_0(\beta)|\end{aligned}$$

First term $|\hat{S}_n(\beta)-\tilde{S}_n(\beta)| = \frac{1}{n}\sum_{i=1}^{n}[2Y_i-\hat{F}(X_i'\beta,\hat{V}_i)-F_0(X_i'\beta,V_{i0})][F_0(X_i'\beta,V_{i0})-\hat{F}(X_i'\beta,\hat{V}_i)]$, therefore

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}|\hat{S}_n(\beta)-\tilde{S}_n(\beta)| &\leq 2\sup_{\beta\in\mathcal{B}}[F_0(X_i'\beta,V_{i0})-\hat{F}(X_i'\beta,\hat{V}_i)] \\ &\leq 2\sup_{\beta\in\mathcal{B}}|F_0(X_i'\beta,V_{i0})-F_0(X_i'\beta,\hat{V}_i)|+2\sup_{\beta\in\mathcal{B}}|F_0(X_i'\beta,\hat{V}_i)-\hat{F}(X_i'\beta,\hat{V}_i)|\end{aligned}$$

$\sup_{\beta\in\mathcal{B}}|F_0(X_i'\beta,\hat{V}_i)-\hat{F}(X_i'\beta,\hat{V}_i)| = o_p(1)$ because of the uniform convergence of kernel estimator. $\sup_{\beta\in\mathcal{B}}|F_0(X_i'\beta,V_{i0})-F_0(X_i'\beta,\hat{V}_i)| = o_p(1)$ because differentiability of F_0 and $max_i|\hat{V}_i-V_{0i}| = o_p(1)$.

Second term, by uniform law of large number, $\sup_{\beta\in\mathcal{B}}|\tilde{S}_n(\beta)-S_0(\beta)| = o_p(1)$.

Therefore, $\sup_{\beta\in\mathcal{B}}|\hat{S}_n(\beta)-S_0(\beta)| = o_p(1)$.

□

The proof of asymptotic normality is basically done by verifying the conditions given in Theorem 2 in Chen, Linton, and Van Keilegom (2003). In their chapter, they start with moment conditions. So we first need to form the corresponding first order condition of our criterion function. Let

$$M(\beta,h(\cdot,\beta)) = E[(Y_i-F(X_i'\beta,V_i))\cdot\frac{\partial F(X_i'\beta,V_i)}{\partial\beta}]$$

where $h\equiv(F,V)$. Also, $M(\beta,h_0)=0$ at $\beta=\beta_0$. Define $M_n(\beta,h)=\frac{1}{n}\sum_{i=1}^{n}[(Y_i-F(X_i'\beta,V_i))\cdot\frac{\partial F(X_i'\beta,V_i)}{\partial\beta}]$.

Lemma 1.10. $\|M_n(\hat{\beta},\hat{h})\| = inf_{\beta\in\mathcal{B}}\|M_n(\beta,\hat{h})\|+o_p(1/\sqrt{n})$

Proof. Because $M_n(\hat{\beta},\hat{h})$ is the partial derivative of $\hat{S}_n$ with respect to β. By first order condition of optimization, $\|M_n(\hat{\beta},\hat{h})\|=0$, which obviously satisfies this condition.

□

Lemma 1.11. *The ordinary derivative $\Gamma_1(\beta, h_0) \equiv \frac{\partial M(\beta,h_0)}{\partial \beta}$ in β of $M(\beta, h_0)$ exists for a neighborhood of β_0, and is continuous at $\beta = \beta_0$.*

Proof.

$$\begin{aligned}\Gamma_1(\beta, h_0) = & E[(Y_i - F_0(X_i'\beta, V_{i0}))\frac{\partial^2 F_0(X_i'\beta, V_{i0})}{\partial \beta \partial \beta'} - \frac{\partial F_0(X_i'\beta, V_{i0})}{\partial \beta}\frac{\partial F_0(X_i'\beta, V_{i0})}{\partial \beta'}] \\ = & - E\frac{\partial F_0(X_i'\beta, V_{i0})}{\partial \beta}\frac{\partial F_0(X_i'\beta, V_{i0})}{\partial \beta'}\end{aligned}$$

This lemma is then proved with Assumption NOR(b).

□

Lemma 1.12. *The matrix $\Gamma_1 \equiv \Gamma_1(\beta_0, h_0)$ is of full rank.*

Proof. This directly follows Assumption NOR(c).

□

Lemma 1.13. *For all $\beta \in \mathbb{B}$ the pathwise derivative $\Gamma_2(\beta, h_0)[h - h_0]$ of $M(\beta, h_0)$ exists in all directions $[h - h_0]$ and for all (β, h) with a positive sequence $\delta_n = o(1)$: (i) $\|M(\beta, h) - M(\beta, h_0) - \Gamma_2(\beta, h_0)[h - h_0]\| \leq c\|h - h_0\|_{\mathcal{H}}$ for a constant $c \geq 0$; (ii) $\|\Gamma_2(\beta, h_0)[h - h_0] - \Gamma_2(\beta_0, h_0)[h - h_0]\| \leq o(1)\delta_n$.*

Proof. We first calculate the pathwise derivatives. By definition,

$$\Gamma_2(\beta, h)[\bar{h} - h] = \lim_{\tau \to 0} \frac{M(\beta, h + \tau(\bar{h} - h)) - M(\beta, h)}{\tau}$$

We obtain that

$$\begin{aligned}\Gamma_2(\beta, h)[\bar{h} - h] = E[& -\frac{\partial F(X_i'\beta, V_i)}{\partial V_i}\frac{\partial F(X_i'\beta, V_i)}{\partial \beta}[\bar{V} - V] \\ & -\frac{\partial F(X_i'\beta, V_i)}{\partial \beta}[\bar{F} - F] \\ & -(Y_i - F(X_i'\beta, V_i))\frac{\partial^2 F(X_i'\beta, V_i)}{\partial \beta \partial V_i}[\bar{V} - V]\end{aligned}$$

Since $E(Y \mid X'\beta, V) = F(X'\beta, V)$, using law of iterated exceptions on the pathwise derivative above, we can get

$$\Gamma(\beta_0, h_0)[h - h_0] = E[-\frac{\partial F(X_i'\beta, V_i)}{\partial V_i}\frac{\partial F(X_i'\beta, V_i)}{\partial \beta}[V - V_0] - \frac{\partial F(X_i'\beta, V_i)}{\partial \beta}[F - F_0]$$

Because our function F satisfy a Lipschitz property, the two inequalities holds.

□

Lemma 1.14. *$\hat{h} \in \mathcal{H}$ with probability tending to one and $\|\hat{h} - h_0\|_{\mathcal{H}} = o_p(n^{-1/4})$.*

Proof. $\|\hat{V} - V_0\| = o_p(n^{-1/4})$ comes directly from Assumption NOR(e).

$\| \hat{F}(X'\beta, \hat{V}) - F(X'\beta, V) \| \leq \| \hat{F}(\cdot, \hat{V}) - \hat{F}(\cdot, V) \| + \| \hat{F}(\cdot, V) - F(\cdot, V) \|$. Consider $\| \hat{F}(\cdot, \hat{V}) - \hat{F}(\cdot, V) \|$ and $\| \hat{F}(\cdot, V) - F(\cdot, V) \|$, respectively.

$$\begin{aligned}\| \hat{F}(X'\beta, \hat{V}) - \hat{F}(X'\beta, V) \| =& \| \frac{\frac{1}{n}\sum_{i=1}^n K_1(\frac{X'\beta - X_i'\beta}{h_1})K_2(\frac{\hat{V}-\hat{V}_i}{h_2})Y_i}{\frac{1}{n}\sum_{i=1}^n K_1(\frac{X'\beta - X_i'\beta}{h_1})K_2(\frac{\hat{V}-\hat{V}_i}{h_2})} \\ &- \frac{\frac{1}{n}\sum_{i=1}^n K_1(\frac{X'\beta - X_i'\beta}{h_1})K_2(\frac{V-V_i}{h_2})Y_i}{\frac{1}{n}\sum_{i=1}^n K_1(\frac{X'\beta - X_i'\beta}{h_1})K_2(\frac{V-V_i}{h_2})} \| \\ =& \| \frac{\partial \hat{F}}{\partial V}(\hat{V} - V) \| + o_p(n^{-1/4})\end{aligned}$$

Denote $F = \frac{m}{f}$ and $\hat{F} = \frac{\hat{m}}{\hat{f}}$. Then $\frac{\partial \hat{F}}{\partial V} = \frac{\frac{\partial \hat{m}}{\partial V}f - \frac{\partial \hat{f}}{\partial V}m}{\hat{f}^2}$. Here we only consider$\frac{\partial \hat{f}}{\partial V}m(\hat{V} - V)$, the other term can be treated in the same way.

$$\begin{aligned}\frac{\partial \hat{f}}{\partial V}m(\hat{V} - V) &= \frac{1}{nh_2^{2dv}}\sum_{i=1}^n K_1\frac{\partial K_2}{\partial V}\sum_{j=1}^n Y_jK_1K_2 \cdot (\hat{V} - V) \\ &= \frac{1}{n^2h_2^{dv}}\sum_{i=1}^n K_1\frac{\partial K_2}{\partial V}\sum_{j=1}^n Y_jK_1K_2 \cdot \frac{1}{n}\sum_{k=1}^n g_n(Z, Z_k)\Psi_k + o_p(n^{-1/2}) \\ &= \frac{1}{n^3h_2^{dv}}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n K_{1i}\frac{\partial K_{2i}}{\partial V}Y_jK_{1j}K_{2j}g_n(Z, Z_k)\Psi_k + o_p(n^{-1/2}) \\ &= O_p(n^{-1/2}h^{-dv})\end{aligned}$$

By Assumption NOR(a), $h = cn^{-\delta}$, $O_p(n^{-1/2}h^{-dv}) = O_p(n^{dv\delta - 1/2})$. $dv\delta - \frac{1}{2} < -\frac{1}{4}$ then $\delta < \frac{1}{4dv}$.

