

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

RICE UNIVERSITY

Semi- and Non-Parametric Estimation and Testing
of Economic Models

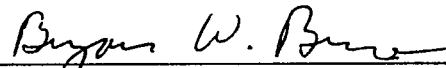
by

Xing Ming

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:



Bryan W. Brown, Chairman
Hargrove Professor of Economics



David Scott
Professor of Statistics



Francis Vella
Assistant Professor of Economics

Houston Texas

April 1995

UMI Number: 9610681

UMI Microform 9610681

Copyright 1996, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI

**300 North Zeeb Road
Ann Arbor, MI 48103**

ABSTRACT

Semi- and Non-Parametric Estimation and Testing of Economic Models

by

Xing Ming

Chapter one provides a new estimator for the ordered polychotomous model. The estimator is based on the use of the average of the standard normal densities with different means as a parametric approximation to the density of the error term. The method also, for the first time, provides a consistent, differentiable estimator of the distribution function of the error term. Chapter two employs the conventional interpretation of endogeneity in econometric models to develop a way of eliminating the inconsistency resulting from endogenous explanators in cross sectional models. The method first obtains an estimate of the unobserved heterogeneity responsible for the endogeneity and then creates a synthetic observation by taking a non-parametric weighted average of nearby observations. The deviations are produced from these synthetic means thereby eliminating the unobserved heterogeneity. The procedure is particularly useful for estimating models when the endogenous regressors are censored or appear non-linearly in the primary equation. Chapter three first calculates the exact distribution of Blum *et al's* (1961) statistic, which is based on a comparison of the sample joint CDF with the product of the sample marginal CDF's, for very small sample size and simulate the distribution quantiles of it for sample size not large enough to employ the asymptotic result. Secondly, the asymptotic distribution of the statistic constructed from residuals and/or predicted values, to test the independence of the error term and the regressors in nonlinear regression models, is obtained. Thirdly, bootstrap technique is used to obtain the distribution quantiles of the statistic constructed from residuals and/or predicted values. The test is nonparametric in that it does not specify the parametric form of distributions of the error term and the regressors.

ACKNOWLEDGMENTS

I thank Dr. Bryan W. Brown, Robin Sickles and Francis Vella for their help and guidance during my stay at Rice University.

I thank Elaine Fortowsky for her proof reading.

I thank the following people for providing me with help during my time-consuming pursuit of the Ph.D. degree.

Professor Adrian Pagan, The University of Rochester and Australian National University

Professor Peide Chen, Colorado State University and the Chinese Academy of Sciences

Professor Michael Jerison, The State University of New York at Albany

Professor Gary M. Quinlivan, St. Vincent College

Professor Wenchuan Mo, Shandong University, China

Professor Rong Wu, Nankai University, China

Mr. Dezheng Tian, Zouping No. 1 High School, China

Mr. Nianwen Wan, , Zouping No. 1 High School and Qingdao No. 9 High School, China

In addition, I thank all my family members, especially my Mom, for their sacrifice and continuing encouragement, without which I would never have had a chance of obtaining a Ph.D. degree.

Contents

Chapter

1	Semi-Parametric Maximum Likelihood Estimation of Discrete Choice Models.....	1
1.1	Introduction	2
1.2	Estimation Methodology	4
1.3	Assumptions	7
1.4	Theoretical Derivations	9
1.5	Monte Carlo Evidence	21
1.6	Extensions to Ordered Discrete Choice Models	25
1.7	Conclusions of Chapter 1	27
2	Semi-Parametric Estimation via Synthetic Fixed Effects.....	28
2.1	Introduction	29
2.2	Model	31
2.3	Estimator	34

2.4	Some Models of Interest	43
2.5	Three Step Synthetic Fixed Effects Estimation	55
2.6	Control Function Procedures versus Synthetic Fixed Effects Proce- dures	57
2.7	Covariance Matrix Estimation	59
2.8	Simulation Evidence	60
2.9	Relationship between synthetic fixed effects estimator and the 2SLS estimator	63
2.10	Conclusions of Chapter 2	63
3	Semi- and Non-Parametric Tests of Independence.....	65
3.1	Introduction	66
3.2	Exact Finite Sample Distributions of Blum <i>et al.</i> 's statistics	68
3.3	Residual-Based Test of Independence of Regressors and Error Term	80
3.4	Test of Serial Independence Using Residuals	94
3.5	Comparisons of Residual-Based Test with Other Tests	97
3.6	Computer Simulations	99
3.7	Conclusions of Chapter 3	103
3.8	Proof of the Theorems in Chapter 3	104
	REFERENCES.....	114

LIST OF TABLES

1.1	Three Estimation Results	24
2.1	Simulation Results for Ordinal Sample Selection Model	61
2.2	Simulation Results for Model with Nonlinear Endogenous Regressors	62
3.1	The Right Hand Tail Exact Distribution of $\pi^4 T_n/2$ for Small Samples	73
3.2	The Exact Distribution Quantiles of $\pi^4 T_n/2$ for Small Samples . . .	74
3.3	Simulated Quantiles of $\pi^4 T_n/2$ for Sample Size Less than or Equal to 200	76
3.4	Power Comparisons with BDS Test	100
3.5	Simulation Results	102

CHAPTER 1

Semi-Parametric Maximum Likelihood Estimation of Discrete Choice Models

Chapter one presents a new semi-parametric method to estimate consistently the parameters of discrete choice models without specifying the distribution function of the error term. The approach utilizes kernel estimation techniques to parameterize the distribution function of the error term and maximum likelihood techniques to obtain the estimates. The estimators are shown to be consistent as long as the number of nuisance parameters goes to infinity as the sample size goes to infinity, and a smooth consistent estimator of the distribution function of the error term is obtained. As the likelihood function of the model can be made as smooth as possible, the approach is easy to implement.

1.1 Introduction

In this chapter we propose a new semiparametric estimator for the binary choice model. Our estimator is shown to be consistent. Furthermore, in the process of estimating the parameters of the model, the distribution function of the error term is simultaneously and consistently estimated. The method is semiparametric in that it makes no assumption concerning the specific distribution generating the disturbances.

The model we consider is given by:

$$y_i^* = V(z_i, \theta_0) - u_i, \tag{1.1}$$

where we only observe $y_i = 1\{y_i^* > 0\}$ and $1\{\cdot\}$ is the indicator function, $V(\cdot, \cdot)$ is a known function, z_i is a vector of exogenous variables, θ_0 is an unknown parameter vector, and u_i is a random disturbance. The subscript i distinguishes observations. In conventional maximum likelihood estimation of this model one has to make assumptions about the distribution of the error term. It is well-known that misspecification of the distribution of the error term may lead to inconsistent parameter estimates.

The problem of distribution-free estimation of binary choice model was first addressed by Manski (1975), who introduced the maximum score estimator and proved its consistency. The most closely related alternative approaches (to the method proposed in this paper) within the class of distribution-free likelihood esti-

mators are Cosslett's (1983), which is an application of the result of Kiefer and Wolfowitz (1956) and Gallant's and Nychka's (1987) which utilized the density function approximation result of Phillips (1983). In the literature of econometrics, very few single models have attracted so many researchers, *e.g.*, Ruud (1983), Cosslett (1987), Han (1987), Ichimura (1987), Manski and Thompson (1986), Horowitz (1992) and Matzkin (1992) and others.

As in Cosslett, our approach provides a consistent estimator of the distribution function of the error term. The fundamental difference in our approach is that Cosslett maximized the likelihood function over the collection of all distribution functions whereas we maximize over a class of distribution functions that can approximate the true distribution function as closely as possible. Because our restricted class of distribution functions can be made as smooth as desired, the method is easy to implement. In addition, our estimator of the distribution function is itself a smooth function, while Cosslett's is a step function. Gallant and Nychka (1987) also maximized over a restricted class of distributions, but their specification of the parametric functions can not guarantee that they are distribution density functions, unless a restriction is imposed, *i.e.* they must restrict the definite integration of the functions to equal to one. This makes it difficult for empirical practitioners to solve the problem numerically. Ruud (1993) proposes an algorithm for computing the semi-parametric maximum likelihood estimator for a class of discrete dependent variable models that include the ordered probit and multinomial choice models with non-parametric distribution functions. In addi-

tion to the theory, this paper also comes up with a very simple approach to the computation of the models mentioned above. The methodology adopted in this paper can also be used in the estimation of other models, *e.g.* Gallant and Nychka (1987).

The most recent investigation of the binary choice model was done by Klein and Spady (1993). Their estimator satisfies all the classical desiderata: consistency, \sqrt{n} -normality, and semiparametric efficiency. But the approach in this paper provides for the first time a consistent smooth estimator of the distribution function of the error term.

This chapter is organized as follows: Section 1.2 is a brief introduction of the estimation method; Section 1.3 enumerates the technical assumptions; Section 1.4 presents the theoretical derivation of the estimator and its properties; Section 1.5 presents a Monte Carlo experiment and Section 1.6 briefly discusses applications of the method to ordered discrete choice models.

1.2 Estimation Methodology

Suppose we know the parametric form of the distribution function of the disturbance term $F(\cdot, \alpha)$, where α is a vector of parameters with α_0 being the truth. The most common way to estimate θ_0 in (1.1) is to apply the method of maximum likelihood under the assumption that u and the exogenous variables z are mutually independent of each other and that u and z are jointly *i.i.d.*. The average

log-likelihood function for this problem is

$$L_N(\theta, F(\cdot, \alpha)) = \frac{1}{N} \sum_{j=1}^N \{y_j \log F[V_j, \alpha] + (1 - y_j) \log(1 - F[V_j, \alpha])\} \quad (1.2)$$

where $V_j = V(z_j, \theta)$. By the strong law of large numbers we know immediately that $L_N(\theta, F)$ goes to $E[L_N(\theta, F(\cdot, \alpha))]$. The explicit form of $E[L_N(\theta, F(\cdot, \alpha))]$ is:

$$\int \{F[V(\theta_0), \alpha_0] \log F[V(\theta), \alpha] + (1 - F[V(\theta_0), \alpha_0]) \log(1 - F[V(\theta), \alpha])\} \mu dz$$

where $\mu \equiv \mu(z)$ is the density function of the explanatory variable vector z and $V(\theta_0) = V(z, \theta_0)$, $V(\theta) = V(z, \theta)$. The above objective function is maximized at (θ_0, α_0) . This observation is the essence of the maximum likelihood methodology. For the sake of completeness we will prove this claim later in this paper.

Like Gallant and Nychka (1987) we will construct a series of parametric distribution functions to approximate the true distribution function. The approximation should be such that it gets as close as possible to the truth when the sample size becomes sufficiently large. Once the series of approximation functions is obtained, it can be substituted in place of the true distribution function in equation (1.2). Estimation of θ and the artificial parameters of the approximating function is then done simultaneously. The estimator $\hat{\theta}$ of θ_0 proves to be consistent. Inserting $\hat{\theta}$ back into the artificially constructed distribution function gives us an estimator of the distribution function. This estimator converges uniformly to the true distribution function. Cosslett (1983) also obtained a consistent estimator of the distribution function, and as a matter of fact, our objective function can in some sense be regarded as a smoothing of Cosslett's objective function.

Define

$$F(t; \underline{\alpha}_n) \equiv F_n(t) \equiv \frac{1}{n} \sum_{i=1}^n K\left(\frac{t - \alpha_i}{h}\right),$$

where we use $\underline{\alpha}_n$ to denote the n -dimensional vector $(\alpha_1, \alpha_2, \dots, \alpha_n)$. $K(\cdot)$ is some cumulative distribution function (currently we regard it as the standard normal cumulative distribution function). and h is the window width¹. For notational simplicity we suppress the dependence on n of h . Given $\underline{\alpha}_n$, $F(t; \underline{\alpha}_n)$ is itself a distribution function.

Some intuition can be gained by writing

$$F(t; \underline{\alpha}_n) \equiv n^{-1} \sum_{i=1}^n K\left(\frac{t - \alpha_i}{h}\right) \equiv \int_{-\infty}^t (nh)^{-1} \sum_{i=1}^n k\left(\frac{s - \alpha_i}{h}\right) ds \equiv F_n(t). \quad (1.3)$$

Assume $(\alpha_1, \alpha_2, \dots, \alpha_n)$ are n random experiments drawn from the population distribution of the error term u ; Then $f_n(s) = (nh)^{-1} \sum_{i=1}^n k\left(\frac{s - \alpha_i}{h}\right)$ is nothing but the nonparametric estimator of the probability density function of the error term u , see Parzen (1962). The uniform convergence of $f_n(s)$ to its target $f(s)$ needs more regularity conditions. In the problem at hand, we use only $F_n(t)$, the integration function of $f_n(s)$. Because of the integration, the uniform convergence of $F_n(t)$ to its target $F(t)$ is guaranteed by the continuity of $F(t)$. The method of this paper is outlined as follows: First, we claim that the non-parametric form distribution function $F(t; \underline{\alpha}_n)$ can approximate any continuous distribution function as closely as possible uniformly; second we plug $F(t; \underline{\alpha}_n)$ into (1.2) in the place of the true

¹ In this paper the only requirement for window-width h is $\lim_{n \rightarrow \infty} h = 0$. From the non-parametric density estimation literature, *e.g.* Pagan and Ullah (1992), we know that the optimal window-width $h = O(n^{-1/5})$. From the perspective of finite sample estimation, this optimality rule for h has no use at all.

distribution function $F(\cdot)$ and maximize the log-likelihood function with respect to both θ and $\underline{\alpha}_n$; third, we prove that providing n increases with the sample size N , the estimator $\hat{\theta}_{nN}$ of θ_0 will be consistent; and finally, we prove that, if we plug the estimator $\hat{\underline{\alpha}}_{nN}$ of $\underline{\alpha}_n$ into $F(t; \underline{\alpha}_n)$, $F(t; \hat{\underline{\alpha}}_{nN})$ is a uniform consistent estimator of $F(\cdot)$.

1.3 Assumptions

As we pointed out earlier, the aforementioned procedure can in some sense be viewed as a smooth version of Cosslett's (1983). Not surprisingly the technical assumptions we are making here are almost the same as those made by Cosslett².

Assumption 1.1. The parameter space \mathbf{Q} is a compact subset of a Euclidean space, with θ_0 being an interior point of \mathbf{Q} .

Assumption 1.2. $V(z, \theta)$ is continuous in θ for each z , and is measurable in z for each θ .

Assumption 1.3. The density function $\mu(z)$ is a measurable function of z . One way of achieving this is to suppose that each component of z is either an absolutely continuous distribution or a discrete distribution, and that $\mu(z)$ is continuous with respect to the continuous components of z .

Assumption 1.4. The probability distribution of $V(z, \theta)$ is absolutely continuous and has full support on the real line.

Assumption 1.5. (Identification Condition): (i) If $F_0[V(z, \theta_0)] = F_1[V(z, \theta_1)]$

² Readers are referred to Cosslett (p.769) for discussion of the following assumptions.

for almost all z (with respect to μ), then $\theta_0 = \theta_1$ and $F_0 = F_1$, where $\theta_0, \theta_1 \in \mathbb{Q}$ and F_0, F_1 are two distribution functions. (ii) $V(z, \theta_0)$ is not homogeneous of degree one in θ_0 .

If $V(z, \theta_0)$ is homogeneous of degree one in θ_0 , *e.g.* the linear case, this assumption can be interpreted as normalizing one of the arguments of θ_0 to one. For discussion of this problem, readers are referred to Cosslett (1983). For more general identification discussions for binary choice model, readers are also referred to Manski (1988) and Matzkin (1992).

Assumption 1.6. The distribution function of the disturbance term has continuous density function $f(\cdot)$ and satisfies $\int t f(t) dt = 0$.

The second part of this assumption seems to be implicit. Because the general form of our parameterization of the density function is free of mean, we single it out for emphasis. Without this assumption even the following model can not be identified:

$$y = 1 \left(1 + \theta x^2 + \varepsilon > 0 \right)$$

Because the constant can be absorbed into the error term such that the mean of the error term becomes 1 and the original regression function becomes linear in θ . In the following, whenever we mention either the distribution space of the error term or the parameterization of it, we always assume implicitly that its mean is zero.

1.4 Theoretical Derivations

From now on let Φ denote the set of continuous distribution functions, *i.e.* $\Phi = \{F(\cdot); F \text{ nondecreasing, continuous, and satisfying } F(-\infty) = 0, F(+\infty) = 1\}$. For any sufficiently small $\sigma > 0$, define the σ -truncation $F^\sigma(t)$ ³ of a function $F(\cdot)$ valued between 0 and 1 as:

$$F^\sigma(t) = \begin{cases} \sigma & \text{if } F(t) < \sigma \\ F(t) & \text{if } \sigma \leq F(t) \leq 1 - \sigma \\ 1 - \sigma & \text{if } F(t) > 1 - \sigma \end{cases}$$

Define $\Phi^\sigma = \{F^\sigma(\cdot); F(\cdot) \in \Phi\}$, $\Phi_n = \{F(\cdot, \underline{\alpha}_n), \underline{\alpha}_n \in \mathbf{R}^n\}$, $\Phi_n^\sigma = \{F^\sigma(\cdot, \underline{\alpha}_n), \underline{\alpha}_n \in \mathbf{R}^n\}$, where $F(\cdot, \underline{\alpha}_n)$ is defined by (1.3) for some given symmetric kernel $k(\cdot)$. By definition, all sets Φ_n^σ are subsets of Φ^σ .

Lemma 1.1. Assume $F(t) \in \Phi$, then there exists a real sequence $\{\alpha_i\}$ such that $F(t; \underline{\alpha}_n) = n^{-1} \sum_{i=1}^n K\left(\frac{t - \alpha_i}{h}\right)$ converges to $F(t)$ uniformly; and $F^\sigma(t; \underline{\alpha}_n)$ converges to $F^\sigma(t)$ uniformly, *i.e.* the σ -truncated sequence also goes to the σ -truncation of $F(\cdot)$ uniformly.

Proof: (i) Let α_i be the i -th realization of the random experiment according to population distribution $F(\cdot)$. Once the experiment is done, $\{\alpha_i\}$ becomes a non-stochastic real sequence. I claim that any such sequence $\{\alpha_i\}$ satisfies our

³ For any given distribution function $F(t)$, $F^\sigma(t)$ is not a distribution function. By controlling σ we can easily control the absolute distance of the two functions, *i.e.*

$$\sup_{t \in \mathbf{R}} |F(t) - F^\sigma(t)| \leq \sigma$$

requirements. By construction

$$\lim_{n \rightarrow \infty} K\left(\frac{t - \alpha_i}{h}\right) = \begin{cases} 1 & \text{if } \alpha_i < t \\ 1/2 & \text{if } \alpha_i = t \\ 0 & \text{if } \alpha_i > t \end{cases}$$

Thus

$$\begin{aligned} \lim_{n \rightarrow \infty} F(t; \underline{\alpha}_n) &= \lim_{n \rightarrow \infty} \left\{ \sum_{\alpha_i < t} 1 \{\alpha_i < t\} + \sum_{\alpha_i = t} 1 \{\alpha_i = t\} / 2 \right\} / n \\ &= E[1 \{\alpha_i < t\}] + E[1 \{\alpha_i = t\}] / 2 \\ &= F(t) \end{aligned}$$

The second equality is by the strong law of large numbers. When we use the strong law of large numbers we treat $\{\alpha_i\}$ as a random variable with distribution function $F(\cdot)$. The third equality is because $E[1 \{\alpha_i = t\}] = 0$ due to the fact that $F(\cdot)$ is continuous.

By continuity of $F(\cdot)$ and page 21, Problem 3 of Billingsley (1968) the uniform convergence result can be obtained.

Define a metric $d(\cdot, \cdot)$ on the space Φ (Φ^σ) as

$$d(F_1(\cdot), F_2(\cdot)) = \int_0^\infty |F_1(t) - F_2(t)| e^{-|t|} dt.$$

The completion of Φ (Φ^σ) with respect to the metric d is denoted by $\bar{\Phi}$ ($\bar{\Phi}^\sigma$), which includes all the nondecreasing functions valued between 0 (σ) and 1 ($1 - \sigma$)⁴.

Define $\bar{\Gamma} = \mathbf{Q} \times \bar{\Phi}$, which is also a compact space with respect to the metric δ in $\Gamma = \mathbf{Q} \times \Phi$. δ is defined as the sum of the distance d plus the Euclidean distance in

⁴ This claim will be proven in Lemma 3.

Q. The metric $d(\cdot, \cdot)$ and δ can also be defined on the spaces Φ^σ and $\Gamma^\sigma = \mathbf{Q} \times \Phi^\sigma$ in exactly the same way as above. We use bars to denote the completion of the corresponding metric, *e.g.* $\bar{\Gamma}^\sigma$ denotes the completion of Γ^σ .

Lemma 1.2. For given σ , $\lim_{n \rightarrow \infty} d(F_n^\sigma, F^\sigma) = 0$ if and only if $\lim_{n \rightarrow \infty} F_n^\sigma(t_0) = F^\sigma(t_0)$ for all continuous points t_0 of $F^\sigma(\cdot)$, i.e. convergence in metric $d(\cdot, \cdot)$ is equivalent to pointwise convergence at all continuity points of the limiting function.

Proof: Suppose $\lim_{n \rightarrow \infty} F_n^\sigma(t_0) \neq F^\sigma(t_0)$ for some continuous point t_0 of $F^\sigma(\cdot)$. Without loss of generality we assume that for all n , $F_n^\sigma(t_0) - F^\sigma(t_0) > \alpha > 0$. Because of continuity of $F^\sigma(t)$ at t_0 , there exists a $d > 0$ such that whenever $t \in [t_0, t_0 + d]$, $F^\sigma(t) - F^\sigma(t_0) < \alpha/2$. Thus $F_n^\sigma(t) - F^\sigma(t) \geq F_n^\sigma(t_0) - F^\sigma(t) \geq \alpha/2$ for all n and all $t \in [t_0, t_0 + d]$. We have

$$\begin{aligned} \lim_{n \rightarrow \infty} d(F_n^\sigma, F^\sigma) &= \lim_{n \rightarrow \infty} \int_0^\infty |F_n^\sigma(t) - F^\sigma(t)| e^{-|t|} dt \\ &\geq \lim_{n \rightarrow \infty} \int_{t_0}^{t_0+d} |F_n^\sigma(t) - F^\sigma(t)| e^{-|t|} dt \\ &\geq \alpha/2 \int_{t_0}^{t_0+d} e^{-|t|} dt > 0 \end{aligned}$$

which is a contradiction.

Suppose $\lim_{n \rightarrow \infty} F_n^\sigma(t_0) = F^\sigma(t_0)$ for all continuous point t_0 of $F^\sigma(\cdot)$. Due to the fact that $F^\sigma(\cdot)$ is a monotone function, the set of its discontinuity points is countable. A countable set has a zero Lebesgue measure. Thus by the dominated convergence theorem, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} d(F_n^\sigma, F^\sigma) &= \lim_{n \rightarrow \infty} \int_0^\infty |F_n^\sigma(t) - F^\sigma(t)| e^{-|t|} dt \\ &= \int_0^\infty \lim_{n \rightarrow \infty} |F_n^\sigma(t) - F^\sigma(t)| e^{-|t|} dt \end{aligned}$$

$$= 0$$

Lemma 1.3. For any $F^\sigma(\cdot) \in \bar{\Phi}^\sigma$, there exists a subsequence $\{F_{nk}^\sigma(\cdot); F_{nk}^\sigma(\cdot) \in \Phi_{nk}^\sigma\}$ such that $\lim_{n \rightarrow \infty} d(F_{nk}^\sigma, F^\sigma) = 0$, i.e. $\cup_{n=1}^\infty \Phi_n^\sigma$ is a dense subset of $\bar{\Phi}^\sigma$ under metric $d(\cdot, \cdot)$.

Proof: This is an extension of Lemma 1.1 and Lemma 1.2. If $F^\sigma(\cdot) \in \Phi^\sigma$, then by Lemma 1.1 and Lemma 1.2 there exists a sequence $\{F_n^\sigma(\cdot); F_n^\sigma(\cdot) \in \Phi_n^\sigma\}$ satisfying our requirement. The only thing that needs to be proven is that for any nondecreasing function $F(\cdot)$, which might be discontinuous, there exists a sequence $F_n^\sigma(\cdot) \in \Phi^\sigma$ such that $d(F_n^\sigma, F^\sigma) = 0$, or we can say $\bar{\Phi}^\sigma$ contains all the nondecreasing functions between σ and $1 - \sigma$. The following proof is also a justification of footnote 3.

First we notice that the discontinuity points of a nondecreasing function $F^\sigma(\cdot)$ are countable. We order them as $\{x_i\}$. By Lemma 1.2 the only thing we need to prove is that for any continuous point t of $F^\sigma(\cdot)$ we can find a sequence of continuous functions such that $\lim_{n \rightarrow \infty} F_n^\sigma(t) = F^\sigma(t)$. Next we will construct this functional sequence.

For any $\epsilon > 0$, we use O_i to denote the neighborhood $(x_i - \epsilon/2^i, x_i + \epsilon/2^i)$ of x_i . The sum of the Lebesgue measures of these neighborhoods is less than ϵ . Denote each connection of these neighborhood as (a_i, b_i) , which is at most countable. Define $F_\epsilon^\sigma(\cdot)$ as

$$F_\epsilon^\sigma(x) = \begin{cases} \text{linear function connecting } F(a_i) \text{ and } F(b_i) & \text{if } x \in (a_i, b_i) \\ F^\sigma(x) & \text{otherwise} \end{cases}$$

By construction $F_\epsilon^\sigma(x)$ is continuous. If we choose a positive decreasing ϵ_n having a limit 0, we obtain a functional sequence $F_n^\sigma(\cdot)$ corresponding to ϵ_n . I claim $\lim_{n \rightarrow \infty} F_n^\sigma(t) = F^\sigma(t)$ for any continuity point of $F^\sigma(t)$. Suppose t is a continuity point of $F^\sigma(\cdot)$. There exists N such that for all $n > N$, no discontinuity points of $F^\sigma(\cdot)$ belong to $(t - \epsilon_n, t + \epsilon_n)$, from which we know that t does not belong to any (a_i, b_i) corresponding to ϵ_n . By definition of $F_n^\sigma(\cdot)$, we get $F_n^\sigma(t) = F^\sigma(t)$ for all $n > N$ i.e. $\lim_{n \rightarrow \infty} F_n^\sigma(t) = F^\sigma(t)$. By Lemma 1.2 this pointwise convergence is equivalent to convergence in metric $d(\cdot, \cdot)$.

Lemma 1.4. (Identification Lemma) Assumption 1.5, (i) is equivalent to the following: there exists small enough $\sigma > 0$ such that if $F_0^\sigma[V(z, \theta_0)] = F_1^\sigma[V(z, \theta_1)]$ for almost all z (with respect to μ), then $\theta_0 = \theta_1$ and $F_0^\sigma = F_1^\sigma$.

Proof: We need only to prove that Assumption 1.5, (i) implies this Lemma. If there does not exist a $\sigma > 0$ such that it makes $\theta_0 = \theta_1$ and $F_0^\sigma = F_1^\sigma$, then we let σ go to zero, $\theta_0 = \theta_1$ and $F_0 = F_1$ can not hold even we have $F_0[V(z, \theta_0)] = F_1[V(z, \theta_1)]$. This is a contradiction to Assumption 1.5, (i).

Lemma 1.5. (Information Inequality)⁵ Assume that u and the exogenous variables z are mutually independent of each other, that u and z are jointly i.i.d., and that the true distribution function of the error term is F , then under Assumption 1.5, we can claim (θ_0, F) is the only solution to $\max_{(\theta, F)} E[L_N(\theta, F)]$ for $(\theta, F) \in \bar{\Gamma}$;

⁵ The most commonly used information inequality is the following

$$\int f(x, \theta_0) \log f(x, \theta) dx \leq \int f(x, \theta_0) \log f(x, \theta_0) \text{ for any } \theta,$$

where $f(x, \theta)$ is any probability density function, e.g. Robinson (1991).

and (θ_0, F^σ) is the only solution to $\max_{(\theta, F)} E[L_N(\theta, F)]$ for $(\theta, F) \in \bar{\Gamma}^\sigma$, where

$$L_N(\theta, F) = \frac{1}{N} \sum_{j=1}^N \{y_j \log F[V(z_j, \theta)] + (1 - y_j) \log(1 - F[V(z_j, \theta)])\}$$

Proof: We prove only the first part of this Lemma. The second part can be proven accordingly. First we notice that if α is between 0 and 1 then α solves

$$\max_x [\alpha \log(x) + (1 - \alpha) \log(1 - x)]$$

But

$$\begin{aligned} & \max_{(\theta, F)} E[L_N(\theta, F)] \\ &= \max_{(\theta, F)} \int \{F[V(\theta_0)] \log F[V(\theta)] + (1 - F[V(\theta_0)]) \log(1 - F[V(\theta)])\} \mu dz \\ &= \int \max_{(\theta, F)} \{F[V(\theta_0)] \log F[V(\theta)] + (1 - F[V(\theta_0)]) \log(1 - F[V(\theta)])\} \mu dz \end{aligned}$$

i.e. (θ_0, F) is one solution. Where $V(\theta_0) = V(z, \theta_0)$ and $V(\theta) = V(z, \theta)$. Assume there is another solution $(\bar{\theta}, F)$ to this problem. Then we have $F[V(z, \bar{\theta})] = F[V(z, \theta_0)]$. By Identification Assumption 1.5, this is impossible.

$$\int \{F[V(z, \theta_0)] \log F[V(z, \theta_0)] + (1 - F[V(z, \theta_0)]) \log(1 - F[V(z, \theta_0)])\} \mu dz$$

is termed as the “entropy” of this problem. This Lemma says the maximum value of the expected log-likelihood is the value of its entropy. Some authors, *e.g.* Gallant and Nychka (1987), call $E[L(\theta, F)] \leq E[L(\theta_0, F)]$ the “Information Inequality”.

Lemma 1.6. Assume the number of elements of $\mathbf{F} \subset \bar{\Phi}^\sigma$ is infinite. Then there exists a sequence $\{F_n^\sigma(\cdot)\} \subset \mathbf{F}$ and a nondecreasing real function $F_0^\sigma(\cdot) \in \bar{\Phi}^\sigma$ such that $F_n^\sigma(t)$ goes to $F_0^\sigma(t)$ for all the continuous points of $F_0^\sigma(\cdot)$, *i.e.* $\bar{\Phi}^\sigma$ is compact under metric $d(\cdot, \cdot)$.

Proof: We use the diagonal method to prove our claim. Suppose all rational numbers are ordered by $\{r_n\}$. We can find $\{F_{1n}^\sigma(\cdot)\} \subset \mathbf{F}$ such that $F_{1n}^\sigma(r_1)$ has a limit v_1 ; we can also find a subsequence $\{F_{2n}^\sigma(\cdot)\} \subset \{F_{1n}^\sigma(\cdot)\}$ such that $F_{2n}^\sigma(r_2)$ has a limit v_2 ; \dots . When we continue this process we can find a subsequence of $F_{k-1n}^\sigma(\cdot)$, $F_{kn}^\sigma(\cdot)$ such that when n goes to infinity $F_{kn}^\sigma(r_s)$ go to v_s for all $s \leq k$. Choose $F_n^\sigma(\cdot) = F_{nn}^\sigma(\cdot)$ and define $F_0^\sigma(r_s) = v_s$. Then by construction $F_n^\sigma(\cdot)$ goes to $F_0^\sigma(\cdot)$ for all rational numbers. The value of $F_0^\sigma(\cdot)$ for all irrational numbers can be defined by the right limit of the values of $F_0^\sigma(\cdot)$ at the rational points. Function $F_0^\sigma(\cdot)$ satisfies our requirements.

Theorem 1.1. $-L_N(\theta, F)$ converges to $E[-L_N(\theta, F)]$ uniformly with respect to $(\theta, F) \in \bar{\Gamma}^\sigma$. Where

$$\begin{aligned}
& -L_N(\theta, F) \\
&= -\frac{1}{N} \sum_{j=1}^N \left\{ 1\{y_j^* > 0\} \log F[V(z_j, \theta)] + 1\{y_j^* < 0\} \log(1 - F[V(z_j, \theta)]) \right\} \\
&\equiv -\frac{1}{N} \sum_{j=1}^N \left\{ 1\{V(z_j, \theta_0) - u_j > 0\} \log F[V(z_j, \theta)] \right. \\
&\quad \left. + 1\{V(z_j, \theta_0) - u_j \leq 0\} \log(1 - F[V(z_j, \theta)]) \right\} \\
&\equiv \frac{1}{N} \sum_{j=1}^N \{-s(u_j, z_j, \theta, F, \theta_0)\} \\
& \\
& E[-L_N(\theta, F)] \\
&= - \int \{F[V(\theta_0)] \log F[V(\theta)] + (1 - F[V(\theta_0)]) \log(1 - F[V(\theta)])\} \mu dz
\end{aligned}$$

Proof: We prove this theorem using Theorem 1 of Burguete *et. al* (1982).

Noting that all the summands and the integrands are positive, we choose the dominant function $b(\epsilon, z) \equiv -2\log(\sigma) > 0$. By the dominated convergence theorem,

the target function $E[-L_N(\theta, F)]$ is continuous in (θ, F) with respect to the metric δ of $\bar{\Gamma}^\sigma$. This is the place we have to restrict the function $F(\cdot) \in \bar{\Phi}^\sigma$. Without this restriction we can not use the well-known dominated convergence theorem. Assumptions 1, 2 and 3 of Burguete, Gallant and Souza (1982) are satisfied trivially.

Due to the fact that the indicator function appearing in $-s(\epsilon, z, \theta, F, \theta_0)$ is not continuous in its arguments, Theorem 1 of Bruguete *et al.* can not be utilized directly. We shall approximate the discontinuous function $-s(\epsilon, z, \theta, F, \theta_0)$ by a continuous function $-s^\gamma(\epsilon, z, \theta, F, \theta_0)$ to which Theorem 1 of Burguete *et al.* (1982) applies and show that the approximation error can be made arbitrarily small. The following technique is adopted from Gallant and Nychka (1987).

Let $l(x)$ be a continuous function with $0 \leq l(x) \leq 1$, $l(x) = 1$ for $x \geq 0$, and $l(x) = 0$ for $x \leq -\gamma$. Let $m(x)$ be a continuous function with $0 \leq m(x) \leq 1$, $m(x) = 1$ for $x \leq 0$, and $m(x) = 0$ for $x \geq \gamma$. The two functions satisfy $l(x) \geq 1\{x > 0\}$, $m(x) \leq 1\{x \leq 0\}$ and $\lim_{\gamma \rightarrow 0} l(x) = 1\{x > 0\}$, $\lim_{\gamma \rightarrow 0} m(x) = 1\{x \leq 0\}$.

Let

$$\begin{aligned} & -s^\gamma(u, z, \theta, F, \theta_0) \\ = & -l(V(z, \theta_0) - u) \log F[V(z, \theta)] - m(V(z, \theta_0) - u) \log(1 - F[V(z, \theta)]) \end{aligned}$$

which is a continuous function dominated by continuous function $-2\log(\sigma)$. Let

$$L_N^\gamma(\theta, F) = \frac{1}{N} \sum_{j=1}^N \{s^\gamma(u_j, z_j, \theta, F, \theta_0)\}$$

then

$$E[L_N^\gamma(\theta, F)]$$

$$\begin{aligned}
&= E[s^\gamma(u, z, \theta, F, \theta_0)] \\
&= \iint \{l\{V(z, \theta_0) - u > 0\} \log F[V(z, \theta)] \\
&\quad + m\{V(z, \theta_0) - u \leq 0\} \log(1 - F[V(z, \theta)])\} \mu(z) du dz
\end{aligned}$$

$$\begin{aligned}
&\sup |L_N(\theta, F) - E[L_N(\theta, F)]| \\
&\leq \sup |L_N^\gamma(\theta, F) - E[L_N^\gamma(\theta, F)]| + \sup |L_N^\gamma(\theta, F) - L_N(\theta, F)| \\
&\quad + \sup |E[L_N(\theta, F)] - E[L_N^\gamma(\theta, F)]|
\end{aligned}$$

where all the suprema are taken with respect to $(\theta, F) \in \bar{\Gamma}^\sigma$

The first term of the right hand side satisfies the requirements of Theorem 1 of Burguete *et al.* (1982); Thus we have

$$\lim_{n \rightarrow \infty} \sup_{(\theta, F) \in \bar{\Gamma}^\sigma} |L_N^\gamma(\theta, F) - E[L_N^\gamma(\theta, F)]| = 0.$$

For the second term we have

$$\begin{aligned}
&\sup_{(\theta, F) \in \bar{\Gamma}^\sigma} |L_N^\gamma(\theta, F) - L_N(\theta, F)| \\
&= \sup_{(\theta, F) \in \bar{\Gamma}^\sigma} \frac{1}{N} \left| \sum_{j=1}^N \log F[V_j(\theta)] \{l(V_j(\theta_0) - u_j) - 1\{V_j(\theta_0) > u_j\}\} \right| \\
&\quad + \sup_{(\theta, F) \in \bar{\Gamma}^\sigma} \frac{1}{N} \left| \sum_{j=1}^N \log(1 - F[V_j(\theta)]) \{m(V_j(\theta_0) - u_j) - 1\{V_j(\theta_0) \leq u_j\}\} \right| \\
&\leq -\frac{1}{N} \log(\sigma) \left| \sum_{j=1}^N \{l(V_j(\theta_0) - u_j) - 1\{V_j(\theta_0) > u_j\}\} \right| \\
&\quad - \frac{1}{N} \log(\sigma) \left| \sum_{j=1}^N \{m(V_j(\theta_0) - u_j) - 1\{V_j(\theta_0) \leq u_j\}\} \right| \\
&\leq -\frac{1}{N} \log(\sigma) \sum_{j=1}^N \{1\{V_j(\theta_0) \leq u_j \leq V_j(\theta_0) + \gamma\} \\
&\quad + 1\{V_j(\theta_0) - \gamma \leq u_j \leq V_j(\theta_0)\}\}
\end{aligned}$$

This last RHS term can not exceed

$$-\frac{1}{N} \log(\sigma) \sum_{j=1}^N 1\{V_j(\theta_0) - \gamma \leq u_j \leq V_j(\theta_0) + \gamma\}$$

which converges to

$$K(\gamma) \equiv -\log(\sigma) \int \{F[V(z, \theta_0) + \gamma] - F[V(z, \theta_0) - \gamma]\} \mu(z) dz$$

$K(\gamma)$ is continuous in γ by the dominated convergence theorem and $K(0) = 0$.

For any N the third term is also less than $K(\gamma)$. Thus we have

$$\lim_{N \rightarrow \infty} \sup_{(\theta, F) \in \bar{\Gamma}^\sigma} |L_N(\theta, F) - E[L_N(\theta, F)]| \leq 2K(\gamma).$$

$K(\gamma)$ can be made smaller than any given $\epsilon > 0$. The left hand side is fixed and ϵ is arbitrary so the limit is zero, *i.e.*

$$\lim_{N \rightarrow \infty} \sup_{(\theta, F) \in \bar{\Gamma}^\sigma} |L_N(\theta, F) - E[L_N(\theta, F)]| = 0.$$

From now on we use $(\hat{\theta}_{nN}, \hat{F}_{nN})$ to denote the solution to $\max L_N(\theta, F(\cdot))$, $(\theta, F) \in \mathbf{Q} \times \Phi_n$, and $(\hat{\theta}_{nN}^\sigma, \hat{F}_{nN}^\sigma)$ to denote the solution to $\max L_N(\theta, F(\cdot))$, $(\theta, F) \in \mathbf{Q} \times \Phi_n^\sigma$, where

$$L_N(\theta, F(\cdot)) = \frac{1}{N} \sum_{j=1}^N \{y_j \log F[V(z_j, \theta)] + (1 - y_j) \log(1 - F[V(z_j, \theta)])\}$$

Theorem 1.2. (σ -truncation estimation) For small enough $\sigma > 0$ such that the Identification Lemma 1.4 is satisfied, we have

$$\lim_{n, N \rightarrow \infty} |\hat{\theta}_{nN}^\sigma - \theta_0| = 0, \text{ almost surely,}$$

$$\lim_{n, N \rightarrow \infty} d(\hat{F}_{nN}^\sigma, F^\sigma) = 0, \text{ almost surely.}$$

Proof: By Theorem 0 of Gallant and Nychka (1987). We will check that all the requirements of the Theorem are satisfied.

(a) Compactness: $\bar{\Phi}^\sigma$ is compact with respect to metric $d(\cdot, \cdot)$ by Lemma 1.6.

(b) Denseness: By Lemma 1.3, $\cup_{n=1}^{\infty} \Phi_n^\sigma$ is a dense subset of $\bar{\Phi}^\sigma$ under metric $d(\cdot, \cdot)$. (note: In Theorem 0 of Gallant and Nychka (1987) condition (b) requires that the subsets Φ_n^σ are nondecreasing. In our definition the Φ_n^σ do not satisfy this requirement. When going through the proof of Theorem 0, I found out that the proof is still valid if Φ_n^σ is inserted in place of their subset sequences.)

(c) Uniform convergence: This is by Theorem 1.1.

(d) Identification: This is guaranteed by the σ we have chosen.

Theorem 1.2 has the following implications:

(i) Theorem 1.2 says that provided n and N go to infinity the estimators will be strongly consistent, but it does not specify which, n or N , should go to infinity faster or which should be larger. If the number of nuisance parameters n is smaller than the sample size N , then intuitively everything should be fine. What if the number of nuisance parameters is larger than the sample size? This is also acceptable. First the existence of a solution is guaranteed because of the extreme nonlinearity of the objective function; second our maximization can be regarded as a restricted version of Cosslett's (1983). Cosslett maximized the objective function with respect to all distribution functions and still came up with consistent estimators. We maximize the objective function with respect to a restricted class of distribution functions, so of course the maximum problem has a solution.