Now turn to $\| \hat{F}(\cdot, V) - F(\cdot, V) \|$, by Masry (1996), the uniform rate of convergence of constant kernel regression is

$$\sup_x |\hat{F}(x) - F(x)| = O(\frac{(\ln n)^{1/2}}{(nh_1 \cdots h_q)^{1/2}} + \sum_{s=1}^{q} h_s^2) \ \ almost\, surely$$

Therefore, $\| \hat{F}(X'\beta, V) - F(X'\beta, V) \| = O_p(\frac{(\ln n)^{1/2}}{(nh^{1+dv})^{1/2}} + dv \cdot h^2) = O_p(\frac{(\ln n)^{1/2}}{(nn^{-\delta dv - \delta})^{1/2}} + dv \cdot n^{-2\delta})$. Let $n^{-2\delta} < n^{-1/4}$, $\delta > \frac{1}{8}$. Let $\frac{(\ln n)^{1/2}}{(nn^{-\delta dv - \delta})^{1/2}} > n^{-1/4}$, first $\frac{(\ln n)^{1/2}}{(nn^{-\delta dv - \delta})^{1/2}} > \frac{1}{n^{\frac{1}{2} - \frac{1}{2}\delta dv - \delta}}$ for enough large n ($n > e$), then $\frac{1}{n^{\frac{1}{2} - \frac{1}{2}\delta dv - \frac{1}{2}\delta}} > n^{-1/4}$, $\delta < \frac{1}{2dv+1}$. As stated in Assumption NOR(a).

□

Lemma 1.15. *For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$*

$$\sup_{\|\beta - \beta_0\| \le \delta_n, \|h - h_0\|_{\mathcal{H}} \le \delta_n} \|M_n(\beta, h) - M(\beta, h) - M_n(\beta_0, h_0)\| = o_p(n^{-1/2})$$

Proof. Theorem 3 in Chen, Linton, and Van Keilegom (2003) propose primitive conditions for stochastic equicontinuity. We use it here to prove this condition. First, we need to prove that

$$E[\sup_{\|\beta' - \beta\| \le \delta, \|h' - h\|_{\mathcal{H}} \le \delta} |(Y_i - F'(X_i'\beta', V_i')) \cdot \frac{\partial F'(X_i'\beta', V_i')}{\partial \beta} - (Y_i - F(X_i'\beta, V_i)) \cdot \frac{\partial F(X_i'\beta, V_i)}{\partial \beta}|^2]^{1/2} \le K\delta$$

for all $(\beta, h) \in \mathcal{B} \times \mathcal{H}$, all $\delta > 0$ and for some $K > 0$. This follows from the mean value theorem and the differentiability of $(Y_i - F(X_i'\beta, V_i)) \cdot \frac{\partial F(X_i'\beta, V_i)}{\partial \beta}$. Denote $m(\beta, F, V) = (Y_i - F(X_i'\beta, V_i)) \cdot \frac{\partial F(X_i'\beta, V_i)}{\partial \beta}$

$$\begin{aligned}
&|(Y_i - F'(X_i'\beta', V_i')) \cdot \frac{\partial F'(X_i'\beta', V_i')}{\partial \beta} - (Y_i - F(X_i'\beta, V_i)) \cdot \frac{\partial F(X_i'\beta, V_i)}{\partial \beta}| \\
=&|m(\beta', F', V') - m(\beta, F, V)| \\
\le&|m(\beta', F', V') - m(\beta, F', V')| + |m(\beta, F', V') - m(\beta, F, V')| + |m(\beta, F, V') - m(\beta, F, V)|
\end{aligned}$$

And then

$$
\begin{aligned}
&|m(\beta', F', V') - m(\beta, F, V)|^2 \\
\leq &|m(\beta', F', V') - m(\beta, F', V')|^2 + |m(\beta, F', V') - m(\beta, F, V')|^2 + |m(\beta, F, V') - m(\beta, F, V)|^2
\end{aligned}
$$

The first term

$$
\begin{aligned}
\sup_{\|\beta'-\beta\|\leq\delta} |m(\beta', F', V') - m(\beta, F', V')|^2 = &\sup_{\|\beta'-\beta\|\leq\delta} |\frac{\partial m(\tilde{\beta}, F', V')}{\partial \beta}(\beta' - \beta)|^2 \\
\leq &\delta^2 \cdot \sup_{\|\beta'-\beta\|\leq\delta} |\frac{\partial m(\tilde{\beta}, F', V')}{\partial \beta}|^2
\end{aligned}
$$

$$
\frac{\partial m}{\partial \beta} = -\frac{\partial F}{\partial \beta}\frac{\partial F}{\partial \beta'} + (Y - F)\frac{\partial^2 F}{\partial \beta \partial \beta'}
$$

is bounded almost for every X because the density of F is bounded by assumption, thus $E\sup_{\|\beta'-\beta\|\leq\delta} |m(\beta', F', V')-$ $m(\beta, F', V')|^2 \leq K_1\delta^2$ for some $K_1 > 0$.

The second term

$$
\begin{aligned}
&\sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} |m(\beta, F', V') - m(\beta, F, V')|^2 \\
= &\sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} |(Y_i - F'(X_i'\beta, V_i')) \cdot \frac{\partial F'(X_i'\beta, V_i')}{\partial \beta} - (Y_i - F(X_i'\beta, V_i')) \cdot \frac{\partial F(X_i'\beta, V_i')}{\partial \beta}|^2 \\
\leq &\sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} |(Y_i - F + \frac{\delta}{2}) \cdot \frac{\partial F}{\partial \beta} - (Y_i - F) \cdot \frac{\partial F}{\partial \beta}|^2 \\
= &\sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} \frac{\delta^2}{4}|\frac{\partial F}{\partial \beta}|^2
\end{aligned}
$$

is also bounded almost everywhere. Therefore,

$$
E \sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} |m(\beta, F', V') - m(\beta, F, V')|^2 \leq K_2\delta^2
$$

for some $K_2 > 0$.

The third term

$$|m(\beta, F, V') - m(\beta, F, V)|^2 = |\frac{\partial m(\beta, F, \tilde{V})}{\partial V}(V' - V)|^2$$

by mean value theorem.

$$\frac{\partial m}{\partial V} = -\frac{\partial F}{\partial V}\frac{\partial F}{\partial \beta} + (Y - F)\frac{\partial^2 F}{\partial \beta \partial V}$$

is bounded almost everywhere. Therefore, $\sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} |m(\beta, F, V') - m(\beta, F, V)|^2 \leq K_3\delta^2$ for some $K_3 > 0$.

Thus,

$$E[\sup_{\|\beta'-\beta\|\leq\delta, \|h'-h\|_{\mathcal{H}}\leq\delta} |(Y_i - F'(X_i'\beta', V_i')) \cdot \frac{\partial F'(X_i'\beta', V_i')}{\partial \beta} - (Y_i - F(X_i'\beta, V_i)) \cdot \frac{\partial F(X_i'\beta, V_i)}{\partial \beta}|^2]^{1/2}$$
$$\leq (K_1 + K_2 + K_3)\delta$$

Second, we need to show that $\mathcal{B}$ is compact, and $\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{H}, \|\cdot\|_{\mathcal{H}})} d\epsilon < \infty$. This follows from Corollary 2.7.4 in Van der Vaart and Wellner (1996), together with Assumption NOR(g).

□

Lemma 1.16. *For some finite matrix V, $\sqrt{n}\{M_n(\beta_0, h_0) + \Gamma_2(\beta_0, h_0)[\hat{h} - h]\} \implies N(0, V)$.*

Proof. We have that

$$\begin{aligned}\Gamma_2(\beta_0, h_0)[\hat{h} - h] = &E[-\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}[\hat{V}_i - V_{i0}] - \frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}[\hat{F} - F_0]] \\ =&E[-\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}(\hat{V}_i - V_{i0}) \\ &- \frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}(\hat{F}(X_i'\beta, V_i) - F_0(X_i'\beta, V_i))]\end{aligned}$$

Consider $E(\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta} \mid X_i'\beta_0, V_{i0})$, by Klein and Spady (1993), this term equals zero. Therefore, by law of iterated expectation

$$\Gamma_2(\beta_0, h_0)[\hat{h} - h] = E[-\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}(\hat{V}_i - V_{i0})]$$

By Assumption NOR(d),

$$\begin{aligned}\Gamma_2(\beta_0, h_0)[\hat{h} - h] &= -E[\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}\frac{1}{n}\sum_{j=1}^{n} g_n(Z_i, Z_j)\Psi_j + o_p(n^{-1/2})] \\ &= -\frac{1}{n}\sum_{j=1}^{n} E[\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta} g_n(Z_i, Z_j)\Psi_j] + o_p(n^{-1/2}) \\ &= -\frac{1}{n}\sum_{j=1}^{n} E[\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta} g_n(Z_i, Z_j)\Psi_j \mid Z_j] + o_p(n^{-1/2}) \\ &= -\frac{1}{n}\sum_{j=1}^{n} E[\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial V_i}\frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta} g_n(Z_i, Z_j) \mid Z_j]\Psi_j + o_p(n^{-1/2})\end{aligned}$$

Note that the expectations above are all with respect observation i. Next, we have

$$\begin{aligned}&\sqrt{n}\{M_n(\beta_0, h_0) + \Gamma_2(\beta_0, h_0)[\hat{h} - h]\} \\ &= \sqrt{n}\{\frac{1}{n}\sum_{i=1}^{n}[(Y_i - F_0(X_i'\beta_0, V_{i0})) \cdot \frac{\partial F_0(X_i'\beta_0, V_{i0})}{\partial \beta}] \\ -&\frac{1}{n}\sum_{i=1}^{n} E[\frac{\partial F_0(X'\beta_0, V_0)}{\partial V}\frac{\partial F_0(X'\beta_0, V_0)}{\partial \beta} g_n(Z, Z_i) \mid Z_i]\Psi_i\}\end{aligned}$$

Because Ψ_i and $Y_i - F_0(X_i'\beta_0, V_{i0})$ are orthogonal, by central limit theorem,

$$\sqrt{n}\{M_n(\beta_0, h_0) + \Gamma_2(\beta_0, h_0)[\hat{h} - h]\} \implies N(0, V)$$

where $V = V_1 + V_2$.

□

Proof of Theorem 4:

Proof. Lemma 1-7 show that the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003) are satisfied. Together with the consistency result, Theorem 4, the asymptotic distribution of the semiparametric least square estimator is implied by Theorem 2 in Chen, Linton, and Van Keilegom (2003).