(ii) Theoretically the σ -truncation of all the functions is not a problem as it appears to be. First, it does not affect the consistency of the estimator $\hat{\theta}_{nN}$ of θ ; secondly, the distribution function estimator $F^\sigma(\cdot, \hat{\alpha}_{nN})$ can be made to approximate the true distribution function $F(\cdot)$ as closely as we want by prechoosing a small enough $\sigma > 0$. For empirical practitioners this σ -truncation makes the objective function no longer differentiable, and will cause computation difficulties. However σ -truncation is nothing but a restriction of the tail behavior of the distribution function. With σ -truncation the tails of the distribution function are prevented from going to 0 or 1 respectively, so that the dominated convergence theorem can be applied. In fact, the tails of the approximating functions must only go to 0 or 1 at a slower rate than the tails of the distribution function of the error term do in order for the dominated convergence theorem to be valid. Intuitively, if the tails of the error term distribution are not “fat”, the σ -truncation will not be necessary. Assumption (C.4a)⁶ of Klein and Spady (1993) (p. 392) guarantees another situation which the σ -truncation is not required for strong consistency of our estimator. Thus one suggestion for empirical practitioners is to apply this technique directly without the σ -truncation. If there is no truncation involved in the objective function, it is very easy for empirical practitioner to implement. General software packages can solve this problem. Another justification for direct implementation is that there is no difference between the σ -truncation estimation and

⁶ The assumption is as follows: There exist \underline{P} and \bar{P} that do not depend on z such that $0 < \underline{P} \leq F[V(z, \theta_0)] \leq \bar{P} < 1$. This assumption, which requires that $V(z, \theta_0)$ be a bounded random variable, serves to bound the probability function $F[V(z, \theta_0)]$ away from 0 and 1. Our σ -truncation serves this purpose.

the non-truncation estimation from the perspective of small sample estimation.

(iii) Ad Hoc inference. Fix $n = 5, 10$ or whatever point at which we have strong confidence that the approximating function $F_n(t) \equiv \frac{1}{n} \sum_{i=1}^n \Phi\left(\frac{t-\alpha_i}{h}\right)$ will approximate the true $F(t)$ to a satisfactory degree such that we can use the aforementioned approximation as the true distribution function of the error term and go ahead to obtain the MLE estimator $\hat{\theta}_N$. Then $\hat{\theta}_N$ should be normally distributed and do inferences based on $\hat{\theta}_N$ and its variance.

1.5 Monte Carlo Evidence

The model we are considering is

$$y = 1 \left(-10.5 + 7z + \theta_0 x^3 > u \right)$$

where z is distributed as $\chi^2(1)$, x is distributed as $U[-4, 4]$ and $u = \varepsilon^3$, where ε is the uniform distribution on $[-3, 3]$. Under the assumption that $\theta_0 = 1$, the average log-likelihood function for this problem is

$$L = \frac{1}{N} \sum_{j=1}^N \{y_j \log F(V_j) + (1 - y_j) \log (1 - F(V_j))\}$$

where $V_j = -10.5 + 7z + x_j^3$, $F(\cdot)$ the distribution function of the error term u .

If we parameterize $F(\cdot)$ as

$$F[t, \alpha] = \frac{1}{n} \left\{ \sum_{i=1}^{n-1} \Phi\left(\frac{t - \alpha_i}{h}\right) + \Phi\left(\frac{t + \sum_{i=1}^{n-1} \alpha_i}{h}\right) \right\}$$

where $\Phi(\cdot)$ is the CDF of standard normal distribution. This kind of parameterization satisfies $\int t dF[t, \alpha] = 0$, the requirement of Assumption 1.6. Then the

derivative of L with respect to θ is

$$\frac{\partial L}{\partial \theta} = \frac{1}{N} \sum_{j=1}^N \left\{ \frac{y_j f[V_j, \alpha]}{F[V_j, \alpha]} - \frac{(1 - y_j) f[V_j, \alpha]}{1 - F[V_j, \alpha]} \right\} x_j^3$$

where

$$f[t, \alpha] = \frac{1}{nh} \left\{ \sum_{i=1}^{n-1} \phi\left(\frac{t - \alpha_i}{h}\right) + \phi\left(\frac{t + \sum_{i=1}^{n-1} \alpha_i}{h}\right) \right\}$$

The derivatives with respect to the nuisance parameters are:

$$\begin{aligned} \frac{\partial L}{\partial \alpha_i} = & \frac{1}{N} \sum_{j=1}^N \left\{ \frac{(1 - y_j) \left[\phi\left(\frac{V_j - \alpha_i}{h}\right) - \phi\left(\frac{V_j + \sum_{i=1}^{n-1} \alpha_i}{h}\right) \right]}{nh(1 - F[V_j, \alpha])} \right. \\ & \left. - \frac{y_j \left[\phi\left(\frac{V_j - \alpha_i}{h}\right) - \phi\left(\frac{V_j + \sum_{i=1}^{n-1} \alpha_i}{h}\right) \right]}{nhF[V_j, \alpha]} \right\} \end{aligned}$$

with $\phi(\cdot)$ the density function of the standard normal.

Let

$$\begin{aligned} A(\theta; \alpha) &\equiv \left(\frac{\partial L}{\partial \theta}(\theta; \alpha); \frac{\partial L}{\partial \alpha_1}(\theta; \alpha), \dots, \frac{\partial L}{\partial \alpha_n}(\theta; \alpha) \right)' \\ A_j(\theta; \alpha) &\equiv \left(\frac{\partial L_j}{\partial \theta}(\theta; \alpha); \frac{\partial L_j}{\partial \alpha_1}(\theta; \alpha), \dots, \frac{\partial L_j}{\partial \alpha_n}(\theta; \alpha) \right)' \\ B_1(\theta; \alpha) &= \frac{\partial A(\theta; \alpha)}{\partial(\theta; \alpha')} \\ B_2(\theta; \alpha) &= \frac{1}{N} \sum_{j=1}^N [A_j(\theta; \alpha)] [A_j(\theta; \alpha)]' \end{aligned}$$

with

$$\begin{aligned} \frac{\partial L_j}{\partial \theta} &= \left\{ \frac{y_j f[V_j, \alpha]}{F[V_j, \alpha]} - \frac{(1 - y_j) f[V_j, \alpha]}{1 - F[V_j, \alpha]} \right\} x_j^3 \\ \frac{\partial L_j}{\partial \alpha_i} &= \frac{(1 - y_j) \left[\phi\left(\frac{V_j - \alpha_i}{h}\right) - \phi\left(\frac{V_j + \sum_{i=1}^{n-1} \alpha_i}{h}\right) \right]}{nh(1 - F[V_j, \alpha])} \\ &\quad - \frac{y_j \left[\phi\left(\frac{V_j - \alpha_i}{h}\right) - \phi\left(\frac{V_j + \sum_{i=1}^{n-1} \alpha_i}{h}\right) \right]}{nhF[V_j, \alpha]} \end{aligned}$$

We will use a Quasi-Newton method to solve the nonlinear equations: $A(\theta; \alpha) =$

0. The Taylor series expansion of $A(\theta; \alpha)$ around an arbitrary $\gamma_0 \equiv (\theta_0; \alpha_0)$ is

$$A(\gamma) \simeq A(\gamma_0) + B_1(\gamma_0) (\gamma - \gamma_0) = 0$$

Solving for γ and then equating γ to γ_{i+1} and γ_0 to γ_i , we obtain the iteration

$$\gamma_{i+1} = \gamma_i - [B_1(\gamma_i)]^{-1} A(\gamma_i)$$

We substitute $B_2(\gamma_i)$ for $B_1(\gamma_i)$ and use the following iteration

$$\gamma_{i+1} = \gamma_i - [B_2(\gamma_i)]^{-1} A(\gamma_i)$$

I have a Gauss program to solve this problem. The program can be modified to solve any binary choice model semi-parametrically with this method and is available on request. The following is a report on my experiments:

(1) The number of the nuisance parameters can not be large. Otherwise chances are that the matrix B_2 will become singular during the process of iteration. This may be because that the variance of the regressor is not big enough.

(2) The estimate is very sensitive to the window width h . This is due to the possibilities that some h will make the target function have many local maximum points and that some h will make the parameterized CDF closer to the true CDF of the error term than others.

(3) The Gauss program is fairly easy to write because we are using the Quasi-Newton method instead of the Newton's method. If we use the Newton's method, we have to calculate the second derivatives of the maximum likelihood function

Initial value θ	Initial value α	Initial value h	$\hat{\theta}$
0.9	3.0	1.440	0.943
1.2	3.0	1.432	0.965
1.1	3.5	1.435	0.977

Table 1.1: Three Estimation Results

with respect to the unknown parameters and the nuisance parameters. It is really involved.

In Table 1 we report three estimation results. All of them are obtained by setting $n = 2$, *i.e.* we choose the parameterized distribution function of the error term as

$$F(t) = \frac{1}{2} \left[\Phi \left(\frac{t - \alpha}{h} \right) + \Phi \left(\frac{t + \alpha}{h} \right) \right]$$

The first column of Table 1.1 is the estimated value of θ_0 , $\hat{\theta}$, second column the initial value for θ when we do our iteration, third column the initial value for the nuisance parameter and fourth the initial value for window width h . The estimation is conducted when the sample size is equal to 500.

Our three estimates are not very good for a sample size of 500. In fact we are still subject to misspecification, because we are only using the average of two normal distribution functions to represent the true distribution function of the error term.

1.6 Extensions to Ordered Discrete Choice Models

This method can be applied to the estimation of other ordered choice models⁷.

In this section I will present the implementation of our method in ordered discrete choice models. Proofs concerning the validity of the implementation are the same as those given earlier for the binary choice model. Before we proceed we shall define a set of constants γ_i such that $\gamma_1 = -\infty$, $\gamma_m = +\infty$, and $\gamma_1 < \gamma_2 < \dots < \gamma_m$. $\gamma = (\gamma_2, \dots, \gamma_{m-1})$ is regarded as a set of parameters to be estimated.

The ordered choice model is

$$Y_i = V(z_i, \theta) - u_i,$$

where Y is the underlying response variable, z is a vector of exogenous variables, θ is a vector of parameters, and u is the error term with unknown distribution function $F(\cdot)$. Y is not observed, but we know that if it belongs to the j th category it will satisfy $\gamma_{j-1} < Y < \gamma_j$, $j = 1, 2, \dots, m$. Because Y is observed only ordinally, we have to be careful about the identification problem, *i.e.* we have to assume that $V(z, \theta)$ is not homogeneous of degree one in θ . If $V(z, \theta)$ is homogeneous of degree one in θ we just normalize any argument of θ to be one.

We shall define a set of ordinal variables:

$$l_{ij} = \begin{cases} 1 & \text{if } \gamma_{j-1} < Y_j < \gamma_j \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \dots, N; j = 1, \dots, m$$

⁷ Maddala (1983) (Chapter 2) studies extensively discrete regression models. Under regularity conditions, our method can be used to estimate the discrete regression models semi-parametrically.

The log likelihood function for this model is

$$L_N = \frac{1}{N} \sum_{i,j} l_{ij} \log\{F[\gamma_j - V(z_i, \theta)] - F[\gamma_{j-1} - V(z_i, \theta)]\}.$$

As in the estimation of the binary choice model, let

$$F_n(t) \equiv F_n(t; \underline{\alpha}_n) \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{K}\left(\frac{t - \alpha_i}{h}\right).$$

Substitute $F_n^\sigma(t)$ into the above equation for $F(\cdot)$ to get L_N^σ and maximize L_N^σ with respect to α_n , $\gamma = (\gamma_2, \dots, \gamma_{m-1})$ and θ . Denote the corresponding estimators by $\hat{\underline{\alpha}}_{nN}^\sigma$, $\hat{\gamma}_{nN}^\sigma$ and $\hat{\theta}$. Our estimator of the distribution function $F(t)$ is $\hat{F}_{nN}(t) \equiv F_n^\sigma(t; \hat{\underline{\alpha}}_{nN}^\sigma)$.

Theorem 1.3. (σ -truncation estimation) Under the assumptions of 1.1-1.6, and for small enough $\sigma > 0$ such that the Identification Lemma 1.4 is satisfied, we have

$$\lim_{n,N \rightarrow \infty} |\hat{\gamma}_{nN}^\sigma - \gamma| = 0, \text{ almost surely,}$$

$$\lim_{n,N \rightarrow \infty} |\hat{\theta}_{nN}^\sigma - \theta_0| = 0, \text{ almost surely,}$$

$$\lim_{n,N \rightarrow \infty} d(\hat{F}_{nN}, F^\sigma) = 0, \text{ almost surely.}$$

For empirical practitioners, we have the same advice as in the estimation of the binary choice model, namely, that the σ -truncation be ignored. As we pointed earlier in the estimation of binary choice model, provided that the tails of the approximating functions go to 0 and 1 at a slower rate than the tails of the distribution function of the error do or that Assumption (C.4a) of Klein and Spady is satisfied, we can get the strongly consistent estimator without the σ -truncation.

1.7 Conclusions of Chapter 1

Kernel estimation techniques and maximum likelihood method are used jointly to estimate discrete choice models without specifying the distribution of the error term. Thus the misspecification problem is avoided and the estimators are proven to be strongly consistent. The insight of this paper is that nonparametric kernel estimation formula is directly used to parameterize the distribution function of the error term. From the kernel estimation literature, we know this is a very good parameterization. A Monte Carlo study shows the method works.

Under stronger regularity conditions, this method may be used in the estimation of commonly-encountered models semiparametrically, *e.g.* any models that can be estimated by maximum-likelihood method unless the density function of the disturbance term is not known. The estimation strategy is to approximate the *density* function parametrically with kernel estimation techniques and to maximize the likelihood function with respect to the parameters of the model and the nuisance parameters of the density function simultaneously. The kinds of regularity conditions needed to be imposed are being investigated in the estimation of the models studied by Gallant and Nychka (1987).

CHAPTER 2

Semi-Parametric Estimation via Synthetic Fixed Effects

This chapter employs the conventional interpretation of endogeneity in econometrics to develop a way of eliminating the inconsistency resulting from endogenous explanators in cross-sectional models. We obtain an estimate of the unobserved heterogeneity responsible for the endogeneity and re-order the data such that the distance between individuals is increasing in the difference in the unobserved heterogeneity. We create a synthetic “average” observation for each individual by taking a non-parametric weighted average of nearby observations. We produce deviations from these synthetic means thereby eliminating the unobserved heterogeneity. While our approach is applicable to the conventional simultaneous equation model it is most attractive, due to its relatively weak distributional assumptions, for models with censored endogenous regressors or selection bias. Our procedure is also useful for models when the endogenous regressor appears non-linearly in the primary equation.

2.1 Introduction

This chapter employs the conventional interpretation of endogeneity in econometrics to develop a new way of eliminating the inconsistency resulting from endogenous explanators in cross sectional models. We do so by adapting some ideas from panel data estimation to cross sectional models. We argue that cross sectional data can be organized such that the unobserved heterogeneity generating the endogeneity can be eliminated through appropriate data transformations. Unlike panel data, however, where the arrangement of the data is obvious and generally done on the assumption that the unobserved heterogeneity is individual specific, our approach requires an estimate of the unobserved heterogeneity. With this estimate we re-order the data as though we have observations on “increasingly dissimilar” individuals and create a synthetic average observation for each individual by taking a non-parametric weighted average of nearby observations. Deviations from synthetic means thus eliminate the unobserved heterogeneity. As this methodology is closely linked to the fixed effects, or within estimator, for panel data we call our approach synthetic fixed effects estimation.

While our procedure is applicable to the conventional simultaneous model system it is most attractive for models with censored endogenous regressors (see, for example, Heckman 1978 and Vella 1993) and variations on the sample selection model pioneered by Heckman (1979) and extended by others (see, for example, Olsen 1980, Lee 1982, Garen 1984 and Vella 1993). The reason for this appeal is

related to distributional assumptions. Models with censored endogenous regressors or models estimated over non-randomly chosen subsets of the data typically require strong distributional assumptions. These distributional assumptions are employed to; i) estimate the reduced form equation of the censored regressor or the variable generating the selectivity; and ii) express the error from the primary equation as some known function of the reduced form error. In estimating the sample selection models our approach requires no distributional assumption regarding the reduced form error. It also relaxes the assumption that the relationship between the two error terms is known, as assumed in Heckman (1979) and others. Nor do we need to specify the manner in which the relationship is approximated as is required in Lee (1982), Gallant and Nychka (1987), Newey (1988) and Vella (1993). Our approach is more in the spirit of Powell (1989) and Ahn and Powell (1993) which eliminate the unobserved heterogeneity. In this sense the estimator is semi-parametric.

While the advantages of our approach are also enjoyed by the methodology employed in Powell (1989) and Ahn and Powell (1993) our estimator can also be applied to a wider range of models involving censored endogenous regressors and systems with non-conventional forms of selectivity bias. It can also be employed for the estimation of treatment effects and for models with alternative forms of censoring for the endogenous regressors. As several of these models can be estimated without any distributional assumptions our estimator is semi-parametric.

Our procedure is also useful for an additional family of models in which the endogenous regressor appears non-linearly in the conditional mean of the primary

equation. While these models can be estimated by instrumental variables our approach avoids the search for instruments. This is a substantial attraction given the inaccuracies induced by poor instruments (see, for example, Pagan and Jung 1993).

The following section outlines the generic model under focus. Section 2.3 derives the estimation procedure and outlines its properties. Section 2.4 examines some models, characterized by endogenous censored regressors and sample selectivity, where our procedure is attractive. We also consider models where the endogenous regressor appears non-linearly in the conditional mean of the endogenous variable of primary interest. Section 2.5 outlines how our procedure can be augmented with an additional step to estimate parameters which are unidentified with the estimator in Section 2.3. Section 2.6 discusses the advantages of our estimator compared with the control function procedure which is often employed for several of the models described in Section 2.4. Section 2.7 outlines a strategy for estimating the covariance matrix. Section 2.8 provides some simulation evidence and Section 2.9 presents concluding comments.

2.2 Model

Consider the following model:

$$y_i^* = x_i\beta + z_i\gamma + u_i, \quad i = 1, \dots, n \quad (2.1)$$

$$z_i^* = w_i\theta + v_i, \quad i = 1, \dots, N \quad (2.2)$$

$$y_i = l(y_i^*) \quad (2.3)$$

$$z_i = g(z_i^*) \quad (2.4)$$

$$D_{ji} = 1 \text{ iff } z_i^* \in A_j \text{ and } D_{ji} = 0 \text{ otherwise, } j = 1, \dots, J \quad (2.5)$$

where y_i^* and z_i^* are endogenous latent variables; w_i is a row vector of exogenous variables; x_i is subset of w_i ; β , γ and θ are parameters to be estimated; l and g are functions mapping y_i^* and z_i^* into the observed values y_i and z_i ; A_j are subsets of the real line; D_{ji} is the indicator function of the event $z_i^* \in A_j$; u_i and v_i are zero mean error terms. Equation (2.1) is of primary interest and equation (2.2) is the reduced form representation of the endogenous explanator. N denotes the entire sample whereas $n \leq N$ represents some systematically chosen subset. Furthermore, we make the following assumptions regarding the structure of the model:

Assumption A:

(A1) $E(v_i|w_i) = 0$, $E(u_i|w_i) = 0$ and $cov(u_i, v_i) \neq 0$.

(A2) The structural error can be expressed as a function of the reduced form error plus some random component

$$u_i = E(u_i|v_i) + [u_i - E(u_i|v_i)] = f(v_i) + e_i \quad (2.6)$$

where $f(t) = E(u_i|v_i = t)$ is an unknown function; $e_i = u_i - E(u_i|v_i)$ and e_i is independent of $(z_j, w_j, v_j)_{j=1}^N$ for all $i = 1, \dots, N$.

(A3) (Identification) $\dim(x_i) \leq \dim(w_i) - 1$.

Using (A2) we substitute equation (2.6) into (2.1) to get

$$y_i^* = x_i\beta + z_i\gamma + f(v_i) + e_i. \quad (2.7)$$

The ordinary least squares (OLS) estimates from (2.1) will be inconsistent due

to the endogeneity of z_i by (A1). The conventional way to estimate (2.1), when $y^* = y$ and (A3) is satisfied, is instrumental variables. However, instrumental variables is not applicable when observations on y_i^* are not available for ranges of z_i^* (*i.e.* sample selection). Accordingly, it is useful to consider an estimator which is appropriate for a wider class of models.

Consider the following strategy while assuming that y_i^* is observed. First note it is useful to consider v_i as an unobserved individual heterogeneity which simultaneously influences y_i^* and z_i . It can be loosely considered a “fixed” individual effect as it is individual specific. Accordingly, if we had multiple observations for the same individual we could eliminate this effect through appropriate data transformations. However, even with a single cross section we can eliminate the heterogeneity if we can identify observations with values of v which are “near” to each other. For example suppose we identify multiple pairs of observations i and j which have similar values for the error term v . Thus the following transformed equation

$$(y_i^* - y_j^*) \approx (x_i - x_j)\beta + (z_i - z_j)\gamma + (e_i - e_j) \quad (2.8)$$

can be consistently estimated by OLS as the unobserved heterogeneity responsible for the inconsistency, $f(v_i) - f(v_j)$, has been eliminated. This methodology is employed in panel data estimation as the metric of “closeness” is obvious. Deaton (1985) also employs this approach to construct artificial panels on the basis of cohort membership over repeated cross sections. In single cross sectional econo-

metric studies, however, it has been confined to the work of Powell (1989), Ahn and Powell (1993), and few others. The following section outlines a estimation procedure based on residuals to define closeness of observations.