□

Proof of Theorem 5:

Proof. This proof will generally verify the conditions of Theorem B in Chen, Linton, and Van Keilegom (2003).

The first condition is that with P^*-probability tending to one, $\hat{h}^* \in \mathcal{H}$, and $\parallel \hat{h}^* - \hat{h} \parallel_{\mathcal{H}} = o_{p^*}(n^{-1/4})$. This condition can be verified in a similar way as Lemma 6 above.

The second condition is

$$\sup_{(\beta,h)\in\mathcal{B}_{\delta_n}\times\mathcal{H}_{\delta_n}} \parallel M_n^*(\beta,h) - M_n(\beta,h) - \{M_n^*(\beta_0,h_0) - M_n(\beta_0,h_0)\} \parallel = o_{p^*}(n^{-1/2})$$

for all positive values $\delta_n = o(1)$. This can be verified in the same way as Lemma 7 above.

The third condition is

$$\sqrt{n}\{M_n^*(\hat{\beta},\hat{h}) - M_n(\hat{\beta},\hat{h}) + \Gamma_2(\hat{\beta},\hat{h})[\hat{h}^* - \hat{h}]\} = N(0,V) + o_{p^*}(1)$$

Note that

$$\Gamma_2(\hat{\beta},\hat{h})[\hat{h}^* - \hat{h}] = -E\{\frac{\partial \hat{F}(X_i'\hat{\beta},\hat{V})}{\partial V}\frac{\partial \hat{F}(X_i'\hat{\beta},\hat{V})}{\partial \beta}(\hat{V}^* - \hat{V})$$
$$+ \frac{\partial \hat{F}(X_i'\hat{\beta},\hat{V})}{\partial \beta}(\hat{F}^* - \hat{F})\}$$

Following Chen, Linton, and Van Keilegom (2003), under standard regularity conditions the bias of $\hat{F}^*$, $E^*\hat{F}^* - \hat{F}$ can be majorized by some bounded continuous function times $o(n^{-1/2})$. Also by

Assumption NOR(d), $\hat{V}^* - \hat{V} = \frac{1}{n}\sum_i g_n(Z^*, Z_i^*)\Psi_i^* + o_p(n^{-1/2})$,

$$\begin{aligned}
\Gamma_2(\hat{\beta}, \hat{h})[\hat{h}^* - \hat{h}] = -E\{ & \frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial V}\frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial \beta}\frac{1}{n}\sum_i g_n(Z^*, Z_i^*)\Psi_i^* \\
& + \frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial \beta}(\hat{F}^* - E^*\hat{F}^*)\} + o_p(n^{-1/2}) \\
& = -\, E\frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial V}\frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial \beta}\frac{1}{n}\sum_i g_n(Z^*, Z_i^*)\Psi_i^* \\
& - \frac{1}{nh_1h_2}\sum_{i=1}^n E\frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial \beta}[\frac{Y_i^* K_1(\frac{X'\beta - X_i^{*'}\beta}{h_1})K_2(\frac{\hat{V} - \hat{V}_i^*}{h_2})}{f} \\
& - E^*\frac{Y_i^* K_1(\frac{X'\beta - X_i^{*'}\beta}{h_1})K_2(\frac{\hat{V} - \hat{V}_i^*}{h_2})}{f}] \\
& = -\, \frac{1}{n}\sum_{i=1}^n E\frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial V}\frac{\partial \hat{F}(X'\hat{\beta}, \hat{V})}{\partial \beta} g_n(Z^*, Z_i^*)\Psi_i^* \\
& - \frac{1}{n}\sum_{i=1}^n \frac{\partial \hat{F}(X_i^{*'}\hat{\beta}, \hat{V}_i)}{\partial \beta}Y_i^* - E^*[\frac{\partial \hat{F}(X_i^{*'}\hat{\beta}, \hat{V}_i)}{\partial \beta}Y_i^*] + o_{p^*}(n^{-1/2})
\end{aligned}$$

where the last equal sign follows from change of variables and Taylor expansion. In the last equation, the first three terms are independent and zero mean random variables. Therefore, they satisfies central limit theorem and hence the third condition is satisfied.

□

Proof of Theorem 6:

Proof. First we need to find the asymptotic joint distribution of $\hat{\beta}^1$ and $\hat{\beta}^2$. Since $\sqrt{n}(\hat{\beta}^i - \beta_0) \rightsquigarrow N(0, \Omega_i)$,

$$\sqrt{n}[\begin{pmatrix} \hat{\beta}^1 \\ \hat{\beta}^2 \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \beta_0 \end{pmatrix}] \rightsquigarrow N(0, \Omega_{12}),$$

where Ω_{12} is the asymptotic covariance matrix of $\hat{\beta}^1$ and $\hat{\beta}^2$, and has the form $\Omega_{12} = \begin{pmatrix} \Omega_1 & A \\ A' & \Omega_2 \end{pmatrix}$, A is the covariance between $\hat{\beta}^1$ and $\hat{\beta}^2$. The influence functions of these two estimators are

$$infl_i^1 = \Sigma^{-1}[(Y_i - F_0)\cdot\frac{\partial F_0}{\partial \beta} - E[\frac{\partial F_0}{\partial V}\frac{\partial F_0}{\partial \beta}g_n(Z, Z_i) \mid Z_i]]\Psi_i$$

$$infl_i^2 = \Sigma^{-1}[\frac{Y_i - F_0}{F_0 \cdot (1 - F_0)} \cdot \frac{\partial F_0}{\partial \beta} - E[\frac{\partial F_0}{\partial V}\frac{\partial F_0}{\partial \beta} \cdot \frac{g_n(Z, Z_i)}{F_0 \cdot (1 - F_0)} \mid Z_i]]\Psi_i.$$

Therefore $A = E(infl_i^1 \times infl_i^{2'})$.

$$\hat{\beta}^{MA} = \lambda\hat{\beta}^1 + (1 - \lambda)\hat{\beta}^2$$

By continuous mapping theorem,

$$\sqrt{n}(\hat{\beta}^{MA} - \beta_0) \rightsquigarrow N(0, \Omega)$$

where $\Omega = (\lambda, 1 - \lambda)\Omega_{12} \begin{pmatrix} \lambda \\ 1 - \lambda \end{pmatrix} = \lambda^2\Omega_1 + (1 - \lambda)^2\Omega_2 + \lambda(1 - \lambda)(A + A')$

□

Proof of Corollary 1:

Proof. $AVar(\hat{\beta}_k^{MA})$ is no greater than both $AVar(\hat{\beta}_k^1)$ and $AVar(\hat{\beta}_k^2)$ when using optimal weight λ_k^*. The reason is simple that the minimization is conducted on the interval $[0, 1]$. If $\lambda_k^* \in (0, 1)$, it means the value of $AVar(\hat{\beta}_k^{MA})$ at λ_k^* is smaller than the values at $\lambda_k = 1$ and $\lambda_k = 0$, which imply $AVar(\hat{\beta}_k^1)$ and $AVar(\hat{\beta}_k^2)$, respectively. Otherwise, there would be a contradiction with the operation of minimization. □

Chapter 2

A Semiparametric Estimator for Binary Response Models with Endogenous Regressors

2.1 Introduction

This chapter is concerned with estimating a semiparametric binary response model with endogenous explanatory variables. In particular, a triangular simultaneous equations model with binary outcome is discussed here. The binary response model we consider is represented of the form

$$Y = I[X'\beta - U \geq 0]$$

where X is an observed vector of explanatory variables and Y is an indicator of the event that the value of $X'\beta - U$ is non-negative. In addition, U is the unobserved error term and β is the coefficient vector of interest. If X and U are independent and U follows some known distribution, this binary response model can be estimated via standard parametric estimation procedures, such as Probit or Logit, see McFadden (1984) for a detailed survey.

However, when the distribution of U is unknown, misspecification would in general cause the estimators inconsistent even if the independence between X and U still holds. Therefore, semiparametric estimators of β have been proposed in the literature, where the conditional distribution of U given X is not specified. For example, Manski (1975) and Manski (1985) introduce the maximum score estimator for β. It requires that along with the full support of X, the conditional distribution of U given X is median independent, i.e. $Med(U|X) = 0$. This identification condition is rather weak compared to previous parametric estimators. The convergence rate is $n^{-1/3}$ by Kim and Pollard (1990). Horowitz (1992) maintains the median independence condition for identification and modifies Manski's maximum score estimator by smoothing the objective function with kernel functions. As a result, under smoothness conditions, Horowitz's smoothed maximum score estimator can attain a faster convergence rate, at least $n^{-2/5}$ with asymptoticly normal distribution. Unlike linear regression models, conditional mean independence is generally not enough to achieve identification of the binary response model, as noted by Manski (1988) and Horowitz (2009). Ichimura (1993) and Klein and Spady (1993) employ the single-index restriction and propose semiparametric estimators for the binary response model.

When some components of X are endogenous, we need more information, for example some instrumental variables, to account for the endogeneity. Unlike the separable models, the parameters in the binary response model, one type of non-separable models, are not universally identified under the standard independence assumption between the instruments and the error term U, see Blundell and Powell (2003), Chesher (2010), Chesher and Rosen (2013) and references therein. Alternatively, the control function approach has been widely used in the estimation of the simultaneous equation models with discrete dependent variable. Smith and Blundell (1986) and Rivers and Vuong (1988) introduce a two-stage Probit estimator (2SProbit) for the binary response triangular system with continuous endogenous regressors by specifying the joint distribution of error terms as normal distribution. Their estimation procedure take the residuals from the reduced equation for the endogenous regressors as the covariates in the binary response structural equation to account for endogeneity. However, the shortcoming of this type of parametric estimators is the same as mentioned above. Newey, Powell, and Vella (1999) propose a two-step nonparametric estimator for a separable triangular system using the control function approach. Blundell and Powell (2004) firstly use the control function to account for endogeneity in the semiparametric binary response model. The estimator they consider is the extension of the matching estimator proposed by Ahn, Ichimura, and Powell (1996). Rothe (2009) extends the Klein and Spady (1993) estimator to the endogenous case by forming a triangular system and estimates it using two-step semiparametric maximum likelihood. The endogeneity is also accounted for by a control variable obtained from the first-step reduced form regression. These studies in the semiparametric settings rely on conditional independence for identification, i.e. the endogenous explanatory variable X and the error term U are independent conditional on a control variable. Lee (2007) applies a control function approach to a linear triangular simultaneous equations models with conditional quantile independence restrictions. He proposes a two-step semiparametric estimation procedure in which the control function is estimated by series estimator. Liao (2012) proposes a two-step estimator for the triangular system with binary response. The key identification condition of this estimator relies on is that conditional on a control variable, the quantile of the error term U is independent of X as well as the instruments.