2.3 Estimator

To derive our estimation procedure assume that y_i^* and z_i^* are observed and $z_i = z_i^*$. Taking the expectation of equation (2.7) conditional on v_i gives

$$E[y_i^*|v_i] = E[x_i|v_i]\beta + E[z_i|v_i]\gamma + f(v_i). \quad (2.9)$$

We can now eliminate the unobserved heterogeneity responsible for the endogeneity of z_i by subtracting (2.9) from (2.7) to get

$$y_i^* - E[y_i^*|v_i] = [x_i - E(x_i|v_i)]\beta + [z_i - E(z_i|v_i)]\gamma + e_i. \quad (2.10)$$

In general v_i is unobserved. However, if we obtain a \sqrt{N} -consistent estimate of θ , denoted $\hat{\theta}$, we can substitute the residual $\hat{v}_i = z_i^* - w_i\hat{\theta}$ in place of v_i in equation (2.10). We rewrite equation (2.10) as

$$y_i^* - E[y_i^*|\hat{v}_i] = [x_i - E(x_i|\hat{v}_i)]\beta + [z_i - E(z_i|\hat{v}_i)]\gamma + \varepsilon_i \quad (2.11)$$

where $\varepsilon_i = e_i + f(v_i) - E[f(v_i)|\hat{v}_i]$. Equation (2.11) is consistently estimated by OLS provided ε_i is asymptotically uncorrelated with these constructed regressors.

This requires $E[f(v_i)|\hat{v}_i] \rightarrow E[f(v_i)|v_i] = f(v_i)$.

Lemma 2.1. Assume $g(\hat{\theta}, z_i^* - w_i\hat{\theta}) = E[f(v_i)|z_i^* - w_i\hat{\theta}]$ and $G_i(\theta^*) = g_1(\theta^*, z_i^* - w_i\theta^*) - g_2(\theta^*, z_i^* - w_i\theta^*)w_i$ are bounded, then¹

¹ $E[f(v_i)|z_i^* - w_i\hat{\theta}]$ is typically only a function of $z_i^* - w_i\hat{\theta}$, e.g. $E[f(v_i)|z_i^* - w_i\hat{\theta} = \cdot]$

(a)

$$E[f(v_i) | \hat{v}_i] \rightarrow E[f(v_i) | v_i] = f(v_i)$$

(b)

$$E[m(\hat{v}_i) | x_i] \rightarrow E[m(v_i) | x_i]$$

where $g_1(s, t) = \partial g(s, t) / \partial s$, $g_2(s, t) = \partial g(s, t) / \partial t$ and $m(\cdot)$ is any continuous function. Convergence is *in probability* or *in probability 1* depending on whether the convergence of $\hat{\theta}$ to θ is *in probability* or *in probability 1*.

Proof: (a) By Taylor's series expansion

$$g(\hat{\theta}, z_i^* - w_i \hat{\theta}) = f(v_i) + G_i(\theta^*)(\hat{\theta} - \theta)$$

where $f(v_i) = g(\theta, z_i^* - w_i \theta)$ and θ^* lies between $\hat{\theta}$ and θ . By assumption

$$G_i(\theta^*) = g_1(\theta^*, z_i^* - w_i \theta^*) - g_2(\theta^*, z_i^* - w_i \theta^*) w_i$$

is bounded. Thus $g(\hat{\theta}, z_i^* - w_i \hat{\theta}) = f(v_i) + o_p(1)$.

(b) $E[m(\hat{v}_i) | x_i]$ is continuous in \hat{v}_i and $\hat{v}_i = z_i^* - w_i \hat{\theta}$ is continuous in $\hat{\theta}$. Thus $E[m(\hat{v}_i) | x_i]$ is continuous in $\hat{\theta}$.

Lemma 2.1 implies that equation (2.11) satisfies the orthogonality conditions for OLS *asymptotically*. Thus under fairly general conditions the OLS estimates of equation (2.11) will provide consistent estimates of β and γ . While this procedure provides no advantages over existing instrumental variables estimators in the $= g(\cdot)$. However different $\hat{\theta}$ will generate different $g(\cdot)$ functions. Accordingly we index $E[f(v_i) | z_i^* - w_i \hat{\theta} = \cdot]$ by $g(\hat{\theta}, \cdot)$.

conventional setting we outline below many models where it is attractive².

Before proceeding to these models we state our main theorem which is employed throughout the paper. First, however, we require some additional assumptions regarding the generation of the conditional expectations **(K)**, the choice of bandwidth h **(H)**, the data generating process **(D)**, and the manner in which the \hat{v}_i are estimated **(Θ)**:

Assumptions K:

(K1) $K(c)$ is a non-negative function on the real line bounded by K^* .

(K2) $K(c)$ has compact support A .

(K3) $K(c) \geq \tau I_B$ for some $\tau > 0$ and some closed and connected interval B centered at the origin and having positive Lebesgue measure.

(K4) $K(c)$ is continuously differentiable and $\max_c |K'(c)| \leq \bar{K}$, a finite constant.

(K5) $\int K(c) dc = 1$ and $\int |c \log |c||^{1/2} |dK(c)| < \infty$.

Assumption H:

$\lim_{N \rightarrow \infty} h = 0$ and $\lim_{N \rightarrow \infty} Nh^4 = \infty$.

Assumptions D:

(D1) All the exogenous variables are bounded.

(D2) The density function of the error term v , $p_v(v)$, is uniformly continuous.

Assumption Θ: The parameters θ and their estimators $\hat{\theta}$ belong to a compact set Θ and $\sqrt{N}(\hat{\theta} - \theta) = O_p(1)$.

² The exact relationship between our estimator and two stage least squares is outlined in the Appendix B.

Definition. For any series of data observations $(x_i, y_i)_{i=1}^N$, we define

$$\begin{aligned} p_N(x, t) &= \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_i - t}{h}\right) \\ r_N(y, x, t) &= \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_i - t}{h}\right) y_i. \end{aligned}$$

The Nadaraya-Watson non-parametric kernel estimate of the conditional expectation of y given x is defined as

$$\hat{E}[y|x=t] = r_N(y, x, t) / p_N(x, t) \quad (2.12)$$

where h is the bandwidth satisfying Assumption H.

Lemma 2.2. If Assumptions K, H and (D2) are satisfied the non-parametric density estimator of v will converge to the true density uniformly, i.e.

$$\sup_t |p_N(v, t) - p_v(t)| \rightarrow 0 \text{ a.s.} \quad \text{as } N \rightarrow \infty.$$

Proof: From Assumption H, we have $Nh^2 \rightarrow \infty$. Thus Theorem A in Silverman (1978) applies.

Lemma 2.3. If Assumptions K, H, D and Θ are satisfied, then

$$\begin{aligned} & \max_j \left| \hat{E}(\xi_i | \hat{v}_j) - \hat{E}(\xi_i | v_j) \right| \\ &= \max_j \left| \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{\hat{v}_i - \hat{v}_j}{h}\right) \xi_i \right. / \left. \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{\hat{v}_i - \hat{v}_j}{h}\right) \right. \\ & \quad \left. - \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{v_i - v_j}{h}\right) \xi_i \right. / \left. \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{v_i - v_j}{h}\right) \right| = o_p(1) \end{aligned}$$

where $(\xi_i) = (w_i, z_i)$ and $\hat{v}_i = z_i^* - w\hat{\theta}$ for $i = 1, 2, \dots, N$.

Proof: Taylor's series expansion gives

$$\begin{aligned} & K\left(\frac{\hat{v}_i - \hat{v}_j}{h}\right) \\ &= K\left(\frac{v_i - v_j}{h}\right) - \frac{1}{h} K' \left(\frac{(z_i^* - z_j^*) - (w_i - w_j)\theta^*}{h} \right) (w_i - w_j) (\hat{\theta} - \theta) \end{aligned}$$

where θ^* is between $\hat{\theta}$ and θ . Thus

$$\begin{aligned}\hat{E}(\xi_j|\hat{v}_j) &= \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{\hat{v}_i - \hat{v}_j}{h}\right) \xi_i \Big/ \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{\hat{v}_i - \hat{v}_j}{h}\right) \\ &= \left[\frac{1}{Nh} \sum_i K\left(\frac{v_i - v_j}{h}\right) \xi_i - \frac{1}{\sqrt{N}h^2} \Pi_N^j \sqrt{N} (\hat{\theta} - \theta) \right] \\ &\quad \div \left[\frac{1}{Nh} \sum_i K\left(\frac{v_i - v_j}{h}\right) - \frac{1}{\sqrt{N}h^2} \Pi_{N\xi}^j \sqrt{N} (\hat{\theta} - \theta) \right]\end{aligned}$$

where

$$\begin{aligned}\Pi_N^j &= \frac{1}{N} \sum_i K' \left(\frac{(z_i^* - z_j^*) - (w_i - w_j) \theta^*}{h} \right) (w_i - w_j) \\ \Pi_{N\xi}^j &= \frac{1}{N} \sum_i K' \left(\frac{(z_i^* - z_j^*) - (w_i - w_j) \theta^*}{h} \right) (w_i - w_j) \xi_i.\end{aligned}$$

By Assumption Θ , $\sqrt{N}(\hat{\theta} - \theta) = O_p(1)$. By Assumption (D1)

$$\Pi_N^j \text{ and } \Pi_{N\xi}^j.$$

are bounded uniformly in j . Thus

$$\begin{aligned}\max_j |\Pi_N^j \sqrt{N} (\hat{\theta} - \theta)| &= O_p(1) \\ \max_j |\Pi_{N\xi}^j \sqrt{N} (\hat{\theta} - \theta)| &= O_p(1).\end{aligned}$$

By Assumption H,

$$\begin{aligned}\frac{1}{\sqrt{N}h^2} \max_j |\Pi_N^j \sqrt{N} (\hat{\theta} - \theta)| &= o_p(1) \\ \frac{1}{\sqrt{N}h^2} \max_j |\Pi_{N\xi}^j \sqrt{N} (\hat{\theta} - \theta)| &= o_p(1).\end{aligned}$$

So

$$\begin{aligned}\hat{E}(\xi_j|\hat{v}_j) &= \left[\frac{1}{Nh} \sum_i K\left(\frac{v_i - v_j}{h}\right) \xi_i + o_p(1) \right] \\ &\quad \Big/ \left[\frac{1}{Nh} \sum_i K\left(\frac{v_i - v_j}{h}\right) + o_p(1) \right]\end{aligned}$$

holds uniformly in j . By Lemma 2.2, when N is large, $\frac{1}{Nh} \sum_i K\left(\frac{v_i - v_j}{h}\right)$ converges to $f_v(v_j)$ uniformly in j , and must be strictly positive. Thus our claim follows, *i.e.*

$$\max_j \left| \hat{E}(\xi_i | \hat{v}_j) - \hat{E}(\xi_i | v_j) \right| = o_p(1).$$

Lemma 2.4. If Assumptions K, H, D and Θ are satisfied, then

$$\frac{1}{N} \sum_{j=1}^N \left[E(\xi_j | v_j) - \hat{E}(\xi_j | \hat{v}_j) \right]^2 = o_p(1)$$

Proof: From Lemma 2.3

$$\begin{aligned} & \frac{1}{N} \sum_{j=1}^N \left[E(\xi_j | v_j) - \hat{E}(\xi_j | \hat{v}_j) \right]^2 \\ &= \frac{1}{N} \sum_{j=1}^N \left[E(\xi_j | v_j) - \hat{E}(\xi_j | v_j) + o_p(1) \right]^2 \\ &= \frac{1}{N} \sum_{j=1}^N \left[E(\xi_j | v_j) - \hat{E}(\xi_j | v_j) \right]^2 \\ & \quad + \frac{2}{N} \sum_{j=1}^N \left[E(\xi_j | v_j) - \hat{E}(\xi_j | v_j) \right] \cdot o_p(1) + o_p(1). \end{aligned}$$

From Theorem 1 of Devroye and Wagner (1980), if $E|\xi_j|^\alpha < \infty$, which is guaranteed by (D1) for $\alpha = 1, 2$, then

$$\int \left| E(\xi_j | v) - \hat{E}(\xi_j | v) \right|^\alpha F_v(dv) = o_p(1)$$

where $F_v(\cdot)$ is the probability measure induced by the random variable v . By the Law of Large Numbers

$$\begin{aligned} & \frac{1}{N} \sum_{j=1}^N \left[E(\xi_j | v_j) - \hat{E}(\xi_j | v_j) \right]^\alpha \\ &= \int \left| E(\xi_j | v) - \hat{E}(\xi_j | v) \right|^\alpha F_v(dv) + o_p(1) = o_p(1). \end{aligned}$$

This equality holds for $\alpha = 1, 2$.

Theorem 2.1. If the following conditions are satisfied:

Condition 1: Assumption Θ is satisfied and the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta} - \theta)$ is V_θ ;

Condition 2: The conditional expectations in equation (2.11) are estimated by the kernel method shown in equation (2.12);

Condition 3: Assumptions (A1), **K**, **H** and **D** are satisfied;

then

i) The OLS estimate of $\delta = [\beta', \gamma']'$, $\hat{\delta}$, from equation (2.11) is \sqrt{N} -consistent.

ii) $\sqrt{N}(\delta - \hat{\delta}) \xrightarrow{d} N(0, V_\delta)$ with $V_\delta = \sigma_e^2 Q_s^{-1} + Q_s^{-1} Q_{sg} V_\theta Q_{sg}' Q_s^{-1}$, where $\sigma_e^2 Q_s^{-1}$ is given by the OLS variance formula of $\hat{\delta}$; $Q_s = E(s_i' s_i)$, where $s_i = (x_i - E(x_i|v_i), z_i - E(z_i|v_i))$ and $Q_{sg} = E[s_i' G_i(\theta)]$.

Proof: Let $S_i = (\xi_i - \hat{E}(\xi_i|v_i))$, where $\xi = (x, z)$,

$$\sqrt{N}(\hat{\delta} - \delta) = \left(\frac{1}{N} \sum_i S_i' S_i \right)^{-1} \left[\frac{1}{\sqrt{N}} \sum_i S_i' e_i - \left(\frac{1}{N} \sum_i S_i' G_i(\theta^*) \right) \sqrt{N}(\hat{\theta} - \theta) \right]$$

First we prove the following 3 convergence claims:

(a)

$$\frac{1}{N} \sum_i S_i' S_i \xrightarrow{p} E(s_i' s_i)$$

(b)

$$\frac{1}{N} \sum_i S_i' G_i(\theta^*) \xrightarrow{p} E(s_i' G_i(\theta))$$

(c)

$$\frac{1}{\sqrt{N}} \sum_i S_i' e_i \xrightarrow{d} N(0, \sigma_e^2 E(s_i' s_i))$$

Proof of (a):

$$\begin{aligned} S_i &= (\xi_i - \hat{E}(\xi_i|\hat{v}_i)) = [(\xi_i - E(\xi_i|v_i)) + (E(\xi_i|v_i) - \hat{E}(\xi_i|\hat{v}_i))] \\ \frac{1}{N} \sum_i S'_i S_i &= \frac{1}{N} \sum_i [s'_i s_i + s'_i \Delta_i + \Delta'_i s_i + \Delta'_i \Delta_i] \end{aligned}$$

where $\Delta_i = E(\xi_i|v_i) - \hat{E}(\xi_i|\hat{v}_i)$. By the Law of Large numbers

$$\frac{1}{N} \sum_i s'_i s_i \xrightarrow{a.s.} E(s'_i s_i).$$

By Schwartz's Inequality

$$\frac{1}{N} \sum_i s'_i \Delta_i \leq \left[\left(\frac{1}{N} \sum_i s_i'^2 \right) \left(\frac{1}{N} \sum_i \Delta_i^2 \right) \right]^{1/2} = o_p(1).$$

The equality holds since $\sum_i s_i'^2/N \rightarrow E(s'_i)^2$ a.s. and $\sum_i \Delta_i^2/N = o_p(1)$ by Lemma

2.4. By the same logic $\sum_i \Delta'_i s_i/N = o_p(1)$ and $\sum_i \Delta'_i \Delta_i/N = o_p(1)$. Thus

$$\frac{1}{N} \sum_i S'_i S_i = \frac{1}{N} \sum_i s'_i s_i + o_p(1) = E(s'_i s_i) + o_p(1)$$

Proof of (b) is similar to that of (a).

Proof of (c):

$$\frac{1}{\sqrt{N}} \sum_i S'_i e_i = \frac{1}{\sqrt{N}} \sum_i s'_i e_i + \frac{1}{\sqrt{N}} \sum_i (E(\xi_i|v_i) - \hat{E}(\xi_i|\hat{v}_i)) e_i. \quad (2.13)$$

By the Central Limit Theorem

$$\frac{1}{\sqrt{N}} \sum_i s'_i e_i \xrightarrow{d} N(0, \sigma_e^2 E(s'_i s_i)). \quad (2.14)$$

By assumption A1, e_i 's are independent of each other and independent of $E(\xi_j|v_j) - \hat{E}(\xi_j|\hat{v}_j)$ for $j = 1, \dots, N$. Thus the variance of the second part is

$$V_{II} \equiv \sigma_e^2 E \left[\frac{1}{N} \sum_i (E(\xi_i|v_i) - \hat{E}(\xi_i|\hat{v}_i))^2 \right].$$

By assumption (D1), we know that both $E(\xi_i|v_i)$ and $\hat{E}(\xi_i|\hat{v}_i)$ are bounded.

By Lemma 2.4 it will also go to zero in L_1 .

Thus far we have shown

$$\frac{1}{\sqrt{N}} \sum_i \left(E(\xi_i|v_i) - \hat{E}(\xi_i|\hat{v}_i) \right) e_i \xrightarrow{p \text{ and } d} 0. \quad (2.15)$$

From equations (2.13), (2.14) and (2.15), we get

$$\frac{1}{\sqrt{N}} \sum_i S'_i e_i \xrightarrow{d} N\left(0, \sigma_e^2 E(s'_i s_i)\right).$$

By the definition of e_i the correlation between the random variables $\frac{1}{\sqrt{N}} \sum_i S'_i e_i$ and $\left(\sum_i S'_i G_i(\theta^*)\right) \sqrt{N}(\hat{\theta} - \theta)$ is zero. Thus the random variables $N\left(0, \sigma_e^2 E(s'_i s_i)\right)$ and $N\left(0, \sigma_v^2 E(w'_i w_i)\right)$ are asymptotically independent. The independence of the two random variables gives us the two parts of the variance.

If v_i is observed we can take conditional expectations in equation (2.11) with respect to v_i and the asymptotic variance is given by $\sigma_e^2 Q_s^{-1}$. However since the residuals are estimated, the asymptotic variance of the estimator is inflated by $Q_s^{-1} Q_{sg} V_\theta Q'_{sg} Q_s^{-1}$. This latter term characterizes the finite sample correlation between the error term and the regressors in equation (2.11).

Thus the estimation procedure is as follows. First we estimate equation (2.2) by some procedure which provides \sqrt{N} -consistent estimates $\hat{\theta}$ and \hat{v}_i . We then use the kernel method to estimate the conditional expectations of the triple (y_i^*, z_i, x_i) in equation (2.11). We transform the data to produce deviations from synthetic means and perform OLS.

Our estimator is closely linked, in spirit, to the procedures proposed by Robinson (1988) and Powell (1989) although there are some essential differences. First contrast our procedure with Robinson's estimator. The major difference is that we condition on the estimate of v_i , \hat{v}_i , while Robinson assumes that the conditioning set is observed. Conditioning on the unobserved heterogeneity represents a major advantage for the types of models on which we now focus. The motivation for the methodology of Powell is the use of a single index to define the closeness of observations in the estimation of sample selection models inspired by the work of Heckman (1979). Powell argues that the sample selectivity is generated by a single index and thus defines "closeness" on the basis of this single index. In Section 2.4 we show that the estimators of Powell (1989) and Ahn and Powell (1993) based on the single index are in fact the same as based on the residual \hat{v}_i due to the nature of the mapping from the single index to the residual.

2.4 Some Models of Interest

Thus far we have considered endogenous variables which were fully observed. To extend our approach we require additional assumptions regarding the censoring mechanism and the sample selection. In doing so we employ the unusual approach of first examining a model in which our approach is not applicable³. We do so to highlight several features of our procedure and to illustrate its relationship with existing estimators. We then examine alternative models for which our procedure

³ We return to this case in section 5 to outline how this model can be estimated through the use of an additional step.

is attractive. First, however, consider the following model concerned with the estimation of binary treatment effects.

CASE 1. a) $y_i = y_i^*$; b) $z_i = 1(z_i^* > 0)$; c) $n = N$; This model has the form

$$y_i = x_i\beta + z_i\gamma + u_i \quad (2.16)$$

$$z_i^* = w_i\theta + v_i \quad (2.17)$$

$$z_i = 1(z_i^* > 0). \quad (2.18)$$

This model has a dummy endogenous explanator in the structural equation and the reduced form equation has a binary dependent variable. As $n = N$ the estimates of the parameters from both equations are based on the entire sample. Models of this type are considered in Heckman (1978) and Vella (1993) although the treatment in those papers is parametric⁴.

As z_i^* is no longer observed we are unable to employ the residuals from the reduced form equation. Accordingly we project $u_i = f(v_i) + e_i$ onto z_i and w_i . This gives

$$y_i = x_i\beta + z_i\gamma + E(f(v_i)|w_i, z_i) + [f(v_i) - E(f(v_i)|w_i, z_i)] + e_i.$$

The regressors x_i , z_i and $E(f(v_i)|w_i, z_i)$ are orthogonal to the new error term $f(v_i) - E(f(v_i)|w_i, z_i) + e_i$. If the following identity holds for some function $r(\cdot)$

$$E[f(v_i)|w_i, z_i] = r[E(v_i|z_i, w_i)] \quad (2.19)$$

⁴ Vella (1993) also considers where the marginal distribution of v_i is normal and the distribution of u_i is unknown. Using those assumptions Vella approximates the conditional distribution of u_i by taking some suitably chosen expansion around v_i . We discuss the relative merits of this procedure below.

we can eliminate the unobserved heterogeneity $r(\tilde{v}_i)$, where \tilde{v}_i denotes $E(v_i|z_i, w_i)$, through our proposed methodology. The conditional expectation of the error in this model is the generalized error given by

$$E(v_i|z_i, w_i) = I(z_i = 1) \frac{\int_{-w_i\theta}^{\infty} t F'(t) dt}{1 - F(-w_i\theta)} + I(z_i = 0) \frac{\int_{-\infty}^{-w_i\theta} t F'(t) dt}{F(-w_i\theta)} \quad (2.20)$$

where $F(\cdot)$ is the cumulative distribution function of error term v_i and $F'(t) = dF(t)/dt$. We can then compute (2.20) using \sqrt{N} -consistent estimates of θ . If we assume v_i is normally distributed with constant variance we can estimate θ by probit maximum likelihood and (2.20) reduces (see, *Gourieroux et al.* 1987, *Pagan and Vella* 1989) to

$$E(v_i|z_i, w_i) = \frac{F'(w_i\theta) [z_i - F(w_i\theta)]}{[1 - F(w_i\theta)] F(w_i\theta)}.$$

After estimating the expectations conditional on the generalized residuals we estimate the transformed form of equation (2.16) by least squares. The form of the generalized residual shown in equation (2.20) is independent of the parametric assumptions regarding v_i . Thus the normality assumption can be replaced by alternative parametric assumptions.

An important limitation of our procedure for estimating Case 1 and the following Case 2 is captured in Theorem 2.2. For our procedure to be valid we require equation (2.19) to be satisfied. When $f(\cdot)$ is linear it is straightforward to verify that (2.19) holds. However, when we estimate (2.16) over all N observations the only permissible form of $f(\cdot)$ is a linear function. This is stated in the following theorem.

Theorem 2.2. If there is a differentiable function $r(\cdot)$ such that equation (2.19) holds then the function $f(\cdot)$ must be linear when the primary equation is estimated over all N observations and the expectations are taken with respect to the generalized residuals.

Proof: Assume A_j denotes the interval $(\mu_{j-1}, \mu_j]$. Taking mathematical expectation of u , conditional on regressors x_i and D_{ji} , gives

$$E[u_i | x_i, D_i] = E[f(v_i) | x_i, D_i] = \sum_j D_{ji} \frac{\int_{\mu_{j-1}-w_i\theta}^{\mu_j-w_i\theta} f(v) \phi_v(v) dv}{\Phi_v(\mu_j - w_i\theta) - \Phi_v(\mu_{j-1} - w_i\theta)}.$$

If there exists a function r , such that $E[f(v_i) | w_i, D_i] = r[E(v_i | w_i, D_i)]$, then we have for all j ,

$$\frac{\int_{\mu_{j-1}-w_i\theta}^{\mu_j-w_i\theta} f(v) \phi_v(v) dv}{\Phi_v(\mu_j - w_i\theta) - \Phi_v(\mu_{j-1} - w_i\theta)} = r \left[\frac{\int_{\mu_{j-1}-w_i\theta}^{\mu_j-w_i\theta} v \phi_v(v) dv}{\Phi_v(\mu_j - w_i\theta) - \Phi_v(\mu_{j-1} - w_i\theta)} \right].$$

Let $\Phi_v(\mu_j - w_i\theta) - \Phi_v(\mu_{j-1} - w_i\theta) = 1/c$, where c is a constant. If $\phi_v(v)$ is unimodal, we can solve uniquely $w_i\theta = m(c)$. Thus the above equation can be written as

$$c \int_{\mu_{j-1}-m(c)}^{\mu_j-m(c)} f(v) \phi_v(v) dv = r \left[c \int_{\mu_{j-1}-m(c)}^{\mu_j-m(c)} v \phi_v(v) dv \right].$$

This r function exists but depends on j , *i.e.*

$$E[f(v_i) | x_i, D_i] = \sum_j D_{ji} \left[\frac{\int_{\mu_{j-1}-w_i\theta}^{\mu_j-w_i\theta} v \phi_v(v) dv}{\Phi_v(\mu_j - w_i\theta) - \Phi_v(\mu_{j-1} - w_i\theta)} \right]$$

which cannot be written as $r(\tilde{v})$, where \tilde{v} is the generalized residual.

If we only observe the sub-sample of one group, as the sample selection model,

$$r \left[\frac{\int_{\mu_{j-1}-w_i\theta}^{\mu_j-w_i\theta} v \phi_v(v) dv}{\Phi_v(\mu_j - w_i\theta) - \Phi_v(\mu_{j-1} - w_i\theta)} \right].$$

This last expression provides the proof of Theorem 2.3 that follows.

To avoid distributional assumptions we may wish to employ a \sqrt{N} -consistent semi-parametric estimator of θ (see, for example, Klein and Spady 1993, and Powell, Stock and Stoker 1989). With the estimates of θ we can construct the single index $w_i\hat{\theta}$ and compute, through kernel estimation, the expectation $E[z_i|w_i\hat{\theta}]$ which we employ as an estimate of $F(w_i\hat{\theta})$. We then return to our formulae for the generalized residual and compute these “quasi” generalized residuals. We then compute the conditional expectation with respect to these “quasi” generalized residuals and estimate the primary equation by OLS.

A major issue in the Case 1 setting is the inability to estimate γ . With an endogenous binary treatment in the primary equation the generalized residuals generate an ordering of the data where the observations satisfying $z = 0$ comprise the first part of the re-ordered sample while those for which $z = 1$ comprise the remainder. Thus the terms $z_i - E(z_i|\tilde{v}_i)$ have values of 0 and γ is unidentified. In Section 2.5 we outline a method for the estimation of the binary treatment effects.

CASE 2. a) $y_i = y_i^*$; b) $D_{ji} = 1$ iff $z_i^* \in A_j$ and $D_{ji} = 0$ otherwise; $n = N$.

We now extend the binary treatment model to the ordinal treatment case.

$$y_i = x_i\beta + D_i\gamma + u_i$$

$$z_i^* = w_i\theta + v_i$$

$$z_i = j \text{ iff } z_i^* \in A_j$$

$$D_{ji} = 1 \text{ iff } z_i = j \text{ and } D_{ji} = 0 \text{ otherwise}$$

where A_j are subsets of the real line; D_{ji} denotes that the latent variable z_i^* is

in some specific range; $D_i = (D_{1i}, \dots, D_{Ji})$. The difficulty in estimating β and γ is again due to the endogeneity of D_i . An example of this model is the impact of schooling on wages, where z_i^* denotes a latent continuous variable capturing schooling and y_i denotes wages. D_{ji} indicates the individual i has obtained education level in the range of A_j . Taking conditional expectations of u_i with respect to w_i and D_i gives

$$y_i = x_i\beta + D_i\gamma + E(f(v_i) | w_i, D_i) + \varsigma_i$$

where $\varsigma_i = u_i - E(f(v_i) | w_i, D_i)$. If we assume that v_i is normally distributed the generalized residual is given by (see Vella 1993)

$$E[v_i | w_i, D_i] = \sum_{j=1}^J D_{ji} \frac{\phi(\mu_{j-1} - w_i\theta) - \phi(\mu_j - w_i\theta)}{\Phi(\mu_j - w_i\theta) - \Phi(\mu_{j-1} - w_i\theta)}$$

where ϕ and Φ are the probability density and cumulative distribution functions of the standard normal distribution; and the μ 's represent the estimated separation points. This model does not suffer from the non-identifiability of the treatment effects, γ , when the dependent variable is ordinal with more than 2 outcomes, as the value of the generalized residuals is no longer unique to a certain value of the dependent variable. The limitations imposed by Theorem 2.2, however, regarding the linearity of $f(\cdot)$ are still required. Ideally we would estimate the reduced form equation by some \sqrt{N} -consistent semi-parametric procedure to avoid imposing distributional assumptions. However there are no such procedures currently available. It may be possible, however, to employ alternative parametric assumptions.

CASE 3. a) $y_i = y_i^* \times 1(z_i^* > 0)$; b) $n < N$; This model has the following

form

$$y_i = x_i\beta + u_i$$

$$z_i^* = w_i\theta + v_i.$$

The values of the dependent variable in the primary equation are only observed for the subset for which z_i^* is positive. The model can be estimated by maximum likelihood (see Heckman 1974) under appropriate distributional assumptions. It is also possible to estimate these models, although less efficiently, by various 2-step procedures (see Heckman 1979). The 2-step procedures exploit the distributional assumptions and employ the conditional expectation of the reduced form error as an additional variable in the primary equation. Under joint normality the conditional expectation is the inverse mills ratio⁵. A number of semi-parametric estimators also exist for this model (see, for example, Gallant and Nychka 1987 and Newey 1988). The two semi-parametric estimators closest to our methodology are Powell (1989) and Ahn and Powell (1993). Before considering the relationship of these procedures with our estimator we first outline our approach. First estimate the reduced form by some \sqrt{N} -consistent procedure. We then estimate the primary equation over the transformed values for the sub-sample corresponding to $z_i = 1$. As with Case 1 we are able to employ alternative distributional assumptions for v_i . Alternatively we can employ a semi-parametric procedure to obtain “quasi” generalized residuals. A notable extension from the Cases 1 and 2 to Case

⁵ As discussed by Vella (1993) the inverse mills ratio is the generalized residual for the probit model.

3 model is captured in Theorem 2.3.

Theorem 2.3. In estimating β over a single sub-sample (*i.e.* $D_j = 1$ for some j) we only require that the $f(\cdot)$ function be differentiable for our procedure to be valid.

Proof: See the proof of Theorem 2.2.

Theorem 2.3 relaxes the requirement that the primary equation error term is linearly related to the reduced form error thus extending the estimators of Heckman (1979) and others (see Olsen 1980, for example) which assume this relationship is linear. It also presents an extension to the estimators of Lee (1984) and Vella (1993) who approximate the non-linearity through arbitrarily chosen polynomials.

Our estimator in Case 3 is very closely related to that proposed by Powell (1989) and extended by Ahn and Powell (1993). It is useful to compare the motivation for our estimator with that proposed by Powell (1989). Powell observes that in the sample selection model the error in the primary equation can be written as

$$u_i = k(w_i\theta) + e_i$$

where k is a function mapping the single index $w_i\theta$ into the unknown error generating the selection bias and e_i is uncorrelated with the regressors. Powell proposes defining closeness on the basis of the single index rather than the residual. In the second step Powell employs a more imaginative instrumental variables procedure which employs all possible pairwise deviations assigning decreasing weights to observations far apart in terms of the single index⁶. The essential difference between

⁶ Powell notes that an estimator can be derived based on the difference between actual and

the Powell procedure and our estimator is the use of the single index rather than the residual when defining closeness. Note, however, that the expectations conditional on single index and the residual will be identical when the second step is performed on the sub-sample corresponding to $z_i = 1$. As the sub-sample have the same value for z_i the generalized residual has the form

$$E[v_i|w_i, z_i = 1] = \frac{\int_{-w_i\theta}^{\infty} tF'(t)dt}{1 - F(-w_i\theta)}.$$

The generalized residual is a real function of the single index $w_i\theta$ and Powell's procedure is identical to ours although differences may arise in the second step.

Note our estimator is easily extended to other circumstances in which the Powell procedure is likely to be less effective. This is due to the implicit inclusion of the dependent variable in our conditioning set. For example, consider the following logic. The objective is to eliminate unobserved heterogeneity by defining observations that are close to each other in the unobservables. Consider two individuals which have identical values of w , and thus the single index, but different values of z . This could only occur if they have very different values for v . The Powell approach would consider these observations close while our procedure would, correctly, treat them as far apart. Although we are not critical of the Powell procedure, as he is considering the case where all the z_i 's are 1, we feel our motivation for taking the expectations conditional on the residuals is appropriate for a wider family of models.

expected values, conditional on the single index, of the variables. The theory in his paper, however, focuses on the pairwise comparisons.

Ahn and Powell (1993) estimate the second step in the same manner as proposed by Powell. However the authors relax the single index requirement in the first step by estimating the probability $\Pr[z_i = 1|w_i]$ non-parametrically. They then define closeness through $\Pr[z_i = 1|w_i]$.

CASE 4. a) $y_i = y_i^* \times D_{ji}$; b) $D_{ji} = 1$; c) $n < N$; An extension of the binary selection model is where we return to Case 2 and consider the estimation of β for the sub-samples corresponding to different treatments. However, as is shown in Theorem 2 the inclusion of the treatment effects in estimation imposes linearity between the error terms. When we relax the direct estimation of the treatment effects we can relax the assumption of linearity and invoke Theorem 3. However, to do this we need to examine each sub-sample, corresponding to $D_{ji} = 1$, separately and order the data on the basis of the generalized residuals for each sub-sample in isolation. We then estimate the vector of β for each sub-sample. This procedure is more appealing than that proposed by Vella (1993) which suffers from the arbitrary manner in which the non-linearity is approximated⁷.

Clearly the treatment effects will not be identified when we only have observations on the sub-samples. Through the three stage technique outlined in next section it is possible to recover the treatment effects by using the whole sample.

CASE 5. a) $y_i = y_i^*$; b) $z_i = z_i^*$ if $z_i^* > 0$ and $z_i = 0$ otherwise; c) $n < N$;

$$y_i = x_i\beta + z_i\gamma + u_i$$

$$z_i^* = w_i\theta + v_i$$

⁷ The estimated constant in this model is the sum of the treatment effect and the unconditional expectation of the u_i corresponding to $D_{ji} = 1$, i.e. $E[E(u_i|w_i\theta, D_{ji} = 1)]$.

$$z_i = z_i^* \text{ if } z_i^* > 0, z_i = 0 \text{ otherwise.}$$

Case 5 is a model with a censored endogenous regressor where the regressor is observed when its latent value is positive. Examples of this model are the following. When γ is set equal to zero the latent z_i^* may represent hours worked and the primary equation may be the wage level for those individuals reporting positive hours. This is different to the conventional sample selection model in that we observe the number of hours worked. We can also relax the restriction $\gamma = 0$ and examine how wages are affected by hours worked. Note, however, that for both of these models the second stage estimation is only performed for the observations where $z_i^* > 0$.

To estimate this model we can assume the reduced form error is normally distributed and estimate the reduced form equation by tobit over N observations. We then compute the generalized residuals (see Gourieroux et al. 1987) given by

$$E[v_i|w_i, z_i] = I(z_i = 0) \times \frac{-F'(w_i\hat{\theta})}{1 - F(w_i\hat{\theta})} + I(z_i^* > 0) \times (z_i - w_i\hat{\theta})$$

noting that for the second stage estimation the residuals for the censored observations are, for practical purposes, irrelevant. If we do not wish to make distributional assumptions we can estimate θ semi-parametrically using the procedures of Powell (1984, 1986). We now order the observations according to $z_i - w_i\hat{\theta}$ and then transform the data and estimate by OLS. As we only employ, in the estimation of the primary equation, the observations with the continuously distributed error terms we are able to invoke Theorem 3. Note that while we only consider limited forms of censoring our procedure can be extended to various alternative forms provided

the corresponding reduced form equation can be estimated.

CASE 6. a) $y_i = y_i^*$; b) $z_i = z_i^*$; The final model we examine has the following form:

$$y_i^* = x_i\beta + \gamma(z_i) + u_i$$

$$z_i^* = w_i\theta + v_i$$

where γ now represents a polynomial function of z_i with parameters as its coefficients.

It is possible to estimate this model by instrumental variables methods although it requires more severe exclusion restrictions to ensure identification. Our estimation procedure first estimates the reduced form by OLS to obtain the residuals $z_i^* - w_i\hat{\theta}$. We transform the data and perform OLS to obtain consistent estimates of β and the parameters characterizing γ .

An alternative approach is proposed by Newey, Powell and Vella (1994). In that procedure the f function is approximated by some flexible method. In our approach we eliminate the f function so we by-pass the issues associated with approximating f . The case when the γ function is unknown and our interest is to estimate the parameters β is treated in Ming and Vella (1994a). In that paper we difference out the unknown function $\gamma(z_i)$ and the residual component by conditioning on z_i and \hat{v}_i simultaneously.

2.5 Three Step Synthetic Fixed Effects Estimation

A major limitation of our procedure is its inability to identify the treatment effects in both the binary treatment model and the sub-sample procedure studied in Case 4. This, however, can be overcome through an additional step. First note that we already obtained an \sqrt{N} -consistent estimator $\hat{\beta}_j$ of β_j , where j indexes each sub-sample. Thus, defining the first step residuals as $\hat{u}_{1i}^{(j)} = y_i^{(j)} - x_i^{(j)} \hat{\beta}_j$ we can write

$$\hat{u}_{1i}^{(j)} = D_i^{(j)} \gamma_j + u_{2i}^{(j)}.$$

Assume the index of sub-sample j runs from $N_{j-1} + 1$ to N_j with $N_0 = 0$ and $N_J = N$ and let u_{2i} denote the residuals for $i = N_{j-1} + 1, \dots, N_j$ and $j = 1, \dots, J$. u_{2i} will be asymptotically orthogonal to the exogenous variables w_i . Thus using composite notation we have

$$\hat{u}_{1i} = \left(D_i^{(1)}, \dots, D_i^{(J)} \right) (\gamma_1, \dots, \gamma_J)' + u_{2i}. \quad (2.21)$$

To estimate (2.21) we invoke Theorem 2.4.

Theorem 2.4. Assume we use $T_i = (T_{1i}, T_{2i}, \dots, T_{Ji}) \in w_i$ as instruments for $D_i = (D_i^{(1)}, \dots, D_i^{(J)})$ in equation (2.21). The estimator of γ is \sqrt{N} -consistent, asymptotically normal, and satisfies the asymptotic linearity condition:

$$\begin{aligned} & \sqrt{N} (\hat{\gamma} - \gamma) \\ &= Q_{TD}^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N T_i' u_i - \sum_{j=1}^J P_j E(T_i' x_i | D_i^{(j)} = 1) \sqrt{N} (\hat{\beta}_j - \beta) \right] + o_p(1) \end{aligned}$$

where $P_j = \Pr(D_i^{(j)} = 1) = \Pr(z_i^* \in A_j)$, $A_j = (\mu_{j-1}, \mu_j]^8$, $Q_{TD} = E(T_i' D_i)$.

⁸ The $E(T_i' x_i | D_j = 1)$ will not be the same for all j due to the sample selection. However,

Proof: The data are stacked such that the index of sub-group j is from $N_{j-1}+1$ to N_j with $N_0 = 0$ and $N_J = N$, $J = 1, 2, \dots, J$. Thus

$$\begin{aligned} & \sqrt{N}(\hat{\gamma} - \gamma) \\ &= \left(\frac{1}{N} \sum_{i=1}^N T'_i D_i \right)^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N T'_i u_i - \left(\frac{1}{N} \sum_{j=1}^J \sum_{i=N_{j-1}+1}^{N_j} T'_i x_i \right) \sqrt{N}(\hat{\beta}_j - \beta) \right]. \end{aligned}$$

By the Law of Large Numbers, the following holds

$$\begin{aligned} & \left(\frac{1}{N} \sum_{j=1}^J \sum_{i=N_{j-1}+1}^{N_j} T'_i x_i \right) \sqrt{N}(\hat{\beta}_j - \beta) \\ &= \left(\sum_{j=1}^J \frac{N_j - N_{j-1}}{N} \cdot \frac{1}{N_j - N_{j-1}} \sum_{i=N_{j-1}+1}^{N_j} T'_i x_i \right) \sqrt{N}(\hat{\beta}_j - \beta) \\ &= \sum_{j=1}^J P_j \left[E(T'_i x_i | D_i^{(j)} = 1) + o_p(1) \right] \sqrt{N}(\hat{\beta}_j - \beta) \\ &= \sum_{j=1}^J P_j E(T'_i x_i | D_i^{(j)} = 1) \sqrt{N}(\hat{\beta}_j - \beta) + o_p(1). \end{aligned}$$

So we have

$$\begin{aligned} & \sqrt{N}(\hat{\gamma} - \gamma) \\ &= Q_{TD}^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N T'_i u_i - \sum_{j=1}^J P_j E(T'_i x_i | D_i^{(j)} = 1) \sqrt{N}(\hat{\beta}_j - \beta) + o_p(1) \right] \\ &= Q_{TD}^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^N T'_i u_i - \sum_{j=1}^J P_j E(T'_i x_i | D_i^{(j)} = 1) \sqrt{N}(\hat{\beta}_j - \beta) \right] + o_p(1) \end{aligned}$$

We can see by the Central Limit Theorem, $\frac{1}{\sqrt{N}} \sum_{i=1}^N T'_i u_i$ and $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically normally distributed. Thus their linear combination is also asymptotically normally distributed.

$\sum_{i=1}^N T'_i u_i$ will have similar properties as when we use the whole sample directly to estimate the model.

While the three step approach is not the strongest aspect of our general approach it does have some appeal. First, it generally requires less instruments than direct instrumental variables. For example, if $w = x$, the direct instrumental variables approach is impossible due to the lack of instruments for the treatment effects. Our 3-step instrumental variables procedure is valid, providing the number of exogenous regressors is no less than the number of the treatment dummies. Second, the 3-step procedure allows us to relax the linearity of the f function prior to estimating the treatment effects in the ordinal treatment model. Also note that this three step procedure can be employed to estimate the treatment effects after obtaining the estimates of β via the methodology of Powell (1989) and Ahn and Powell (1994).