This chapter proposes a new semiparametric estimator for the binary response model with endogeneity. The identification conditions we employ here are mainly following the model setting in Liao (2012), because the quantile independence condition is not as restrictive as full independence used by previous studies.. However, one primary difference lies in that in our procedure not only the structural parameters β but also the choice probability can be consistently estimated. The estimator proposed by Liao (2012) follows Horowitz's smoothed maximum score estimation procedure, which can only provide an estimator for β. However, in some applications, such like policy evaluation, it is the choice probability and even the marginal effects of the covariate change that are more practical and useful. The basic idea here is to obtain more information on the characteristics of the error term distribution via nonparametric estimating additional functional structures attached to the model. In the appendix of Manski (1988), the existence of the dual models with binary outcomes is proved. It implies the possibility of transforming the linear-index binary response model under quantile independence assumption into a class of models whose error terms follow some known distributions. Khan (2013) exhibits that under mild assumptions, the binary response model without endogeneity under Manski's median independence condition is equivalent to a multiplicative heteroskedastic binary response model. Under this equivalence, it suffices to estimate the binary response model by maximizing standard Probit/Logit criterion functions while the heteroskedasticity is pinned down using sieve estimation. This chapter follows this approach to restore the choice probability estimates when some components of the explanatory variables are endogenous.

The remainder of this chapter is organized as follows. Section 2.2 describes the model specification. Section 2.3 establishes an equivalence result between the binary response model with endogeneity and a heteroskadastic one we use for estimation, and identification can be achieved . Section 2.4 describes our semiparametric estimation procedure. In Section 2.5, asymptotic properties of our estimator are analyzed. Section 2.6 studies finite sample properties of our estimator via Monte Carlo simulations. Theorem proofs are provided in the Appendix.

2.2 Model

The binary response model we consider has the form

$$Y_i^* = X_i'\beta_0 - U_i \tag{2.1}$$

$$Y_i = I[Y_i^* \geq 0] \tag{2.2}$$

where $I[\cdot]$ is the indicator function, y_i is the observed binary response variable, and $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ is a vector of observed covariates. Y^* is a latent variable which can be affected by the change of X and determines Y through its sign. U is the scalar unobserved error term whose conditional distribution given other observables can impact the identification of the parameters of interest. β_0 is the d_x-dimensional unknown parameter of interest. If U is assumed to be independent of X and follow some parametric distribution, $E[Y \mid X = x] = F_U(x'\beta_0)$, where $F_U(\cdot)$ is the cumulative distribution function of U. Therefore, β_0 can be estimated by maximum likelihood estimation. However, when endogeneity and heteroscedasticity arise, the independence between X and U no longer holds and other conditions for identification are needed. Moreover, the assumption that U follows some specific parametric distribution is restrictive in economic applications, and weaker conditions on the distribution of U need to be imposed.

Denote $X = (X_1, Z_1')'$, where X_1 is a scalar endogenous variable and X_1 is a $d_{z_1}-$dimensional sub-vector of the instruments $Z = (Z_1', Z_2')' \in \mathcal{Z} \subseteq \mathbb{R}^{d_{z_1}+d_{z_2}}$. In a triangular simultaneous equations model, X_1 is assumed to be determined by a reduced form equation

$$X_1 = h_0(Z, \eta) \tag{2.3}$$

where h_0 is a real-valued measurable function mapping the exogenous instruments Z and a disturbance $\eta \in \mathbb{R}$ to the endogenous explanatory variable X_1. Here η is assumed independent of Z and h_0 is strictly monotone in η. Following Blundell and Powell (2004) and Liao (2012), the condition imposed for the identification of this model is through a control variable $V \in \mathbb{R}$ satisfying

$$Q_{U|X,Z}(\tau \mid x, z) = Q_{U|V,Z}(\tau \mid v, z) = Q_{U|V}(\tau \mid v) \equiv \lambda_\tau(v) \tag{2.4}$$

where $Q_U(\tau)$ is the τth quantile of the random variable U, i.e. $Q_U(\tau) = \inf\{u : F_U(u) \geq \tau\}$, and correspondingly $Q_{U|X,V}(\tau \mid x, v)$ is the τth quantile of U conditional on X and V. The condition in (2.4), known as the quantile exclusion restriction, is crucial to the identification of our semiparametric binary response model with endogeneity defined in (2.1), (2.2) and (2.3). It essentially requires that the control variable V can adjust for the endogeneity in the structural equation (2.1). This quantile exclusion restriction is more flexible than full independence between U and X conditional on control variables. For example, it admits heteroskedasticity in (2.1). This model is a semiparametric binary response version of Lee (2007).

How to construct a control variable used in this model depends on the structure of the reduced form equation for X_1 in (2.3). If the disturbance term η is additive in function h_0. A natural way to get the control variable is to obtain the residuals from regressing the endogenous regressor X_1 on the instrumental variables Z, e.g. Smith and Blundell (1986); Rivers and Vuong (1988); Blundell and Powell (2004); Rothe (2009) *et al.* The function form of $E(X_1 \mid Z = z)$ could either be parametric specified like the single index $z'\alpha_o$ or be obtained by nonparametric regressions. When η_i is not additive in h_0, Imbens and Newey (2009) demonstrated that the control variable V can be constructed as the conditional cumulative distribution of the endogenous variable given the instruments, i.e. $V_i \equiv F_{X_1|Z}(x_1 \mid z) = F_\eta(\eta)$ if h_0 is assumed strictly monotone in η and $U, \eta \perp Z$.

Define $\epsilon \equiv U - \lambda_\tau(V)$, and since ϵ is strictly increasing in U,

$$Q_{\epsilon|X,Z}(\tau \mid X = x, Z = z) = Q_{\epsilon|X,V}(\tau \mid X = x, V = v) = Q_\epsilon(\tau \mid V = v) = Q_U(\tau \mid V = v) - \lambda_\tau(v) = 0$$

Then the binary response model can be rewritten as

$$Y = I[X'\beta_0 - \lambda_\tau(V) - \epsilon \geq 0] \tag{2.5}$$

2.3 Identification

This section aims to establish an observational equivalence between the binary response triangular system with quantile exclusion restriction (2.4) and a heteroskedastic binary response model. The main idea is based on Khan (2013).

Consider the binary response model described in (2.5) with the following assumptions:

Assumption 2.1. *At lease one component of $X \in \mathbb{R}^{d_x}$ with nonzero coefficient, given other components and the control variable V, has positive density everywhere on $\mathbb{R}$ with respect to Lebesgue measure.*

Assumption 2.2. *The function $\lambda_\tau(\cdot)$ is continuously differentiable almost everywhere.*

Assumption 2.3. *$F_{\epsilon|X,Z}(\cdot \mid X = x, Z = z)$ is the conditional CDF of ϵ given X and Z. $F_{\epsilon|X,Z}$ is continuous on $\mathbb{R} \times \mathcal{X} \times \mathcal{Z}$. $f_{\epsilon|X,Z}(\epsilon \mid X = x, Z = z) = \frac{\partial F_{\epsilon|X,Z}(\epsilon|X=x,Z=z)}{\partial \epsilon}$ exists and is continuous and positive on $\mathbb{R}$ for all $(x, z) \in \mathcal{X} \times \mathcal{Z}$. $F_{\epsilon|X,Z}(0|X = x, Z = z) = \tau$ for all $(x, z) \in \mathcal{X} \times \mathcal{Z}$.*

Theorem 2.4. *Under Assumption 2.1 - Assumption 2.3, the binary response model defined by (2.5) is observationally equivalent with the model defined by*

$$Y = I[X'\beta_0 - \lambda_\tau(V) - \sigma_0(X, Z) \cdot \xi \geq 0] \tag{2.6}$$

under the following conditions:

Condition 2.5. Assumption 2.1 is satisfied by (2.6).

Condition 2.6. σ_0 is a continuous measurable function and positive a.e. on $\mathcal{X} \times \mathcal{Z}$. ξ is independent of (X, Z) and follows some known distribution . $Q_\xi(\tau) = 0$. $f_\xi(\cdot)$ exists and is positive and continuous on $\mathbb{R}$.

Moreover, the identification of the parameters in (2.6) can be achieved. Based on the identification results by Liao (2012), the parameters β_0 and $\lambda_\tau(\cdot)$ can be identified up to scale under Assumption 2.1 - Assumption 2.3. Thus due to Theorem 2.4, β_0 and λ_τ in (2.6) stay the same as in (2.5) and then can be identified up to scale. In addition, since $E(Y|X = x, Z = z) = \Pr(Y = 1|X = x, Z = z) = G_\xi(\frac{x'\beta_0 - \lambda_\tau(v)}{\sigma_0(x,z)})$ and $G_\xi(\cdot)$ is strictly increasing on the real line, for some $C > 0$, $C{\cdot}\sigma_0(\cdot)$ can be pinned down based on identification of β_0 and λ_τ as well as $E(Y|X = x, Z = z)$. Therefore, β_0, γ_τ and σ_0 in (2.6) can be identified up to scale.