2.6 Control Function Procedures versus Synthetic Fixed Effects Procedures

Some of the models outlined above can be estimated by control function procedures pioneered by Heckman (1979), for the binary treatment effects under normality, and subsequently extended to alternative distributional assumptions (see, for example Olsen 1980) and a wider range of censoring rules (see, for example, Vella 1993). It is useful to discuss the relative advantages of our proposed procedure over this methodology.

The two primary features of our approach, also enjoyed by Powell (1989) and Ahn and Powell (1993) for the sample selection model, are related to the lack of distributional assumptions. First, when we are able to avoid the use of distributional assumptions in the reduced form estimation our procedure is robust to

misspecification of the error distribution in this step. Second, as we do not need to specify the nature of the f function, or the manner in which it is approximated, we avoid the possibility of misspecification in this area.

Another advantage is associated with the greatest misgiving with the control function procedures. Models estimated by control function procedures are typically identified through the non-linear mapping from the z_i to \hat{v}_i . While this is a valid form of identification it has received a great deal of criticism (see, for example, Little 1985). The major objection is the collinearity between all the regressors x_{ki} and \hat{v}_i , which contaminates the estimates of β . While this is untrue when the second step estimation is over the entire sample (*i.e.* $n = N$), since $\sum_{i=1}^N x_{ki} \hat{v}_i = 0$ by definition⁹, it is potentially a major advantage when the second step estimation is over a subset of observation in which any one, or more, of the x_{ki} are highly correlated with the \hat{v}_i . Moreover this problem occurs whenever we examine any sub-sample irrespective of whether the original treatment variable is dichotomous or polychotomous. However, while this collinearity is a major issue in the control function procedure when estimating sample selection models the Monte Carlo evidence in Ming and Vella (1994b) suggests the synthetic fixed effects estimator is less sensitive. This is due to the additional non-linearity induced through the conditional expectations operation.

⁹ This condition is precisely the first order conditions defining $\hat{\theta}$.

2.7 Covariance Matrix Estimation

In order to conduct inference we require a consistent estimate of the covariance matrix given in Theorem 1. As there is no general covariance matrix for all the models which can be estimated by our procedure we outline a strategy for computing model specific covariance matrices. From Theorem 1

$$V_\delta = \sigma_e^2 Q_s^{-1} + Q_s^{-1} Q_{sg} V_\theta Q_{sg}' Q_s^{-1},$$

where σ_e^2 and Q_s are obtained from OLS estimation and V_θ is derived from the reduced form. The only term we require is Q_{sg} .

By definition, $g(\hat{\theta}, z_i^* - w_i \hat{\theta}) = E[f(v_i) | z_i^* - w_i \hat{\theta}]$. Thus we require an estimate of $f(v)$. This can be estimated non-parametrically using the residuals \hat{u}_i and \hat{v}_i by

$$\hat{f}(v) = \sum_{i=1}^N K\left(\frac{\hat{v}_i - v}{h}\right) \hat{u}_i / \sum_{i=1}^N K\left(\frac{\hat{v}_i - v}{h}\right)$$

where $\hat{u}_i = y_i^* - x_i \hat{\beta} - z_i \hat{\gamma}$; $\hat{v}_i = z_i^* - w_i \hat{\theta}$; $\hat{\beta}$ and $\hat{\gamma}$ are the final stage estimators¹⁰. The terms $g(\hat{\theta}, z_i^* - w_i \hat{\theta}) = E[f(v_i) | z_i^* - w_i \hat{\theta}]$ and $G_i(\theta) = dg(\hat{\theta}, z_i^* - w_i \hat{\theta})/d\hat{\theta}$ can also be estimated consistently by kernel estimation technique. The formulae are

$$\begin{aligned} \hat{g}(\hat{\theta}, z_i^* - w_i \hat{\theta}) &= \hat{E}[f(v_i) | \hat{v}_i] \\ &= \sum_j K\left(\frac{\hat{v}_j - \hat{v}_i}{h}\right) \hat{f}(\hat{v}_j) / \sum_j K\left(\frac{\hat{v}_j - \hat{v}_i}{h}\right) \\ \hat{G}_i(\hat{\theta}) &= d\hat{g}(\hat{\theta}, z_i^* - w_i \hat{\theta})/d\hat{\theta} \end{aligned}$$

¹⁰ Recall that we assume y_i^* and z_i^* are observable in Theorem 1. Thus if the reduced form dependent variable is censored we employ the generalized residuals.

$$= \frac{d}{d\hat{\theta}} \left[\sum_j K \left(\frac{\hat{v}_j - \hat{v}_i}{h} \right) \hat{f}(\hat{v}_j) \right] / \sum_i K \left(\frac{\hat{v}_j - \hat{v}_i}{h} \right) \Bigg].$$

The derivative of $\hat{\theta}$ in the second equation is with respect to the $\hat{\theta}$ in $\hat{v}_i = z_i^* - w_i \hat{\theta}$ and $\hat{v}_j = z_j^* - w_j \hat{\theta}$ within the kernel $K(\cdot)$, not the $\hat{\theta}$ in \hat{v}_j within the $\hat{f}(\cdot)$. When these estimations are appropriate, we conjecture that $Q_{sg} = E[s_i' G_i(\theta)]$ can be estimated consistently by

$$\hat{Q}_{sg} = \frac{1}{N} \sum_{i=1}^N \hat{s}_i' \hat{G}_i(\hat{\theta}).$$

We do not, however, provide a proof of the consistency of this estimator.

2.8 Simulation Evidence

We now investigate the performance of our procedure in a controlled and simple setting by examining Cases 4 and 6. First we consider the following model:

$$y_i = x_{1i} + d_{1i} + d_{2i} + u_i; \quad i = 1, \dots, n$$

$$z_i^* = x_{1i} + x_{2i} + v_i; \quad i = 1, \dots, N$$

$$z_i = 0 \text{ if } z_i^* \leq \mu_1; z_i = 1 \text{ if } \mu_1 < z_i^* \leq \mu_2; z_i = 2 \text{ if } z_i^* > \mu_2$$

$$d_{1i} = 1 \text{ iff } z_i = 1; d_{2i} = 1 \text{ iff } z_i = 2.$$

To simulate this model we generate the data in the following manner. x_1 and x_2 are independently and uniformly distributed over the interval $[-2.5, 2.5]$; $\mu_1 = -0.5$ and $\mu_2 = 0.5$; $v_i \sim N(0, 1)$; $e_i \sim N(0, 1.5)$; $\sigma_{ve} = 0$ and $u_i = v_i^2 - 1 + v_i^3 + e_i$. We choose $N = 1000$ and the value of n varies by the replication depending on the draws of the errors. We employ the density function of the standard normal

	SFE	OLS	CF
MEAN	1.001	0.421	1.007
MSE	0.138	0.612	0.167

Table 2.1: Simulation Results for Ordinal Sample Selection Model

distribution as our kernel and set the bandwidth equal to 0.1. We generated 1000 replications.

To examine the performance of the procedure we estimated the coefficient on x_1 in the primary equation for the sub-sample $d_1 = 1$. We computed the mean value of the estimate from the replications and the simulated mean squared error computed as $\sum_{r=1}^{1000}(\hat{\beta}_r - 1)^2/1000$. To provide comparisons we compute the corresponding values for the OLS and control function estimates. These results are reported in Table 2.1. Note that as the function f is not linear, only the synthetic fixed effects procedure produces consistent estimates.

Table 2.1 strongly suggests that the synthetic fixed effects procedure works well for this model and its performance is superior to the control function estimator. Despite the surprisingly good performance of the control function procedure the synthetic fixed effects procedure has a smaller bias and a clearly smaller simulated mean squared error. Finally, the values for OLS indicate that the bias is substantial

The second model we consider is an example of Case 6:

$$y = x_{1i} + z_i + z_i^2 + z_i^3 + u_i; \quad i = 1, \dots, N$$

$$z_i = x_{1i} + x_{2i} + v_i; \quad i = 1, \dots, N$$

	x_1	z	z^2	z^3
MEAN				
SFE	1.00360	0.99525	1.00045	1.00001
OLS	0.20074	1.80310	1.26560	1.00025
2SLS	0.95806	0.96620	1.27652	1.00179
MSE				
SFE	0.13792	0.11358	0.02065	0.00153
OLS	0.81441	0.81025	0.26588	0.00281
2SLS	0.20737	1.27630	0.27741	0.03703

Table 2.2: Simulation Results for Model with Nonlinear Endogenous Regressors

and the error terms have the following form; $v_i \sim N(0, 3)$, $e_i \sim N(0, 3)$ and $\sigma_{ve} = 0$; $u_i = v_i + 0.5 \times (v_i^2 - 1) + e_i$. We choose $h = 0.05$. The simulation results are presented in Table 2.2. We report the OLS estimates to illustrate the strength of the bias. We also report the 2SLS estimates which employ x_{1i} , x_{2i} , x_{1i}^2 , x_{2i}^2 , $x_{1i}x_{2i}$, $x_{1i}^2x_{2i}$, $x_{1i}x_{2i}^2$ and $x_{1i}^2x_{2i}^2$ as instruments.

Table 2.2 provides strong evidence of the good performance of our estimator. The OLS estimates suggests that with the exception of the coefficient for z_i^3 the estimates are badly biased. While the 2SLS estimates, which are also consistent, greatly reduce this bias the synthetic fixed effects estimates are superior in terms of bias reduction. Furthermore, the simulated mean squared errors also support the better performance of the synthetic fixed effects estimator.

2.9 Relationship between synthetic fixed effects estimator and the 2SLS estimator

Assuming all the data is observed and $\hat{\theta}$ is the OLS estimator of θ we have

$$z_i = w_i \hat{\theta} + \hat{v}_i.$$

The 2SLS estimator is the OLS estimator of the following regression

$$y_i^* = x_i \beta + z_i \gamma + \hat{v}_i \delta + e_i. \quad (2.22)$$

For explicitness, we copy equation (11) here

$$y_i^* - E[y_i^* | \hat{v}_i] = [x_i - E(x_i | \hat{v}_i)] \beta + [z_i - E(z_i | \hat{v}_i)] \gamma + \varepsilon_i \quad (2.23)$$

where $\varepsilon_i = f(v_i) - E[f(v_i) | \hat{v}_i] + e_i$. If we generate $E[y_i^* | \hat{v}_i]$ and $E(z_i | \hat{v}_i)$ via an OLS linear projection, the OLS estimates of β from (2.22) and (2.23) will be identical.

Thus when the following two conditions hold the 2SLS and synthetic fixed effects estimates are equivalent:

- (i) $\hat{\theta}$ is the OLS estimator of θ ;
- (ii) $E[y_i^* | \hat{v}_i]$ and $E(z_i | \hat{v}_i)$ are estimated with OLS by regressing y_i^* and z_i on \hat{v}_i .

2.10 Conclusions of Chapter 2

This Chapter develops a new procedure for eliminating the inconsistency resulting from endogenous regressors in cross sectional models. We do so by obtaining an

estimate of the unobserved heterogeneity responsible for the endogeneity and the performing appropriate data transformations to eliminate the endogeneity. Given this approach is often employed in panel data our procedure is closely related to several panel data estimation methods.

While our approach is applicable to the conventional simultaneous equation model it is perhaps most useful in dealing with models with sample selection bias and censored endogenous regressors. This is due to its ability to relax important distributional assumptions required by most alternative procedures available for these models. Our approach is also particularly useful when the endogenous independent variable appears in the conditional mean of the primary equation in a non-linear manner.

CHAPTER 3

Semi- and Non-Parametric Tests of Independence

Semi- and nonparametric tests of independence of two random variables are considered. First, the exact finite-sample distributions of Blum *et al.*'s (1961) discretely distributed statistics, based on directly observable variables, are calculated for small sample sizes, $n = 1, 2, \dots, 8$. The calculation becomes prohibitively expensive for $n > 10$. Second, Monte Carlo simulations are used to approximate the quantiles for $9 \leq n \leq 200$. On the basis of these simulations, we find that even for $n = 200$, the quantiles of the distribution are still significantly different from Blum *et al.*'s (1961) asymptotic quantiles. Third, we obtain the asymptotic distribution of the statistics which is based on variables such as residuals or predictions from a regression model that are not directly observable. This latter test can be useful in specification testing for a class of models. Fourth, since the residual based asymptotics depends on regression function, distribution parameters of the regressors and error term, and the estimator, we prove that the bootstrapped statistics has the same asymptotic distribution as that of the residual and /or predicted value based statistics. Fifth, we extend our residual-based independence test to test serial independence of regression error terms. Monte Carlo evidence shows that bootstrap works well.

3.1 Introduction

Test of independence finds its new applications in econometric literature in recent years. In linear and nonlinear regression and simultaneous equation models, the distinction between independence and conditional moment restrictions has an impact on semiparametric efficiency bounds (see Begun *et al.* 1982, Chamberlain 1987 and Newey 1990, 1989). Also in the literature of residual-based estimation and prediction (see Brown 1993, 1992 and 1990, Newey 1992a and 1992b and Robinson 1991a), the independence of the error term and regressors is crucial for obtaining best estimators and predictors. In time series literature, an appropriate way of testing the random walk hypothesis is to perform a test of serial independence of its first order difference sequence. At the end of section 3, we list out some concrete examples where test of independence is important.

Hoeffding (1948) first published his paper on the test of independence of random variables. Blum *et al.* (1961) modified the statistics of Hoeffding and obtained its asymptotic distribution. The tests in both papers are nonparametric in that they used the empirical distribution functions of random variables to construct the test.

Recently some authors began to reinvestigate other sorts of independence test, including Brock and Dechert (1989), Robinson (1991b), Cameron and Trivedi (1993) Brock and Dechert (1988) and Brock *et al.* (1995). Our paper can be viewed as an extension of Blum *et al.*'s (1961) result to the cases where one or

both of the tested random variables are not observable, *e.g.* residuals or predicted values. This paper makes the following contributions:

Section 3.2 calculates the exact distributions of Blum *et al.*'s (1961) statistics for very small samples, $n = 1, 2, \dots, 8$ and simulates the its distribution quantiles for $9 \leq n \leq 200$. In theory, if computer has enough memory and there is enough time, our computer program can compute the exact distribution of the statistics for any size of the sample. Because the statistics is discretely valued and it takes $n!$ potential differently values, the exact distribution is beyond consideration for $n > 10$. Thus we use simulation technique to obtain the quantiles for $9 \leq n \leq 200$. Our finding is that even for $n = 200$, the quantiles we get are still significantly different from Blum *et al.*'s (1961) asymptotic quantiles. Thus, when sample size is small, for inference purposes our table will be more accurate than Blum *et al.*'s (1961) table which consists of the asymptotic quantiles of the statistics.

Section 3.3 extends the independence test to the case where the samples are not directly observed, *e.g.* residuals and/or predicted values from a regression model. We find that the asymptotic distribution of the statistics constructed from residuals and regressors depends on the parameters of the regression model: the regression function, the distribution of the error term and the regressors. Because the distribution parameters are not known to us, the test is not feasible. To make it feasible, we prove that the bootstrapped statistics has the same asymptotic distribution as that of the original statistics. This independence test can be used as misspecification test for some class of models.

Section 3.4 extends the approach to test the serial independence of regression error terms using residuals. The asymptotic distribution of the statistics constructed from residuals depends also on the parameters of the regression model: the regression function, the distribution of the error term and the regressors.

Section 3.5 conducts some Monte Carlo experiments to show that our bootstrapped procedure works.

This paper is organized in the above order.

3.2 Exact Finite Sample Distributions of Blum *et al.*'s statistics

3.2.1 Definition of the statistics The sufficient and necessary condition for random variables X and Y to be independent of each other is $F_{xy}(c_1, c_2) = F_x(c_1)F_y(c_2)$ for all $(c_1, c_2) \in R^2$, where $F_{xy}(\cdot, \cdot)$ is the joint distribution function of random variables X and Y , and $F_x(\cdot)$, $F_y(\cdot)$ the marginal distribution functions of X and Y respectively. Hoeffding (1948, Theorem 3.1) proved the following theorem.

Theorem 3.1. If random variables X and Y are absolutely continuous, i.e. their density functions are continuous, then the sufficient and necessary condition for them to be independent is

$$A \equiv \iint [F_{xy}(x, y) - F_x(x)F_y(y)]^2 dF_{xy}(x, y) = 0$$

Hoeffding (1948) and Blum *et al.* (1961) develop tests of independence based on empirical distribution functions of X and Y . Assume X and Y are absolutely continuous and $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are independent random draws from

the population distribution $F_{xy}(x, y)$. The statistics is defined as follows.

Definition 3.1.

$$T_n = n \int [B_n(t_1, t_2)]^2 dG_{xy}(t_1, t_2) \quad (3.1)$$

where

$$\begin{aligned} B_n(t_1, t_2) &= G_{xy}(t_1, t_2) - G_x(t_1)G_y(t_2) \\ G_x(t) &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq t) \\ G_y(t) &= \frac{1}{n} \sum_{i=1}^n 1(y_i \leq t) \\ G_{xy}(t_1, t_2) &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq t_1, y_i \leq t_2) \end{aligned}$$

T_n is clearly distribution-free for absolutely continuous random variables. Thus we will use uniform distribution on $[0, 1] \times [0, 1]$ to define T_n hereafter. Blum *et al.* (1961) prove the following theorem.

Theorem 3.2. If random variables X and Y are absolutely continuous, i.e. their density functions are continuous, then T_n defined in equation (3.1) has the following limiting distribution,

$$T = \int_0^1 \int_0^1 B^2(s, t) ds dt, \quad (3.2)$$

where $B(s, t)$ is a separable Gaussian process with time parameter (s, t) in $[0, 1] \times [0, 1]$, explicitly $B(s, t) = (B_x(s) - sB_x(1))(B_y(t) - tB_y(1))$ with $B_x(s)$ and $B_y(t)$ being independent standard Brownian motions.

Random variable T has tabled p -values by Blum *et al.* (1961).

3.2.2 The Computation of the Exact Finite Sample Distribution of T_n In

this subsection we are going to find the exact finite sample distribution of T_n for

$n = 2, 3, 4, \dots$. In fact, it is almost impossible to list out all the values T_n takes even for n being very small, *e.g.* $n = 10$. Because T_{10} has $10!$ potential different values. Besides, when n is big, T_n becomes very much “continuous” because of the huge numbers of different values it takes. Thus the following paragraph is only for theoretical completion purposes. In addition, based on the following logic, we have compiled a Gauss program¹ to compute the exact finite sample distribution of T_n .

If we sort sample $(x_1, y_1), \dots, (x_n, y_n)$ in the ascending order of (x_i) (we call this sorted sample X -ordered sample) and rewrite T_n as follows,

$$\begin{aligned}
 & T_n \\
 &= \sum_{j=1}^n [G_{xy}(x_j, y_j) - G_x(x_j) G_y(y_j)]^2 \\
 &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, y_i \leq y_j) - \left(\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(y_i \leq y_j) \right) \right]^2 \\
 &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^j 1(y_i \leq y_j) - \frac{j}{n} \left(\frac{1}{n} \sum_{i=1}^n 1(y_i \leq y_j) \right) \right]^2 \\
 &= \sum_{j=1}^n \frac{j^2}{n^2} \left[\frac{1}{j} \sum_{i=1}^j 1(y_i \leq y_j) - \frac{1}{n} \sum_{i=1}^n 1(y_i \leq y_j) \right]^2 \tag{3.3}
 \end{aligned}$$

Under the null, *i.e.* random variables X and Y are uniformly independently distributed on $[0, 1]$, the X -ordered sample of Y , (y_i) , is still a random sample from the uniform distribution on $[0, 1]$. To simulate statistics T_n , expression (3.3) is very advantageous in that we need only to generate random numbers for Y instead of that for X and Y .

¹ This paper is very computer intensive. We write several programs in Gauss to conduct computations, simulations and bootstrapping. All programs are available on request.

For explicitness, we will compute T_2 and T_3 manually.

$$\begin{aligned} T_2 &= \frac{1}{4} \left[1 - \frac{1}{2} (1(y_1 \leq y_2) + 1) \right]^2 \\ &= \frac{1}{16} [1(y_1 > y_2)]^2 = \frac{1}{16} 1(y_1 > y_2) \end{aligned}$$

i.e. T_2 is a binomial distribution: $\Pr(T_2 = 1/16) = 1/2$, $\Pr(T_2 = 0) = 1/2$.

$$\begin{aligned} T_3 &= \frac{1}{9} \left[1 - \frac{1}{3} (1 + 1(y_2 \leq y_1) + 1(y_3 \leq y_1)) \right]^2 \\ &\quad + \frac{4}{9} \left[\frac{1}{2} (1(y_1 \leq y_2) + 1) - \frac{1}{3} (1(y_1 \leq y_2) + 1 + 1(y_3 \leq y_2)) \right]^2 \\ &= \frac{1}{81} \{2 + 6 \cdot 1(21) - 1(13) - 2 \cdot 1(21, 13) - 4 \cdot 1(23, 21)\} \end{aligned}$$

where (ij) denotes the event $(y_i \geq y_j)$ and $1(ij)$ the indicator function $1(y_i \geq y_j)$.