2.4 Estimation

As the extension of the exogenous case in Khan (2013), the previous result implies that a triangular simultaneous equation model with binary outcome can be transformed into a heteroskedastic binary response model while keeping the finite dimensional parameters and conditional probability unchanged. This provides us a potential approach to estimate the endogenous binary response model. Based on (2.6) under Condition 2.5 - Condition 2.6, if the control variable v_i and the function $\lambda_\tau(\cdot)$ are known, the true value of the finite dimensional parameter vector β_0 and the infinite dimensional parameter $\sigma_0(\cdot, \cdot)$ should minimize the following population objective function:

$$S(\beta, \sigma) = E(Y - G_\xi(\frac{X'\beta - \lambda_\tau(V)}{\sigma(X, Z)}))^2 \tag{2.7}$$

Given an i.i.d sample $\{Y_i, X_i, Z_i\}_{i=1}^n$, (β_0, σ_0) can be jointly estimated via nonlinear least squared (NLS) estimation in which the infinite dimensional parameter space is approximated by the sieve space. In addition, both the control function v_i and the function λ_τ are unknown. Note that the control variable V can be estimated at the first stage using parametric or nonparametric regressions. Here, we assume that this part has already been done, and an estimator for V is available denoted as $\hat{V}$. As for the unknown function $\lambda_\tau(\cdot)$, it is natural to incorporate its estimation jointly with estimating β_0 and σ_0. However, all the parameters of interest can only be identified up to scale. Here, the coefficient of X_1 is set as 1 and $\beta_0 = (1, \gamma_0')'$, where $\gamma_0 \in \mathbb{R}^{d_x - 1}$. To ensure σ_0 positive, let $l_0(x, z) = -\ln \sigma_0(x, z)$ and in this case l_0 could be either positive or negative.

Let $\theta_0 \equiv (\gamma_0, \lambda_\tau, l_0) \in \Theta$, where Θ is an infinite dimensional parameter space, and $\theta_n \equiv (\gamma, \lambda_n, l_n) \in \Theta_n$, where Θ_n is a sieve space. We propose the following sieve estimator:

$$\hat{\theta}_n = \arg \min_{\theta_n \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - G_\xi[(X_{1i} + Z_{1i}'\gamma - \lambda_n(\hat{V}_i)) \exp(l_n(X_{1i}, Z_i))]\}^2 \tag{2.8}$$

Let $b^{k_{1n}}(v) = (b_{01}(v), \ldots, b_{0k_{1n}}(v))'$ be a sequence of known basis functions for approximating $\lambda_\tau(v)$, i.e. $\lambda_n(v) = b^{k_{1n}}(v)'\Pi_n$, where Π_n is a k_{1n} dimensional vector of constants. Similarly, let $c^{k_{2n}}(x, z) = (c_{01}(x, z), \ldots, c_{0k_{2n}}(x, z))'$ be the basis functions for approximating $l_0(x, z)$. $l_n(x, z) = c^{k_{2n}}(x, z)'\Psi_n$, where Ψ_n is a k_{2n} dimensional vector of constants. The estimators for Π_n and Ψ_n are denoted as $\hat{\Pi}_n$ and $\hat{\Psi}_n$, and therefore $\hat{\lambda}_n(v) = b^{k_{1n}}(v)'\hat{\Pi}_n$ and $\hat{l}_n(x, z) = c^{k_{2n}}(x, z)'\hat{\Psi}_n$.

2.5 Asymptotic Properties

In this section, we provide large sample properties of our sieve NLS estimator $\hat{\theta}_n$ given in (2.8). First of all, we review some notations which have been widely used in previous literature on nonparametric estimation. For any $1 \times k$ vector $a = (a_1, a_2, \cdots, a_k)$ of non-negative integers, let

$$\nabla^a f(x) = \frac{\partial^{|a|}}{\partial x_1^{a_1} \cdots \partial x_k^{a_k}} f(x)$$

denote the $|a|$-th derivative of a function $f : \mathbb{R}^k \to \mathbb{R}$, where $|a| = \sum_{i=1}^k a_i$. For any $\gamma > 0$, let $[\gamma]$ denote the largest integer smaller than γ and $\| \cdot \|$ denote the Euclidean norm. Then define the γ-th Holder norm:

$$\| f \|_{\Lambda^\gamma} = \sum_{|a|\leq[\gamma]} \sup_{x\in\mathbb{X}} |\nabla^a f(x)| + \sum_{|a|=[\gamma]} \sup_{x\neq\bar{x}} \frac{|\nabla^a f(x) - \nabla^a f(\bar{x})|}{\| x - \bar{x} \|^{\gamma-[\gamma]}}$$

A Holder space with smoothness γ, denoted as $\Lambda^\gamma(\mathcal{X})$, consists of all functions $f \in C^{\gamma-[\gamma]}(\mathcal{X})$ such that $\| f \|_{\Lambda^\gamma}$ is finite. It has been shown that equipped with the γ-th Holder norm, a Holder space is complete. Since the support of x_i and z_i might not be compact in this study, we confine the parameter space to a weighted Holder ball with radius c, $\Lambda^\gamma_c(\mathcal{X}, \omega_1) \equiv \{f \in \Lambda^\gamma(\mathcal{X}) : \| f(\cdot)(1+ \| \cdot \|^2)^{-\omega_1/2} \|_{\Lambda^\gamma} \leq c < \infty\}$, where $\omega_1 \geq 0$, as indicated by Ai and Chen (2003) and Chen, Hong, and Tamer (2005). With the weighting function $(1+ \| \cdot \|^2)^{-\omega_1/2}$, the functions in the weighted Holder ball are allowed to have unbounded derivatives. In addition, define the following two norms:

$$\| f(x) \|_2 = (\int_{\mathcal{X}} f(x)^2 dF_X)^{1/2}$$

$$\| f(x) \|_{\infty,\omega} = \sup_{x\in\mathcal{X}} |f(x)(1+ \| x \|^2)^{-\omega/2}|$$

We provide the following assumptions to show the consistency of our sieve NLS estimator $\hat{\theta}_n$:

Assumption 2.7. *Recall the parameters of interest $\theta_0 \equiv (\gamma_0, \lambda_\tau, l_0) \in \Theta$. Assume that the parameter space Θ consists of all $(\gamma, \lambda(\cdot), l(\cdot))$ such that*

(1) $\gamma \in \Gamma \subset \mathbb{R}^{d_{z_1}}$ and Γ is compact.

(2) $G_\xi((x_1 + z_1'\gamma - \lambda(v)) \exp(l(x_1, z))) \in \Lambda^s_c(\mathcal{X} \times \mathcal{Z}, \omega_1)$ for some $s > 0$ and $\omega_1 \geq 0$.

(3) $\lambda(\cdot)$ is continuously differentiable and its first order derivative $\lambda^{(1)}(\cdot)$ satisfies

$$\sup_{v\in\mathcal{V}} \lambda^{(1)}(v) \leq C < \infty.$$

This assumption imposes regularity conditions on the functional space. Assumption 2.7(1) and (2) ensure the compactness of the parameter space and are relatively standard in the semiparametrics literature, (see Gallant and Nychka (1987) and references therein). Note that a functional of the two unknown functions $\lambda(\cdot)$ and $l(\cdot)$, rather than themselves, is assumed to be compact here, which is denoted as $\phi(w,\theta) \equiv G_\xi((x_1+z_1'\gamma-\lambda(v))\exp(l(x_1,z)))$, where $w \equiv (x_1, z')'$. This requirement, to some extent, is weaker compared to that both λ and l are smooth to certain specific degrees and is enough to ensure the consistency of our sieve estimators with respect to the weighted sup norm. Assumption 2.7(3) states that the functional ϕ admits at least a first order Taylor expansion with respect to the estimator of control variable v_i.

Assumption 2.8. *Let $\mathcal{X}$ denote the support of $X \in \mathbb{R}^{d_x}$ and $\mathcal{Z}$ the support of $Z \in \mathbb{R}^{d_{z_1}+d_{z_2}}$, where $d_{z_1} = d_x - 1 \geq 0$ and $d_{z_2} \geq 1$. Assume the covariates $(Y_i, X_i', Z_i')_{i=1}^n$ satisfy*

(1) The data $(Y,X,Z)'$ are i.i.d, with $V = F_{X_1|Z}(x_1|z)$, satisfying $\Pr(Y=1 \mid X=x, Z=z) = G_\xi((x_1 + z_1'\gamma_0 - \lambda_\tau(v))\exp(l_0(x_1,z)))$.

(2) X_1 conditional on Z has density function with respect to Lebesgue measure which is positive almost everywhere on $\mathbb{R}^1$.

(3) $\mathcal{Z}$ is compact.

(4) $W \equiv (X_1, Z')'$. Let f_W denote the density function of W.

$$\int (1+ \parallel w \parallel^2)^\omega f_W(w)dw < \infty$$

where $\omega > \omega_1 \geq 0$.

(5) $E(b^{\kappa_{1n}}(v)b^{\kappa_{1n}}(v)')$ and $E(c^{\kappa_{2n}}(x_1,z)c^{\kappa_{2n}}(x_1,z)')$ are non-singular for all n.

This assumption collects some conventional support restrictions on the covariates. Large support requirement on the endogenous variable X_1 conditional on all the exogenous variables Z ensures identification of the binary response model discussed above. The compactness of $\mathcal{Z}$ is only assumed for simplicity and can be relaxed.

Assumption 2.9. *Recall that $\hat{V}$ is an estimator of the control variable V. Suppose*

$$\sup_{X_1 \in \mathcal{X}, Z \in \mathcal{Z}} |\hat{V} - V| = O_p(\Delta_v)$$

where $\Delta_v = o_p(1)$.

This is a high level condition on the uniform convergence rate of the control variable estimator $\hat{v}_i$.

Before stating the consistency results of our sieve NLS estimator, we introduce the following notations for simplicity. Denote $\phi(w,\theta) \equiv G_\xi((x_1 + z_1'\gamma - \lambda(v))\exp(l(x_1,z)))$. Define the metric on the parameter space as $\| \theta_1 - \theta_2 \|_2 = \| \phi(\cdot,\theta_1) - \phi(\cdot,\theta_2) \|_2$ and $\| \theta_1 - \theta_2 \|_{\infty,\omega} = \| \phi(\cdot,\theta_1) - \phi(\cdot,\theta_2) \|_{\infty,\omega}$.

Theorem 2.10. *Under Assumption 2.7 - Assumption 2.9, if $\kappa_{1n} \wedge \kappa_{2n} \to \infty$ and $(\kappa_{1n} \vee \kappa_{2n})/n \to 0$, we have*

$$\| \hat{\gamma}_n - \gamma_0 \| = o_p(1)$$

and

$$\| \hat{\theta}_n - \theta_0 \|_\infty = o_p(1)$$

Before we demonstrate the results on the convergence rate of our semi-parametric estimator, several previous assumptions need to be strengthened. These enhancements mainly aim at improving approximation accuracy of the sieve space and the first stage control variable estimation.