The potential $2^3 = 8$ different events are denoted by: $A_1 = (12, 13, 23) (1/6)$; $A_2 = (12, 13, 32) (1/6)$; $A_3 = (12, 31, 23) (0)$; $A_4 = (12, 31, 32) (1/6)$; $A_5 = (21, 13, 23) (1/6)$; $A_6 = (21, 13, 32) (0)$; $A_7 = (21, 31, 23) (1/6)$; $A_8 = (21, 31, 32) (1/6)$, where the number in the second parenthesis is the probability the corresponding event occurs.

If A_1 occurs, $T_3 = 1/81$; If A_2 , $T_3 = 1/81$; If A_4 , $T_3 = 2/81$; If A_5 , $T_3 = 2/81$; If A_7 , $T_3 = 4/81$; If A_8 , $T_3 = 8/81$. Thus the distribution of T_3 is as follows:
 $\Pr(T_3 = 1/81) = \Pr(A_1 \cup A_2 \cup A_5) = 1/2$. $\Pr(T_3 = 2/81) = \Pr(A_4) = 1/6$.
 $\Pr(T_3 = 4/81) = \Pr(A_7) = 1/6$. $\Pr(T_3 = 8/81) = \Pr(A_8) = 1/6$.

Manual computation of T_n becomes prohibitively expensive for $n \geq 6$. We use a computer program written in Gauss to compute the right hand tail quantiles of it. Unfortunately, even with a computer program, the computation can not go any further than $n = 8$ with a PC. The reason is that the number of potential

distinct values T_n takes is $n!$. When $n = 9$, 8-Mb ram memory in a PC is not enough. When $n = 10$, the memory it takes is 10 times as much as it takes as $n = 9$. So when n is large we use simulation technique to get an approximation of the right-tail quantiles. For the sake of theory completion, we formalize the above finite sample computation method of T_n and present it in Theorem 3.3.

Intuitively, any two samples with the same ranking order will give us the same value of T_n . Thus any sample (y_1, y_2, \dots, y_n) satisfying $y_1 \geq y_2 \geq \dots \geq y_n$ will give us a unique value of T_n , say, a . And we have $\Pr(T_n = a) = \Pr(y_1 \geq y_2 \geq \dots \geq y_n) = 1/n!$. If there are other samples with different ranking giving T_n the same value of a , then $\Pr(T_n = a) = (\# \text{ of ranking such that } T_n = a)/n!$.

We take it for granted that $\Pr(\text{each specific ranking}) = 1/n!$. If this piece information is unknown to us, two questions should be addressed: (i) How many different rankings have positive measure? (ii) What is the measure of each specific ranking? These two questions are answered by the Theorem 3.3 and Lemma 3.1. For any sample (y_1, y_2, \dots, y_n) , there are 2^k different ways to define the relations between y_i and y_j , for $i \neq j$ and $i, j = 1, \dots, n$, where $k = n \times (n - 1) / 2$. T_n maps each relation to a well-defined value. The measures of (y_1, y_2, \dots, y_n) with some relations are 0, some are not. The relations of (y_i) that have non-zero measure are called *consistent* relations. The sets defined by *consistent* relations are called *consistent* sets. The strict definition is as follows:

Definition 3.2. Assume we have n independent observations (y_1, y_2, \dots, y_n)

n	$\pi^4 T_n/2$				
2	3.0440(0.5)	0.0000(0.5)			
3	4.8103(.1667)	2.4052(.1667)	1.2026(.1667)	.60129(0.5)	
4	6.4686(.0417)	4.7563(.0417)	3.2343(.0833)	2.4733(.0833)	1.9025(.0417)
5	8.1044(.0083)	6.8576(.0083)	5.4549(.0167)	4.9873(.0083)	4.7536(.0167)
6	9.7334(.0014)	8.7939(.0014)	7.5913(.0028)	6.9900(.0056)	6.6518(.0014)

Table 3.1: The Right Hand Tail Exact Distribution of $\pi^4 T_n/2$ for Small Samples

from the uniform distribution on $[0, 1]$, a set $\Upsilon = \cap_{i>j}^n (y_i \circ y_j)$ is called consistent if $\Pr(\Upsilon) > 0$, where the relation sign is either \geq or $<$.

The reason we call Υ consistent² is that the inequalities in the definition of Υ can not contradict with each other if $\Pr(\Upsilon) > 0$, *i.e.* if $y_i \geq y_j$, we can not deduce from other inequalities that $y_i \leq y_j$, for any i and j .

Theorem 3.3. For fixed n , there are $n!$ different consistent sets each with a measure of $1/n!$.

From the proof of Theorem 3.3 we can see that all the consistent sets are symmetric for fixed n , *i.e.* we can change one into another by changing their indexes. In a n -dimensional unit “cubic”, which has a unit measure, there are $n!$ consistent sets. Thus the measure of each consistent set is $1/n!$. We specify this fact by the following Lemma, which can also be proved by multi-fold integral.

² The terminology “consistent” here is not accurate. Because $y_i \leq y_j$ and $y_i \geq y_j$ are not “inconsistent” due to the existence of equality sign. But even there is one *pure* equality sign in a set, *e.g.* $(y_1 = y_2, y_1 \geq y_3, \dots, y_1 \geq y_n; y_2 \geq y_3, \dots, y_2 \geq y_n; \dots; y_{n-1} \geq y_n)$, the measure of the set is zero.

n	0.900	0.950	0.990	0.999
5	3.5067			
6	3.4950	4.2466		
7	3.2659	4.2193	6.3898	9.1080
8	3.1629	4.1142	6.2189	9.2153

Table 3.2: The Exact Distribution Quantiles of $\pi^4 T_n/2$ for Small Samples

Lemma 3.1. In n -dimensional unit “cubic”, each consistent set has a measure of $1/n!$.

Right hand tail distributions for small samples, $n = 2, 3, 4, 5$ and 6 , are listed in Table 3.1. The value in the parenthesis is its probability value. For $n = 5$ and 6 , Table 3.1 does not exhaust all of their 10% right tail values, which are the values to make inference on; Thus they are still listed in Table 3.2, a quantile table, listing the quantiles corresponding to $p = 0.9, 0.95, 0.99$ and 0.999 , *i.e.* the solutions to $p = \Pr(\pi^4 T_n/2 \leq q)$.³

3.2.3 Simulations of the Quantiles of T_n for Finite Sample Size As afore-

mentioned, direct calculation of the distribution of T_n is prohibitively expensive

³ The only reason for us to scale T_n to $\pi^4 T_n/2$ is to make it comparable to the tables compiled by Blum *et. al* (1961).

Because of the discreteness, $\Pr(\pi^4 T_5/2 > 3.5067) = \Pr(\pi^4 T_5/2 \geq 3.7405) = 0.0833$. $\Pr(\pi^4 T_5/2 \geq 3.5067) = \Pr(\pi^4 T_5/2 > 3.1950) = 0.1167$, *i.e.* $\Pr(\pi^4 T_5/2 = 3.5067) = 0.0333$.

For $n = 6$ there exists the same discreteness problem as for $n = 5$. $\Pr(\pi^4 T_6/2 > 4.2466) = \Pr(\pi^4 T_6/2 \geq 4.2842) = 0.05$. $\Pr(\pi^4 T_6/2 > 3.4574) = \Pr(\pi^4 T_6/2 \geq 3.4950) = 0.1$. Thus the quantiles are not accurate. As a matter of fact, quantiles do not exist for $n \leq 4$, and some of the quantiles do not exist for $n = 5$ and 6 . That is why we compiled the distribution Table 1 and an incomplete quantile Table 2. When $n = 7$ the quantiles can be accurate to the second decimal. When $n = 8$ the quantiles can be accurate to the third decimal. When $n \geq 9$ the distribution

function of T_n becomes “almost” continuous that the discreteness problem disappears.

for large n . Thus we use simulation technique to obtain their right tail quantiles for $9 \leq n \leq 200$. The exact right tail distribution and the calculated quantiles of T_n for $5 \leq n \leq 8$ are listed in Tables 3.1 and 3.2 respectively. Simulated quantiles are listed in Table 3.3. For examining the accuracy of our Monte Carlo simulations, the simulated quantiles for $5 \leq n \leq 8$ are also listed in Table 3.3.

We conduct our simulation as follows: For fixed n , we generate random numbers uniformly distributed on $[0,1]$ and calculate T_n using the formula given by equation (3.3). We repeat this process for S times and obtain S different values of T_n , denoted by T_n^s . In our application, $S = 1,000,000$. Then we take the right tail sample quantiles of $(T_n^s)_{s=1}^S$ as our quantile estimates. When $n = 5, 6, 7$ and 8 , the calculated quantiles in Table 3.2 and the simulated quantiles in Table 3.3 are exactly the same. This fact gives us more confidence in our simulation procedure. Another observation about Table 3.3 is the smoothness of our simulated quantiles. They are monotonic decreasing in sample size n . This fact also strengthens our confidence in our simulation results.

3.2.4 Variances of the Simulated Quantiles Although the simulated quantiles in Table 3.3 seems very accurate, they are still randoms. We will next examine the variances of these random quantiles.

Let $0 < p < 1$ denote a given probability, π_n the exact distribution function of T_n and π_{nS} the empirical distribution function of the simulated samples $(T_n^s)_{s=1}^S$. Brown and Mariano (1991) has the following asymptotic result

$$\pi_{nS}^{-1}(p) - \pi_n^{-1}(p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{Sf_n^2(\pi_n^{-1}(p))}\right), \quad (3.4)$$

n, p	0.9	0.95	0.99	0.999	n, p	0.9	0.95	0.99	0.999
5	3.5067				50	2.4674	3.0997	4.6624	7.0553
6	3.4950	4.2466			52	2.4632	3.0929	4.6483	7.0256
7	3.2862	4.2193	6.3898	9.1080	55	2.4513	3.0720	4.6366	6.9563
8	3.1629	4.1142	6.2189	9.2153	58	2.4444	3.0696	4.6103	6.9046
9	3.1104	3.9938	6.0797	8.9229	61	2.4387	3.0539	4.5928	6.8825
10	3.0489	3.8915	5.9371	8.7766	64	2.4288	3.0472	4.5729	6.9063
11	2.9906	3.8256	5.8149	8.6026	67	2.4223	3.0324	4.5524	6.9166
12	2.9407	3.7557	5.7123	8.5073	70	2.4168	3.0249	4.5419	6.8472
13	2.9007	3.6971	5.6291	8.4122	75	2.4102	3.0161	4.5193	6.7894
14	2.8640	3.6463	5.5429	8.2725	80	2.4007	3.0067	4.5163	6.7950
15	2.8304	3.6020	5.4905	8.2295	85	2.3955	2.9950	4.5040	6.7077
16	2.8010	3.5553	5.4110	8.1363	90	2.3892	2.9887	4.4756	6.7150
18	2.7508	3.4918	5.2956	7.9653	95	2.3831	2.9817	4.4640	6.6929
20	2.7083	3.4334	5.2004	7.8198	100	2.3786	2.9751	4.4495	6.7234
22	2.6750	3.3846	5.1257	7.6940	108	2.3704	2.9666	4.4424	6.6775
24	2.6464	3.3429	5.0784	7.6214	116	2.3691	2.9573	4.4212	6.6576
26	2.6174	3.3071	5.0014	7.5556	124	2.3568	2.9486	4.3941	6.6178
28	2.5974	3.2788	4.9588	7.4576	132	2.3561	2.9391	4.4016	6.5734
30	2.5776	3.2528	4.9126	7.4416	140	2.3596	2.9465	4.4097	6.5879
32	2.5596	3.2349	4.8929	7.3107	150	2.3505	2.9350	4.3832	6.5722
34	2.5451	3.2094	4.8599	7.3038	160	2.3459	2.9196	4.3738	6.5921
36	2.5322	3.1951	4.8155	7.2737	170	2.3394	2.9167	4.3599	6.5483
38	2.5208	3.1739	4.7918	7.2395	180	2.3381	2.9182	4.3511	6.4920
40	2.5084	3.1569	4.7648	7.2004	190	2.3372	2.9167	4.3444	6.5121
43	2.4951	3.1337	4.7138	7.1897	200	2.3356	2.9102	4.3432	6.4970
46	2.4845	3.1166	4.6969	7.1231	∞	2.2860	2.8440	4.2300	6.3200
49	2.4737	3.1077	4.6778	7.0542					

Table 3.3: Simulated Quantiles of $\pi^4 T_n/2$ for Sample Size Less than or Equal to 200

where $\pi_n^{-1}(p)$ is the p -th quantile of random variable T_n ; $\pi_{nS}^{-1}(p)$ is the p -th quantile of the simulated random samples $(T_n^s)_{s=1}^S$ of T_n ; $f_n(\cdot)$ is the density function of T_n . We take $S = 1,000,000$. The only unknown in the variance of our simulated estimator is $f_n(\cdot)$. Fortunately Table I in Blum *et al.* (1961) lists the asymptotic distribution function values, from which we can get a rough approximation of $f(\cdot)$ (the asymptotic density) at the four quantile points. The four values are $f_{0.9} = 0.126$, $f_{0.95} = 0.06$, $f_{0.99} = 0.011$ and $f_{0.999} = 0.0012$, where the indexes denote the p 's. These asymptotic values are used as substitutes of the corresponding finite sample density values. Thus we can approximately obtain the standard error of the simulated quantiles. They are $\sigma_{0.9} = 0.0023$, $\sigma_{0.95} = 0.0036$, $\sigma_{0.99} = 0.009$ and $\sigma_{0.999} = 0.02634$. We see from these standard errors that the 0.9-th quantile can be accurate to the third decimal. The 0.999-th quantile can be accurate to the first or the second decimal. The accuracy of the other two quantiles are in-between. By examination of Table 3, we notice that the 0.999-th quantiles of finite sample size do have some fluctuations, however its first decimal is very smooth. There are only 3 places that the second decimal fluctuates a little bit. Thus the accuracy of the quantiles conforms with what their standard errors suggest. Because we use the asymptotic density function at the 0.999-th quantile to be a substitute of the finite sample density function at the 0.999-th quantile. This substitution may deflate the finite sample density function at the 0.999-th quantile. However the asymptotic random variable is infinite and the finite sample statistics are bounded. At the same 0.999-th quantile, the density function of finite sample random variable may

be larger than that of the asymptotic random variable. From equation (3.4), we know that we may have inflated the variance of the 0.999-th quantile estimates. Thus all the 0.999-th quantiles might be smoother than what their approximated standard error suggests. This fact is also reflected in the pattern the 0.999-th quantile estimates: when n becomes bigger, finite sample statistics becomes closer to the asymptotic one, the fluctuation becomes severe.

3.2.5 Ad Hoc Formula for Calculation of Finite Sample Quantiles As we noticed in Table 3.3 that the simulated quantiles have very smooth patterns. They decrease in sample size n . Also the decreasing speed negatively relates to the sample size n . From these observations, we will assume that the quantile is a linear function of sample size n and its negative powers and it takes the following linear form:

$$\pi_n^{-1}(p) = \alpha_1 + \alpha_2 n^{-1} + \alpha_3 n^{-1/3} \dots + \alpha_k n^{-k/2} \quad (3.5)$$

OLS is used to estimate the coefficients with the simulated quantiles in Table 3.3 for $n \geq 9$ as the dependent variables. There are four sets of different coefficients corresponding to the four different probabilities (0.9, 0.95, 0.99, 0.999).

The criteria to choose the regressors in equation (3.5) are as follows:

- (i) R^2 is close to 1;
- (ii) Constant α_1 is close to the asymptotic quantiles in the last row of Table 3;
- (iii) The standard errors of the predicted values is less than the standard errors of the simulated quantiles so that the noise of the calculated quantiles by (3.5) is dominated by the noise of simulation.

By try and error we choose $(1, n^{-1}, n^{-2}, n^{-5/2})$ as our regressors in equation (3.5) and obtain the following results:

$$\begin{aligned}\pi_n^{-1}(0.900) &= \frac{2.288}{(2370)} + \frac{9.361}{(91.12)} n^{-1} - \frac{21.34}{(-5.986)} n^{-2} + \frac{11.40}{(1.356)} n^{-5/2}, R_{0.900}^2 = 1.000 \\ \pi_n^{-1}(0.950) &= \frac{2.842}{(772.3)} + \frac{13.58}{(34.67)} n^{-1} - \frac{46.26}{(-3.404)} n^{-2} + \frac{51.43}{(1.598)} n^{-5/2}, R_{0.950}^2 = 0.999 \\ \pi_n^{-1}(0.990) &= \frac{4.228}{(1057)} + \frac{23.62}{(55.45)} n^{-1} - \frac{112.9}{(-7.640)} n^{-2} + \frac{151.1}{(4.318)} n^{-5/2}, R_{0.990}^2 = 1.000 \\ \pi_n^{-1}(0.999) &= \frac{6.311}{(455.9)} + \frac{40.61}{(27.56)} n^{-1} - \frac{294.6}{(-5.760)} n^{-2} + \frac{423.5}{(3.498)} n^{-5/2}, R_{0.999}^2 = 0.999\end{aligned}$$

where numbers in the parenthesis are the corresponding t -values.

All the R^2 are approximately equal to or close to 1. The four constant terms are very close to the four asymptotic critical values (2.286, 2.844, 4.23, 6.32). For $n = 18$, the standard errors of the four predicted values are (0.00055, 0.0021, 0.0023, 0.0080), which are smaller than the four simulated standard errors (0.0023, 0.0036, 0.009, 0.02634). And also, the standard errors of the predicted values are decreasing functions of sample size. Thus for $n > 18$, the standard errors of the predicted critical values will become even smaller. So our four formulae above satisfy all the three criteria.

How well do our formulae perform? We pick $n = 18, 100$, compute the four quantiles and compare them with the corresponding quantiles in Table 3.3. The calculated quantiles are (2.75036, 3.49043, 5.30113, 7.96672) and (2.37947, 2.97315, 4.45413, 6.69235). They are very close to the simulated values⁴ in Table 3.3.

⁴ Even there are some discrepancies between the calculated quantiles and the simulated quantiles in Table 3, we suggest to use the calculated quantiles. This is due to the fact that the numbers in Table 3 are simulated. There are some randomness in them. Our formulae, however, smooth out the randomness.

3.3 Residual-Based Test of Independence of Regressors and Error Term

This section extends the independence test to the case where one of the two random variables is not observable. Instead it is the residuals and/or predicted values from a regression. In most economic regression models, its error and regressors are independent of each other almost by the model design. If the error term is known, the previously defined statistics can be used directly to test independence. But the error term is not observable. What we have about the error term is the regression residual. In this section we obtain the asymptotic distribution of the statistics using residuals instead of error term observations. This asymptotic distribution depends on not only the regression model, but also the distribution parameters of the regressors and the error term.

3.3.1 Extension of the Independence Test Using Residuals

Consider the fol-

lowing regression model,

$$\rho(Y_i, X_i, \beta_0) = \varepsilon_i \quad (3.6)$$

which has a reduced form solution

$$Y_i = \pi(X_i, \beta_0, \varepsilon_i). \quad (3.7)$$

where $X_i = (x_{1i}, x_{2i}, \dots, x_{ki})$ is a k -dimensional vector regressor; β_0 is a l -dimensional parameter. We assume that $z_i = (Y_i', X_i')'$ are i.i.d. and that the disturbance term ε_i and X_i are independent.

If the regression function $\rho(Y_i, X_i, \beta_0)$ is misspecified as $\gamma(Y_i, X_i, \theta_0)$, the model

becomes

$$\gamma(Y_i, X_i, \theta_0) = [\gamma(Y_i, X_i, \theta_0) - \rho(Y_i, X_i, \beta_0)] + \varepsilon_i.$$

It is easy to see that regressors X_i and the new error term $\kappa_i = [\gamma(Y_i, X_i, \theta_0) - \rho(Y_i, X_i, \beta_0)] + \varepsilon_i$ are no longer independent as long as $\rho(Y_i, X_i, \beta_0) \neq \gamma(Y_i, X_i, \theta_0)$. Thus if some independence test rejects the null hypothesis, we have very good reason to believe that the model is misspecified.

Most chances we do not know β_0 in regression equation (3.6) but a \sqrt{n} -consistent estimates $\hat{\beta}$ of β_0 . Thus the test of the previous section is not applicable here. This section uses residuals e of the regression to construct the statistics and derives its asymptotic distribution.

We always use x_i to denote *one* of the k regressors of model (3.6), (3.7).

Definition 3.3. The statistics is defined as

$$T_n(\beta) = n \iint [B_n(x, e)]^2 dG_{xe}(x, e) \quad (3.8)$$

where

$$\begin{aligned} B_n(x, e) &= G_{xe}(x, e) - G_x(x) G_e(e) \\ G_x(x) &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq x) \\ G_e(e) &= \frac{1}{n} \sum_{i=1}^n 1(e_i \leq e) \\ G_{xe}(x, e) &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq x, e_i \leq e) \end{aligned}$$

and $e_i = \rho_i(\beta) = \rho(z_i, \beta)$.⁵ Under the independence assumption, the limiting

⁵ If there is not confusion, we also use e_i to denote the residual $\rho_i(\hat{\beta})$.

distribution of $T_n(\beta_0)$ is that of T in equation (3.2). We call $T_n(\hat{\beta})$ the residual-based test statistics if $\hat{\beta}$ is an estimator of β_0 .