Assumption 2.11. *$\lambda(v) \in \Lambda^s_{c_1}(\mathcal{V},\omega_1)$, for some $s > 0$ and $\omega_1 \geq 0$; $l(w) \in \Lambda^s_{c_2}(\mathcal{X} \times \mathcal{Z},\omega_1)$, for some $s > 0$ and $\omega_1 \geq 0$.*

This condition imposes stronger smoothness requirement on two unknown functions λ and l than Assumption 2.7, which only assumes that a functional of λ and l belongs to a weighted Holder ball. However, Assumption 2.11 does not conflict with Assumption 2.7. The same exponent of Holder space s and weighting function ω_1 make it sufficient to conclude $\phi : (\cdot, \lambda, l) \mapsto [0,1] \in \Lambda_c^s(\mathcal{X} \times \mathcal{Z}, \omega_1)$, for some $c_1, c_2 < \infty$. Assumption 2.11 is imposed to detail the accuracy of the sieve space approximating the parameter space.

Assumption 2.12. *The smoothness exponent of Holder space s satisfies $2s > d_z + 1$. For Assumption 2.8, $w > w_1 + s$.*

This condition is a stronger version of Assumption 2.7 and Assumption 2.8. The strengthened smoothness of the parameter space and the further controlled tail behavior of the covariates can improve the approximation accuracy of our sieve space as required by the convergence rate results.

To establish the convergence rate results, we define the following sieve NLS estimator when the control variable v_i is fully observed. Let

$$\tilde{\theta}_n = \arg \min_{\theta_n \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - G_\xi[(X_{1i} + Z'_{1i}\gamma - \lambda_n(V_i)) \exp(l_n(X_{1i}, Z_i))]\}^2$$

Theorem 2.13. *If Assumption 2.7 - Assumption 2.8 and Assumption 2.11 - Assumption 2.12 are satisfied, then*

$$\| \tilde{\theta}_n - \theta_0 \|_2 = O_p(\sqrt{\frac{\kappa_{1n} \vee \kappa_{2n}}{n}} + \kappa_{1n}^{-s} + \kappa_{2n}^{-s/(d_z+1)})$$

Recall that the L^2 metric on the parameter space is actually a measure of the distance between two probability functions over the sample space. Hence this convergence rate result is for the choice probability estimator.

2.6 Monte Carlo

This section examines the finite sample performance of our sieve NLS estimator for the binary response model based on a triangular system. In this experiment, n i.i.d observations,

$\{Y_i, X_i, Z_i\}_{i=1}^n$, are generated through the following data generating process

$$Y_i = I\{X_{1i} + Z_{1i}\gamma_0 - \lambda_\tau(V_i) - \epsilon_i \geq 0\}$$

$$X_{1i} = \alpha_0 + Z_{1i}\alpha_1 + Z_{2i}\alpha_2 + \eta_i$$

The true parameter values are set as $\gamma_0 = 1$, $\alpha_0 = -2$, $\alpha_1 = \frac{1}{2}$ and α_2 =1. The exogenous variables $\{Z_{1i}, Z_{2i}\}$ are assumed to follow normal distributions, where $Z_{1i} \sim N(2,1)$ and $Z_{2i} \sim N(0,1)$. In addition, $\eta_i \sim N(0,1)$. In this study, we focus on the median independence case, i.e. $\tau = 0.5$. For the control function λ_τ and the error distribution in the structural equation ϵ_i, we use the following designs:

1. $\lambda_{0.5}(v_i) = \eta_i$ and $\epsilon_i \sim N(0, 2^2)$
2. $\lambda_{0.5}(v_i) = 1 + 2\eta_i - 2\eta_i^2$ and $\epsilon_i \sim$Student's t distribution with df = 2
3. $\lambda_{0.5}(v_i) = \exp(-\eta_i)$ and $\epsilon_i = Z_{2i}^2 \cdot t_i$, where $t_i \sim N(0,1)$

For each design, we study the performance of the following estimators: two-stage IV Probit estimator by Smith and Blundell (1986) and Rivers and Vuong (1988) (2SProbit); our sieve NLS estimator without endogeneity, i.e. $\lambda_\tau(\cdot)$ is assumed to be known (SNLS-EX); the sieve NLS estimator with the control variable V as given (SNLS-OR); our sieve NLS estimator with the control variable V unknown, i.e. a first-step estimator $\hat{V}$ is used instead of the true value V (SNLS-2S). To simplify the implementation, we utilize the separable property of the reduced form equation and use the residual from linear regression of X_1 on Z as the consistent estimator of the control variable V. For all the designs, we consider the sample size $n = 250, 500, 1000$ and set the number of replications to 500. The polynomial series are used to estimate both $\lambda_\tau(v)$ and $l_0(x_1, z)$. When $n = 250$, both series are set of degree 1; when $n = 500$, λ_n is of degree 2 and l_n degree 1; when $n = 1000$, both are of degree 2. A standard normal distribution is used as the known zero-median distribution for ξ.

Table 2.1: Simulation Design 1

		MEAN	SD	RMSE	MAD	25%	50%	75%
n=250	2SProbit	1.036	0.379	0.380	0.245	0.754	0.986	1.245
	SNLS-EX	0.989	0.151	0.151	0.106	0.881	0.976	1.085
	SNLS-OR	1.076	0.617	0.621	0.293	0.713	0.957	1.297
	SNLS-2S	1.071	0.609	0.613	0.316	0.703	0.960	1.340
n=500	2SProbit	1.037	0.258	0.260	0.167	0.852	1.014	1.186
	SNLS-EX	0.996	0.114	0.114	0.070	0.922	0.984	1.064
	SNLS-OR	1.034	0.322	0.324	0.200	0.807	0.998	1.205
	SNLS-2S	1.034	0.338	0.339	0.207	0.796	1.002	1.208
n=1000	2SProbit	1.017	0.185	0.186	0.122	0.886	1.000	1.127
	SNLS-EX	1.002	0.074	0.074	0.050	0.954	0.998	1.054
	SNLS-OR	1.154	0.395	0.424	0.222	0.891	1.091	1.338
	SNLS-2S	1.193	0.467	0.505	0.223	0.924	1.129	1.379

The simulation results are given in Table 2.1 - Table 2.3. For each design, we focus on the estimator of γ_0 and report the mean value (MEAN), standard deviation (SD), root mean squared error (RMSE), median absolute deviation (MAD) and the sample quartiles of $\hat{\gamma}_n$ from all the replications.

We can drawn some preliminary conclusions from the simulation results. First, in terms of bias and RMSE, our sieve NLS estimator performs generally well across all three designs. There was no unexpected large deviation from the true value of γ_0 in this experiment. Second, the semiparametric setting and the weak assumption for identification in this chapter provide flexibility for our estimator under distinct data generating processes. Meanwhile, the accuracy of the two-stage Probit estimation procedure is sensitive to model misspecification. For example, in Design 3, the 2SProbit estimators exhibit significant upward biases which are around 60%. Meanwhile, our semiparametric estimator is more stable in this setting. Third, the performance of SNLS-EX, the distribution free estimation without endogeneity in Khan (2013), behaves well in terms of both bias and standard deviation compared to 2SProbit

Table 2.2: Simulation Design 2

		MEAN	SD	RMSE	MAD	25%	50%	75%
n=250	2SProbit	1.103	0.479	0.489	0.259	0.782	1.040	1.323
	SNLS-EX	1.008	0.112	0.112	0.072	0.936	0.999	1.079
	SNLS-OR	1.218	0.775	0.804	0.320	0.759	1.048	1.441
	SNLS-2S	1.250	0.945	0.976	0.372	0.714	1.059	1.474
n=500	2SProbit	1.049	0.313	0.317	0.194	0.829	1.003	1.224
	SNLS-EX	1.004	0.077	0.077	0.048	0.952	1.002	1.048
	SNLS-OR	1.014	0.240	0.240	0.165	0.838	0.984	1.171
	SNLS-2S	1.019	0.272	0.272	0.181	0.818	1.008	1.177
n=1000	2SProbit	1.015	0.196	0.197	0.133	0.878	0.997	1.140
	SNLS-EX	1.010	0.056	0.056	0.037	0.974	1.006	1.048
	SNLS-OR	0.945	0.318	0.322	0.218	0.721	0.901	1.129
	SNLS-2S	0.911	0.311	0.324	0.224	0.704	0.859	1.069

Table 2.3: Simulation Design 3

		MEAN	SD	RMSE	MAD	25%	50%	75%
n=250	2SProbit	1.703	0.588	0.917	0.601	1.278	1.601	2.004
	SNLS-EX	0.999	0.052	0.052	0.032	0.964	0.998	1.028
	SNLS-OR	1.215	0.739	0.769	0.263	0.820	1.081	1.379
	SNLS-2S	1.227	0.744	0.777	0.257	0.819	1.049	1.466
n=500	2SProbit	1.642	0.363	0.737	0.602	1.358	1.602	1.896
	SNLS-EX	1.001	0.039	0.039	0.021	0.978	1.000	1.020
	SNLS-OR	1.058	0.357	0.361	0.209	0.805	0.990	1.229
	SNLS-2S	1.070	0.406	0.412	0.200	0.811	0.987	1.221
n=1000	2SProbit	1.600	0.262	0.655	0.574	1.395	1.574	1.771
	SNLS-EX	1.002	0.027	0.027	0.018	0.984	1.002	1.020
	SNLS-OR	0.952	0.218	0.223	0.156	0.795	0.939	1.094
	SNLS-2S	0.954	0.242	0.246	0.165	0.786	0.926	1.081

estimation even in Design 1, where the 2SProbit model is correctly specified. Therefore, comparing the simulation results between SNLS-EX and SNLS-OR, we can conclude that the nonparametric estimation of the control function λ_τ, rather than the heteroskadasticity l_0, may introduce relatively high variability. Fourth, as shown by the difference between SNLS-OR and SNLS-2S, the bias and the variance contributed by the first-step generated regressor $\hat{v}_i$ are small and even negligible in some cases. This is probably due to the fact that in our simulation the separability of the reduced form equation is utilized to get a consistent parametric estimator of v_i which converges at a faster rate than some nonparametric one does when the equation is non-separable.

Finally, like most of the semiparametric estimation procedures, a major concern about our estimator is its computational stability and complexity. Since two unknown functions need to be estimated nonparametrically and one of them is by nature multivariate, the numerical optimization procedure could be computation intensive, and relatively large sample size is required to guarantee the estimation accuracy. Moreover, the objective function is not concave and the computation may potentially suffer from local minimal which could be densely distributed. Therefore, to better implement our semiparametric estimator in application in practical application, some stable global numerical optimization algorithm is preferable.

References

AHN, H., H. ICHIMURA, AND J. L. POWELL (1996): "Simple estimators for monotone index models," *manuscript, Department of Economics, UC Berkeley.*

AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

BLUNDELL, R., AND J. L. POWELL (2003): "Endogeneity in nonparametric and semiparametric regression models," *ECONOMETRIC SOCIETY MONOGRAPHS*, 36, 312–357.

BLUNDELL, R. W., AND J. L. POWELL (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71(3), 655–679.

CHEN, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

CHEN, X., H. HONG, AND E. TAMER (2005): "Measurement error models with auxiliary data," *The Review of Economic Studies*, 72(2), 343–366.

CHEN, X., AND X. SHEN (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, pp. 289–314.

CHESHER, A. (2010): "Instrumental variable models for discrete outcomes," *Econometrica*, 78(2), 575–601.

CHESHER, A., AND A. M. ROSEN (2013): "What Do Instrumental Variable Models Deliver with Discrete Dependent Variables?," *American Economic Review*, 103(3), 557–62.

GALLANT, A. R., AND D. W. NYCHKA (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica: Journal of the Econometric Society*, pp. 363–390.

HOROWITZ, J. (2009): *Semiparametric and nonparametric methods in econometrics*, vol. 692. Springer.

HOROWITZ, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica: Journal of the Econometric Society*, pp. 505–531.

ICHIMURA, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58(1), 71–120.

IMBENS, G. W., AND W. K. NEWEY (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

KHAN, S. (2013): "Distribution free estimation of heteroskedastic binary response models using Probit/Logit criterion functions," *Journal of Econometrics*, 172(1), 168 – 182.

KIM, J., AND D. POLLARD (1990): "Cube root asymptotics," *The Annals of Statistics*, pp. 191–219.

KLEIN, R. W., AND R. H. SPADY (1993): "An efficient semiparametric estimator for binary response models," *Econometrica: Journal of the Econometric Society*, pp. 387–421.

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

LIAO, J.-C. (2012): "Estimation in Triangular Models of Binary Response under Quantile Restrictions," *Job Market Paper*.

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

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

——— (1988): "Identification of binary response models," *Journal of the American Statistical Association*, 83(403), 729–738.

McFadden, D. L. (1984): "Econometric analysis of qualitative response models," *Handbook of econometrics*, 2, 1395–1457.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67(3), 565–603.

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

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

Smith, R. J., and R. W. Blundell (1986): "An exogeneity test for a simultaneous equation Tobit model with an application to labor supply," *Econometrica: Journal of the Econometric Society*, pp. 679–685.

Appendix

Proof of Theorem 2.4

Proof. It is obvious that the second model defined in (2.6), under Condition 2.5 - Condition 2.6, implies the first one defined in (2.5). In order to show the reverse holds, we need to establish that the conditional probability of $Y = 1$ given X and Z in the first model equals the one in the second model, as indicated by Khan (2013). Given (X, Z), $\Pr(Y = 1|X = x, Z = z) = F_{\epsilon|X,Z}(x'\beta_0 - \lambda_\tau(v))$ in (2.5) under Assumption 2.1 - Assumption 2.3.

Suppose the CDF of ξ is $G_\xi(\cdot)$, which is continuous and strictly increasing on the real line. Then define $\sigma_0(x, z) = \frac{x'\beta_0-\lambda_\tau(v)}{G_\xi^{-1}\circ F_{\epsilon|X,Z}(x'\beta_0-\lambda_\tau(v))}$ if $x'\beta_0 - \lambda_\tau(v) \neq 0$. And based on Assumption Assumption 2.1 - Assumption 2.2, $x'\beta_0 - \lambda_\tau(v) = 0$ with zero probability and can be ignored. In addition, it is easy to see that σ_0 is positive a.e. on $\mathcal{X} \times \mathcal{Z}$: when $x'\beta_0 - \lambda_\tau(v) > 0$, $F_{\epsilon|X,Z}(x'\beta_0 - \lambda_\tau(v)) > \tau$ and then $G_\xi^{-1} \circ F_{\epsilon|X,Z}(x'\beta_0 - \lambda_\tau(v)) > 0$; when $x'\beta_0 - \lambda_\tau(v) < 0$, $F_{\epsilon|x,z}(x'\beta_0 - \lambda_\tau(v)) < \tau$ and then $G_\xi^{-1} \circ F_{\epsilon|X,Z}(x'\beta_0 - \lambda_\tau(v)) < 0$. The continuity of σ_0 follows the fact that each component in the definition of σ_0 is continuous.

With $\sigma_0(x, z)$ defined above, the conditional probability of $Y = 1$ given (x, z) equals $G_\xi(\frac{x'\beta_0-\lambda_\tau(v)}{\sigma_0(x,z)}) = G_\xi(\frac{x'\beta_0-\lambda_\tau(v)}{\frac{x'\beta_0-\lambda_\tau(v)}{G_\xi^{-1}\circ F_{\epsilon|X,Z}(x'\beta_0-\lambda_\tau(v))}}) = G_\xi(G_\xi^{-1} \circ F_{\epsilon|X,Z}(x'\beta_0 - \lambda_\tau(v))) = F_{\epsilon|X,Z}(x'\beta_0 - \lambda_\tau(v))$, which is equal to the conditional probability of $Y = 1$ given (x, z) defined in (2.5). □

Proof of Theorem 2.10

Proof. The proof of Theorem 2.10 follows the consistency theorem in Gallant and Nychka (1987). In order to show the consistency of our sieve NLS estimator, we need to verify the following conditions: compactness, denseness, uniform convergence and identification.

The compactness of the closure of our parameter space is ensured by Assumption 2.7. The finite dimensional parameter $\gamma \in \mathbb{R}^{d_{z_1}}$ belongs to a compact set Γ with respect to Euclidean norm $\|\cdot\|$. A functional of the infinite dimensional parameters $\lambda(\cdot)$ and $l(\cdot)$,

$\phi(w,\theta) \equiv G_\xi((x_1+z_1'\gamma-\lambda(v))\exp(l(x_1,z)))$, is assumed to be in a weighted Holder ball. Then the parameter space Θ is compact with respect to the weighted sup norm $\parallel \phi(w,\cdot) \parallel_{\infty,\omega_1}$.

With weighted Holder ball as the parameter space, the denseness of the sieve space as $n \to \infty$, i.e. $\cup_{n=1}^{\infty}\Theta_n$, can be guaranteed by many known finite dimensional linear sieves, including power series, Fourier series, splines and wavelets.

The main task of this proof is to show that the sample object function uniformly converges to some continuous function (the population object function) as the sample size increases to infinity. That is

$$\sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(\hat{w}_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2| = o_p(1)$$

where $\phi(\hat{w},\theta) \equiv G_\xi((x_1 + z_1'\gamma - \lambda(\hat{v}))\exp(l(x_1,z)))$.

First, we prove that $E[(y_i - \phi(w_i - \phi(w_i,\theta))]^2$ is continuous over Θ with respect to the weighted sup norm $\parallel \cdot \parallel_{\infty,\omega_1}$. For any $\theta_1 \in \Theta$,

$$\begin{aligned}
& |E(y_i - \phi(w_i,\theta_1))^2 - E(y_i - \phi(w_i,\theta))^2| \\
\leq\ & E|2y_i - \phi(w_i,\theta_1) - \phi(w_i,\theta)||\phi(w_i,\theta_1) - \phi(w_i,\theta))| \\
\leq\ & 2E|\phi(w_i,\theta) - \phi(w_i,\theta_1)| \\
\leq\ & 2\int |\phi(w_i,\theta) - \phi(w_i,\theta_1)|(1+ \parallel w_i \parallel^2)^{-\omega_1/2}(1+ \parallel w_i \parallel^2)^{\omega_1/2} f_W dw_i \\
\leq\ & 2 \parallel \phi(w_i,\theta) - \phi(w_i,\theta_1) \parallel_{\infty,\omega_1} \int (1+ \parallel w_i \parallel^2)^{\omega_1/2} f_W dw_i \\
\leq\ & C \parallel \phi(w_i,\theta) - \phi(w_i,\theta_1) \parallel_{\infty,\omega_1}
\end{aligned}$$

Thus the continuity of $E[(y_i - \phi(w_i - \phi(w_i,\theta))]^2$ over Θ follows.

Next, to show the uniform convergence of the sample objective function, we need to deal with the generated regressor $\hat{v}_i$, which is a consistent estimator of the control variable

$v_i = F_{x_1|z}(x_{1i}|z_i)$.

$$\begin{aligned}
& \sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(\hat{w}_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2| \\
\leq & \sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(\hat{w}_i,\theta))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2| \\
& +|\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2| \\
\leq & \sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(\hat{w}_i,\theta))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2| \\
& + \sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2|
\end{aligned}$$

By Lemma 2.14 - Lemma 2.15, we can conclude that

$$\sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(\hat{w}_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2| = o_p(1)$$

The last condition that needs to check is identification. As mentioned in Section 2.3, the identification of $(\gamma_0, \lambda_\tau(\cdot), l_0(\cdot))$ is proved. Therefore, we get $\parallel \hat{\gamma}_n - \gamma_0 \parallel = o_p(1)$ and $\parallel \hat{\theta}_n - \theta_0 \parallel_{\infty,\omega_1} = o_p(1)$. Note that $\parallel \hat{\theta}_n - \theta_0 \parallel_{\infty,\omega_1} = \parallel \phi(\cdot,\hat{\theta}_n) - \phi(\cdot,\theta_0) \parallel_{\infty,\omega_1}$ and ϕ is a bounded function (cdf). Therefore, we can also get $\parallel \hat{\theta}_n - \theta_0 \parallel_\infty = o_p(1)$. □

Lemma 2.14. *As $n \to \infty$,*

$$\sup_{w_i\in\mathbb{R}\times\mathbb{Z}} \sup_{\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(\hat{w}_i,\theta))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2| \to 0$$

Proof. Note

$$\begin{aligned}
& |\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(\hat{w}_i,\theta))^2-\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(w_i,\theta))^2| \\
= & |\frac{1}{n}\sum_{i=1}^{n}(2y_i-\phi(\hat{w}_i,\theta)-\phi(w_i,\theta))(\phi(\hat{w}_i,\theta)-\phi(w_i,\theta))| \\
\leq & \frac{1}{n}\sum_{i=1}^{n}|2y_i-\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)||\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)| \\
\leq & \frac{2}{n}\sum_{i=1}^{n}|\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)|
\end{aligned}$$

Because $|\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)|$ is bounded, we get

$$\begin{aligned}
& \sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(\hat{w}_i,\theta))^2-\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(w_i,\theta))^2| \\
\leq & \frac{2}{n}\sum_{i=1}^{n}\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)| \\
= & 2\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)|
\end{aligned}$$

$$\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)|\leq\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|G_\xi^{(1)}((x_{1i}+z_{1i}'\gamma-\lambda(\tilde{v}_i))\exp(l(x_{1i},z_i)))\lambda^{(1)}(\tilde{v}_i)||\hat{v}_i-v_i|$$

where $\tilde{v}_i$ is some value between $\hat{v}_i$ and v_i. By assumptions, the density of ξ is continuous and $\lambda^{(1)}(\cdot)$ is uniformly bounded. Then

$$\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|\phi(\hat{w}_i,\theta)-\phi(w_i,\theta)|\leq C\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}|\hat{v}_i-v_i|$$

By the uniform convergence of $\hat{v}$ to v, we get

$$\sup_{w_i\in\mathbb{R}\times\mathbb{Z}}\sup_{\theta\in\Theta}|\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(\hat{w}_i,\theta))^2-\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(w_i,\theta))^2|\to 0$$

as $n\to\infty$. □

Lemma 2.15. $\sup_{\theta=(\gamma,\tau,l)\in\Theta}|\frac{1}{n}\sum_{i=1}^{n}(y_i-\phi(w_i,\theta))^2-E(y_i-\phi(w_i,\theta))^2|=o_p(1)$

Proof. First, for fixed $\theta \in \Theta$, since $\{w_i\}$ are i.i.d and $|y_i - \phi(w_i,\theta)| \leq 1 < \infty$, by the weak law of large number, $|\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2| = o_p(1)$ pointwisely.

Next, we need to show the sample objective function satisfies the Lipschitz condition in probability. For any $\theta_1, \theta_2 \in \Theta$,

$$\begin{aligned}
& |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta_1))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta_2))^2| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}|2y_i - \phi(w_i,\theta_1) - \phi(w_i,\theta_2)||\phi(w_i,\theta_1) - \phi(w_i,\theta_2)| \\
\leq\ & \frac{2}{n}\sum_{i=1}^{n}|\phi(w_i,\theta_1) - \phi(w_i,\theta_2)| \\
\leq\ & 2 \parallel \phi(w_i,\theta_1) - \phi(w_i,\theta_2) \parallel_{\infty,\omega_1} (1+ \parallel w_i \parallel^2)^{\omega_1/2}
\end{aligned}$$

Because $\int(1+ \parallel w_i \parallel^2)^{\omega_1/2} f_W dw_i$ is bounded, by Markov's Inequality, $(1+ \parallel w_i \parallel^2)^{\omega_1/2} = O_p(1)$.

Therefore, with the fact that Θ is compact and $E(y_i - \phi(w_i,\theta))^2$ is continuous, by Lemma 2.9 in Newey and McFadden (1994),

$$\sup_{\theta=(\gamma,\tau,l)\in\Theta} |\frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(w_i,\theta))^2 - E(y_i - \phi(w_i,\theta))^2| = o_p(1)$$

□

Proof of Theorem 2.13

Proof. To prove this lemma, we can use some well-established results on convergence rates of sieve estimators since we don't take into account of the first stage generated regressor $\hat{v}_i$ here. We mainly follow Theorem 3.2 in Chen (2007).

First of all, we need to verify the conditions required by the theorem. Note that the first condition is trivially satisfied by our i.i.d observations.

Next, we wish to show there is $C_1 > 0$ such that for all small $\epsilon > 0$,

$$\sup_{\{\theta \in \Theta_n : \| \theta_0 - \theta \|_2 \leq \epsilon\}} Var[(y_i - \phi(w_i, \theta))^2 - (y_i - \phi(w_i, \theta_0))^2] \leq C_1 \epsilon^2$$

Denote $\phi_i(\theta) \equiv \phi(w_i, \theta)$. Since

$$(y_i - \phi_i(\theta))^2 - (y_i - \phi_i(\theta_0))^2 = (\phi_i(\theta_0) - \phi_i(\theta))(2y_i - \phi_i(\theta_0) - \phi_i(\theta))$$

we have

$$\begin{aligned} E[(y_i - \phi_i(\theta))^2 - (y_i - \phi_i(\theta_0))^2]^2 &\leq 4E(\phi_i(\theta_0) - \phi_i(\theta))^2 \\ &= 4 \parallel \theta_0 - \theta \parallel_2^2 \leq 4\epsilon^2 \end{aligned}$$

as desired.

The last condition we need to show is that for any $\delta > 0$, there exists a constant $s \in (0, 2)$ such that

$$\sup_{\{\theta \in \Theta_n : \| \theta_0 - \theta \|_2 \leq \delta\}} |(y_i - \phi(w_i, \theta))^2 - (y_i - \phi(w_i, \theta_0))^2| \leq \delta^s U(w_i)$$

with $E([U(w_i)]^\gamma) \leq C_2$ for some $\gamma \geq 2$. Note

$$|(y_i - \phi_i(\theta))^2 - (y_i - \phi_i(\theta_0))^2| \leq 2|\phi_i(\theta_0) - \phi_i(\theta)|$$

Because $|\phi_i(\theta_0) - \phi_i(\theta)|$ is bounded by 2, this condition is trivially satisfied with $U(w_i) = 4$ and for any $\delta > 0$ there must be some s approaching zero as desired.

Therefore by Theorem 3.2 in Chen (2007),

$$\parallel \tilde{\theta}_n - \theta_0 \parallel_2 = O_p(\delta_n \vee \parallel \theta_0 - \pi_n \theta_0 \parallel_2)$$

where for finite dimensional sieves such as power series and B-splines, $\delta_n = C \cdot \sqrt{\frac{\kappa_{1n} \vee \kappa_{2n}}{n}}$ for some constant $C < \infty$, see e.g. Chen and Shen (1998) As for $\| \theta_0 - \pi_n\theta_0 \|_2$, the deterministic approximation error rate, it depends on the smoothness of the functional space and the sieve space we choose. Note

$$\begin{aligned}
& \| \theta_0 - \pi_n\theta_0 \|_2 \\
= & \| \phi_i(\theta_0) - \phi_i(\pi_n\theta_0) \|\ _2 \\
\leq & \| \phi_i(\gamma_0, \lambda_\tau, l_0) - \phi_i(\pi_n\gamma_0, \pi_n\lambda_\tau, l_0) \|_2 \\
& + \| \phi_i(\pi_n\gamma_0, \pi_n\lambda_\tau, l_0) - \phi_i(\pi_n\gamma_0, \pi_n\lambda_\tau, \pi_n l_0) \|_2
\end{aligned}$$

$$\begin{aligned}
& \| \phi(\gamma_0, \lambda_\tau, l_0) - \phi(\pi_n\gamma_0, \pi_n\lambda_\tau, l_0) \|_2 \\
\leq & \sup_{\xi\in\mathbb{R}} G_\xi^{(1)}(\xi) \cdot \| (z_1'(\gamma_0 - \pi_n\gamma_0) - (\lambda_\tau - \pi_n\lambda_\tau)) \exp(l_0) \|_2 \\
= & C \cdot \| (\lambda_\tau - \pi_n\lambda_\tau) \exp(l_0) \|_2 \\
= & C \cdot \{ \int_w [(\lambda_\tau - \pi_n\lambda_\tau) \exp(l_0)]^2 f_W dw \}^{1/2} \\
\leq & \sup_{w_i} \exp(l_0) \cdot C \cdot \| \lambda_\tau - \pi_n\lambda_\tau \|_2 \\
\leq & C \cdot \| \lambda_\tau - \pi_n\lambda_\tau \|_2
\end{aligned}$$

$$\begin{aligned}
& \| \phi_i(\pi_n\gamma_0, \pi_n\lambda_\tau, l_0) - \phi_i(\pi_n\gamma_0, \pi_n\lambda_\tau, \pi_n l_0) \|_2 \\
\leq & \sup_{\xi\in\mathbb{R}} G_\xi^{(1)}(\xi) \cdot \| (x_1 + z_1'\pi_n\gamma_0 - \pi_n\lambda_\tau)[\exp(l_0) - \exp(\pi_n l_0)] \|_2 \\
\leq & C \| l_0 - \pi_n l_0 \|_2
\end{aligned}$$

Since $\lambda \in \Lambda_{c_1}^s(\mathbb{V}, \omega_1)$ and Assumption 2.12 is satisfied, by Ai and Chen (2003) and Chen, Hong, and Tamer (2005), $\| \lambda_\tau - \pi_n\lambda_\tau \|_2 = O_p(\kappa_{1n}^{-s})$ and $\| l_0 - \pi_n l_0 \|_2 = O_p(\kappa_{2n}^{-s/(d_z+1)})$. Therefore,

$$\| \tilde{\theta}_n - \theta_0 \|_2 = O_p(\sqrt{\frac{\kappa_{1n} \vee \kappa_{2n}}{n}} + \kappa_{1n}^{-s} + \kappa_{2n}^{-s/(d_z+1)})$$

□