For any β , we define

$$m_\varepsilon^j(\beta) = m_\varepsilon(z_j, \beta) = E_z[1(\rho(z, \beta) \leq \rho(z_j, \beta))] \quad (3.9)$$

$$m_{x\varepsilon}^j(\beta) = m_{x\varepsilon}(z_j, \beta) = E_z[1(x \leq x_j, \rho(z, \beta) \leq \rho(z_j, \beta))] \quad (3.10)$$

and make the following assumptions.

Assumption 3.1.

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left[m_\varepsilon^{ji}(\beta) - m_\varepsilon^j(\beta) \right] - \left[m_\varepsilon^{ji}(\beta_0) - m_\varepsilon^j(\beta_0) \right] \right\} \\ & \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left[m_{x\varepsilon}^{ji}(\beta) - m_{x\varepsilon}^j(\beta) \right] - \left[m_{x\varepsilon}^{ji}(\beta_0) - m_{x\varepsilon}^j(\beta_0) \right] \right\} \end{aligned}$$

is stochastically equicontinuous at β_0 for given j . Where $m_\varepsilon^{ji}(\beta) = 1(\rho_i(\beta) \leq \rho_j(\beta))$ and $m_{x\varepsilon}^{ji}(\beta) = 1(x_i \leq x_j, \rho_i(\beta) \leq \rho_j(\beta))$.⁶

Assumption 3.2. (a) The partial derivatives $\partial m_\varepsilon^j(\beta) / \partial \beta$ and $\partial m_{x\varepsilon}^j(\beta) / \partial \beta$ exist and are continuous in a neighborhood of β_0 ; (b) $E_{z_j} [\partial m_\varepsilon^j(\beta_0) / \partial \beta' \cdot \partial m_\varepsilon^j(\beta_0) / \partial \beta]$ and $E_{z_j} [\partial m_{x\varepsilon}^j(\beta_0) / \partial \beta' \cdot \partial m_{x\varepsilon}^j(\beta_0) / \partial \beta]$ exist.

Assumption 3.3.

(a) $\hat{\beta}$ is a linear estimator with influence function $\psi(X_i, \varepsilon_i)$, i.e.

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{\sum_i^n \psi(X_i, \varepsilon_i)}{\sqrt{n}} + O_p\left(\frac{1}{\sqrt{n}}\right) \quad (3.11)$$

(b) $E[\psi(X_i, \varepsilon_i) \cdot \psi(X_i, \varepsilon_i)']$ exists.

Assumption 3.4. All the regressors and the error term are uniformly independently distributed on $[0, 1]$.

⁶ For the definition of equicontinuity, see Newey (1991) and Andrews (1994).

Brown and Newey (1992a), Theorem 1, proves the following lemma.

Lemma 3.2. If Assumptions 3.1, 3.2 (a) and 3.3 (a) hold, the following expansions are valid.

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n 1(\rho_i(\hat{\beta}) \leq \rho_j(\hat{\beta})) \\ &= \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j) + \frac{\partial m_\varepsilon(z_j, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \quad (3.12)$$

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, \rho_i(\hat{\beta}) \leq \rho_j(\hat{\beta})) \\ &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, \varepsilon_i \leq \varepsilon_j) + \frac{\partial m_{x\varepsilon}(z_j, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \quad (3.13)$$

Thus if the Assumptions 3.1, 3.2 (a) and 3.3 (a) are satisfied, $T_n(\hat{\beta})$ can be written as

$$\begin{aligned} & T_n(\hat{\beta}) \\ &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, \rho_i(\hat{\beta}) \leq \rho_j(\hat{\beta})) \right. \\ & \quad \left. - \left(\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\rho_i(\hat{\beta}) \leq \rho_j(\hat{\beta})) \right) \right]^2 \\ &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, \varepsilon_i \leq \varepsilon_j) + \frac{\partial m_{x\varepsilon}(z_j, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) \right. \\ & \quad \left. - \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j) + \frac{\partial m_\varepsilon(z_j, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) \right) \right. \\ & \quad \left. \times \left(\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j) \right) + o_p\left(\frac{1}{\sqrt{n}}\right) \right]^2 \\ &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, \varepsilon_i \leq \varepsilon_j) - \left(\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j) \right) \right. \\ & \quad \left. + \left(\frac{\partial m_{x\varepsilon}(z_j, \beta_0)}{\partial \beta'} - x_j \times \frac{\partial m_\varepsilon(z_j, \beta_0)}{\partial \beta'} \right) (\hat{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \right]^2 \\ &= \frac{1}{n} \sum_{j=1}^n \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i \leq x_j, \varepsilon_i \leq \varepsilon_j) - \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j) \right) \right. \\ & \quad \left. \times \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j) \right) + G(X_j, \varepsilon_j) \frac{\sum_i^n \psi(X_i, \varepsilon_i)}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right) \right]^2 \end{aligned} \quad (3.14)$$

where

$$G(X_j, \varepsilon_j) = \frac{\partial m_{x\varepsilon}(z_j, \beta_0)}{\partial \beta'} - x_j \times \frac{\partial m_\varepsilon(z_j, \beta_0)}{\partial \beta'} \quad (3.15)$$

The asymptotic distribution of $T_n(\hat{\beta})$ is given by the following theorem:

Theorem 3.4. If Assumptions 3.1—3.4 hold and, without loss of generality, $\{x_i\}_{i=1}^n$ in the definition of $T_n(\hat{\beta})$ is the first regressor $\{x_{1i}\}$, then $T_n(\hat{\beta})$ has the asymptotic distribution

$$\int_{[0,1]^{k+1}} \left[B_{-1}(1)B(s_1, t) + G(S, t) \int_{[0,1]^{k+1}} \psi(X, \varepsilon) dB(X) dB_\varepsilon(\varepsilon) \right]^2 dS dt \quad (3.16)$$

where

$$B_{-1}(1) = \prod_2^k B_i(1), \quad dB(X) = \prod_1^k dB_i(x_i), \quad dS = \prod_1^k ds_i$$

$B(s, t) = (B_1(s_1) - s_1 B_1(1))(B_\varepsilon(t) - t B_\varepsilon(1))$; B_1, B_2, \dots, B_k and B_ε are standard Brownian motions on $[0, 1]$ and are jointly independent. The integral with Brownian motion is in the sense of Ito integral (see Harrison 1985, Chapter 4).

We make the following remark about Theorem 3.4:

Remark 3.1. It is semiparametric in that: (i) We can always transform the error term of model (3.6), (3.7) such that it is uniformly distributed on $[0, 1]$, $F_\varepsilon[\rho(Y, X, \beta_0)] = F_\varepsilon(\varepsilon)$, a uniformly distributed random variable on $[0, 1]$, where $F_\varepsilon(\cdot)$ is the distribution function of the error term; (ii) The regressors in (3.6), (3.7) can also be transformed such that they are distributed on $[0, 1]^k$:

$$F_\varepsilon[\rho(Y, F_1^{-1} \circ F_1(x_1), \dots, F_k^{-1} \circ F_k(x_k); \beta_0)] = F_\varepsilon(\varepsilon),$$

where $F_i(\cdot)$ is the marginal CDF of x_i and $(F_1(x_1), \dots, F_k(x_k))$ are distributed on $[0, 1]^k$ and the marginal distribution of $F_i(x_i)$, $i = 1, \dots, k$, is the uniform

distribution on $[0, 1]$; (iii) The last transformation is to transform the regressors distributed on $[0, 1]^k$ such that they are independently uniformly distributed on $[0, 1]^k$. Assume (x_1, x_2, \dots, x_k) are distributed on $[0, 1]^k$, not independent, we do the following transformation:

$$\bar{x}_1 = f_1(x_1, x_2, \dots, x_k), \dots, \bar{x}_k = f_k(x_1, x_2, \dots, x_k)$$

such that $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ are independently uniformly distributed on $[0, 1]^k$ and has unique inverse solution:

$$x_1 = g_1(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k), \dots, x_k = g_k(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$$

The existence of this transformation is proved by Lemma 3.3.

Thus if we assume that the regressors are independently uniformly distributed on $[0, 1]^k$, we are making a joint assumptions about the distributions of the regressors and the error term and, about the regression function ρ .

Lemma 3.3. Assume (x_1, x_2, \dots, x_k) are distributed on $[0, 1]^k$ with joint density function $f(x_1, x_2, \dots, x_k)$ and that the marginal distribution of $x_i, i = 1, \dots, k$, is the uniform distribution on $[0, 1]$, then there exists a transformation:

$$\bar{x}_1 = f_1(x_1, x_2, \dots, x_k), \dots, \bar{x}_k = f_k(x_1, x_2, \dots, x_k)$$

such that $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ are independently uniformly distributed on $[0, 1]^k$ and has unique inverse solution:

$$x_1 = g_1(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k), \dots, x_k = g_k(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k).$$

Remark 3.2. When $k > 1$, the influence function (3.11) is a l -dimensional

column vector function. The proof of Theorem should be viewed as a proof with respect to each single argument of the influence column vector function.

Remark 3.3. The asymptotic distribution depends on the number of regressors, the regression function and the distribution parameters of the regressors and the error term. Thus the test is not feasible unless we know the regression function and the distribution parameters of the regressors and the error term. To make this test feasible, we will use bootstrap technique to obtain the corresponding critical values.

Remark 3.4. Theorem 3.4 is an extension of Blum et al.'s (1961) result. When $k = 1$ and $\hat{\beta} = \beta_0$, the asymptotic result⁷ reduces to that of Blum *et al.*

Remark 3.5. Theorem 3.4 is also valid when x is replaced by predicted value $\hat{Y} = \pi(X, \hat{\beta}, 0)$. Regularity assumptions about \hat{Y} is similar to that of the residual \hat{e} .

3.3.2 Bootstrapping the Critical Values of the Statistics

Our test is not feasible unless we know the regression function and the distribution parameters of the regressors and the error term. The following bootstrap procedure is to attain the critical values of the residual-based test statistics to make the test feasible without having to know the aforementioned unknowns. Our bootstrap procedure

⁷ When β is known, the influence function ψ in the formula (3.16) becomes 0. Thus (3.16) becomes

$$\int_{[0,1]^2} [B_{-1}(1)B(s,t)]^2 dSdt$$

which is not exactly the same as that of Blum *et al.* But the Gaussian random process $B_{-1}(1)B(s,t)$ and $B(s,t)$ have the same covariance structure. So their functionals will have the same distribution.

is as follows:

- (i) Randomly draw sample of size n , $(\varepsilon_1^*, \dots, \varepsilon_n^*)$ from residuals (e_1, e_2, \dots, e_n) , $(X_1^*, X_2^*, \dots, X_n^*)$ from the original sample (X_1, X_2, \dots, X_n) with replacement.
- (ii) Calculate $(y_1^*, y_2^*, \dots, y_n^*)$ by $y_i^* = \pi(X_i^*, \varepsilon_i^*, \hat{\beta})$.
- (iii) Use (y_1^*, \dots, y_n^*) from (ii) and $(X_1^*, X_2^*, \dots, X_n^*)$ from (i) to reestimate the model (3.6) and (3.7) and get a new parameter estimate $\tilde{\beta}$ and a new set of residuals (e_1^*, \dots, e_n^*) .
- (iv) Use the x_1^* from (i) and the residuals from (iii) to obtain one value of $T_n^*(\tilde{\beta})$.
- (v) Repeat (i) through (iv) to attain a series of values of $T_{nj}^*(\tilde{\beta})$, for $j = 1, 2, \dots, N$.
- (vi) The estimated critical values are obtained from the empirical distribution of $T_{nj}^*(\tilde{\beta})$, for $j = 1, 2, \dots, N$.

The main task of this section is to find out the asymptotic distributions of $T_n^*(\tilde{\beta})$. With the bootstrapped sample, which is from a discrete population distribution, for any β , the expectations of (3.9) and (3.10) become

$$\begin{aligned}
 m_{\varepsilon}^{*j}(\beta) &= m_{\varepsilon}^*(z_j^*, \beta) \\
 &= E_{z^*} [1(\rho(z^*, \beta) \leq \rho(z_j^*, \beta))] \\
 &= \frac{1}{n} \sum_{i=1}^n 1(\rho(z_i, \beta) \leq \rho(z_j^*, \beta))
 \end{aligned} \tag{3.17}$$

and

$$m_{x\varepsilon}^{*j}(\beta) = m_{x\varepsilon}^*(z_j^*, \beta)$$

$$\begin{aligned}
&= E_{z^*} \left[1 \left(x^* \leq x_j^*, \rho(z^*, \beta) \leq \rho(z_j^*, \beta) \right) \right] \\
&= \frac{1}{n} \sum_{i=1}^n 1 \left(x_i \leq x_j^*, \rho(z_i, \beta) \leq \rho(z_j^*, \beta) \right) \tag{3.18}
\end{aligned}$$

With similar regularity assumptions we prove that the expansions in Lemma 3.2 still hold for the bootstrapped samples. The following assumptions about the bootstrapped samples are analogues of Assumption 3.1 and Assumption 3.3.

Assumption 3.1':

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left[1 \left(\rho_i^*(\beta) \leq \rho_j^*(\beta) \right) - m_{\varepsilon^j}^*(\beta) \right] \right. \\
&\quad \left. - \left[1 \left(\rho_i^*(\hat{\beta}) \leq \rho_j^*(\hat{\beta}) \right) - m_{\varepsilon^j}^*(\hat{\beta}) \right] \right\}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left[1 \left(x_i^* \leq x_j^*, \rho_i^*(\beta) \leq \rho_j^*(\beta) \right) - m_{x\varepsilon^j}^*(\beta) \right] \right. \\
&\quad \left. - \left[1 \left(x_i^* \leq x_j^*, \rho_i^*(\hat{\beta}) \leq \rho_j^*(\hat{\beta}) \right) - m_{x\varepsilon^j}^*(\hat{\beta}) \right] \right\}
\end{aligned}$$

are stochastic equicontinuous at $\hat{\beta}$ for any given z_j^* .

Assumption 3.3': The estimator $\tilde{\beta}$ has the same influence function as $\hat{\beta}$, i.e.

$$\sqrt{n} (\tilde{\beta} - \hat{\beta}) = \frac{\sum_i^n \psi(X_i^*, \varepsilon_i^*)}{\sqrt{n}} + O_p \left(\frac{1}{\sqrt{n}} \right)$$

Lemma 3.4. If there exists a neighborhood $N(\beta_0)$ of β_0 such that

$$\max_{\beta \in N(\beta_0)} E \left| \psi'_{k+1}(X, \rho(X, Y, \beta)) \cdot \partial \rho(X, Y, \beta) / \partial \beta \right| < \infty \tag{3.19}$$

then

$$E [\psi(X_i^*, \varepsilon_i^*) | x_1, \dots, x_n; e_1, \dots, e_n] = \frac{1}{n^2} \sum_{i,j=1}^n \psi(X_i, e_j) = O_p \left(\frac{1}{\sqrt{n}} \right)$$

where $\psi'_{k+1}(X, \varepsilon)$ is the partial derivative of ψ with respect to its last argument.

Lemma 3.5. If Assumptions 3.1, 3.1', 3.2 (a), 3.3 (a) and 3.3' hold, we will have the following expansions

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n 1(\rho_i^*(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) \\ &= \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq \varepsilon_j^*) + \frac{\partial m_\varepsilon(z_j^*, \beta_0)}{\partial \beta'} (\tilde{\beta} - \hat{\beta}) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \quad (3.20)$$

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*, \rho_i^*(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) \\ &= \frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*, \varepsilon_i^* \leq \varepsilon_j^*) + \frac{\partial m_{x\varepsilon}(z_j^*, \beta_0)}{\partial \beta'} (\tilde{\beta} - \hat{\beta}) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \quad (3.21)$$

Lemma 3.6. For fixed j ,

$$\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*) = x_j^* + O_p\left(\frac{1}{\sqrt{n}}\right) \quad (3.22)$$

If Assumptions 3.1 and 3.2 (a) hold, we also have

$$\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq \varepsilon_j^*) = \varepsilon_j^* + O_p\left(\frac{1}{\sqrt{n}}\right) \quad (3.23)$$

Definition 3.4. Our bootstrapped statistics is defined as

$$T_n^*(\beta) = n \iint [B_n^*(x, e)]^2 dG_{x\varepsilon}^*(x, e) \quad (3.24)$$

where

$$\begin{aligned} B_n^*(x, e) &= G_{x\varepsilon}^*(x, e) - G_x^*(x) G_e^*(e) \\ G_x^*(x) &= \frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x) \\ G_e^*(e) &= \frac{1}{n} \sum_{i=1}^n 1(e_i^* \leq e) \\ G_{x\varepsilon}^*(x, e) &= \frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x, e_i^* \leq e) \end{aligned}$$

and $e_i^* = \rho_i^*(\beta) = \rho(z_i^*, \beta)$. Under our bootstrapped procedure, the bootstrapped error term and regressors are stochastically independent.

$T_n^*(\tilde{\beta})$ is our bootstrapped test statistics. The following theorem proves that $T_n^*(\tilde{\beta})$ has the same asymptotic distribution as that of $T_n(\hat{\beta})$.

Theorem 3.5. If Assumptions 3.1, 3.1', 3.2, 3.3, 3.3', 3.4 and condition (3.19) hold, then the bootstrapped statistics $T_n^*(\tilde{\beta})$ has the same asymptotic distribution as that of $T_n(\hat{\beta})$.

Remark 3.6. If the distribution of either the regressors or the error term or both are parameterized, the above bootstrap method is still valid. The practitioner can resample the regressors and/or the error term from their corresponding estimated distribution functions, which are determined by their estimated parameters. Theorem 3.5 still holds if one or both of the samples are drawn from an estimated parameterized distribution, which we assume to be absolutely continuous.

3.3.3 The Power of the Test under Alternatives Assume regressor x_i in model (3.6) and (3.7) are not independent of the error term. First, if the mean of the influence function $\psi(X_i, \varepsilon_i)$ is not zero, then the direct effect of this violation will be that $\sum_{i=1}^n \psi(X_i, \varepsilon_i) / \sqrt{n}$ goes to either positive infinity or negative infinity. By the expression of (3.14), $T_n(\hat{\beta})$ goes to positive infinity when the sample size goes to infinity. Secondly, even if the mean of the influence function $\psi(X_i, \varepsilon_i)$ is zero, the first part in equation (3.14)

$$\sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j, \varepsilon_i \leq \varepsilon_j) - \left(\frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j) \right) \right]^2$$

goes to positive infinity. So the null can be rejected when sample size is bigger enough. Thus this test is *consistent* in that the rejection probability of the null converges to 1 under alternatives.

3.3.4 Motivations of Testing the Independence of Regressors And Error Term

We can now test the independence of regressors and error term of a regression model using residuals. In the literature of semiparametric efficient estimation and residual-based efficient prediction, *e.g.* Brown (1993), (1992) and (1990), Newey (1990) and (1989), Brown and Newey (1992a) and (1992b), Robinson (1991a), among others, semiparametric efficient estimators and predictors are obtained under the assumption that the regressors and the error term of a regression are stochastically independent. This section provides some concrete examples.

Consider, again, estimating the nonlinear simultaneous equation model (3.6) and (3.7). The model has the form

$$\rho(Y, X, \beta_0) = \varepsilon$$

with unique solution

$$Y = \pi(X, \varepsilon, \beta_0).$$

The stochastic assumptions are: (i) $Z_i = (Y_i', X_i')'$ are *i.i.d.* and (ii) ε_i is stochastically independent of regressors X_i .

EXAMPLE 3.1. To estimate the above model, we first consider the best nonlinear three-stage least squares (BNL3S) estimator proposed by Amemiya (1977). This BNL3S estimator is essentially an *IV* estimator. Amemiya chooses the instruments so as to minimize the asymptotic covariance matrix of the *IV* estimator.

Under the assumption (ii), the optimal instruments are given by

$$Q(X) = \Sigma_0^{-1} \bar{R}$$

where

$$\Sigma_0 = E \left[\rho(Y, X, \beta_0) \rho(Y, X, \beta_0)' \right]$$

$$\bar{R} = E_Y [R(Y, X, \beta_0) | X] = E_\varepsilon [R(\pi(X, \varepsilon, \beta_0), X, \beta_0) | X]$$

$$R(Y, X, \beta_0) = \partial \rho(Y, X, \beta_0) / \partial \beta'$$

In order to make this estimator feasible, we must estimate both Σ_0 and \bar{R} . Let $\tilde{\beta}$ be a preliminary consistent estimator of β_0 , the estimator of Σ_0 is obviously the sample average estimated at $\tilde{\beta}$. A number of alternatives have been proposed for the estimation of \bar{R} . Given the assumption of independence, an approach that has been proposed independently by Brown (1990) and Robinson (1991a), is to use residuals to obtain the following estimator

$$\hat{R} = \frac{1}{n} \sum_{j=1}^n R(\pi(X, \tilde{\varepsilon}_j, \tilde{\beta}), X, \tilde{\beta}).$$

According to the discussion in Brown and Newey (1992), this set of instruments are the optimal semiparametric estimates of the target conditional expectation under assumption (ii).

EXAMPLE 3.2. Newey (1990) has several examples of efficient score calculations. All models whose efficient score can be calculated are under either assumption (ii) or assumption that ε is symmetrically distributed conditional on X . If we do not have any confidence in the conditional symmetry distribution of

ε , we would like to know whether ε and X are stochastically independent or not. Newey's (1989) Theorem 3.1 develops locally efficient, residual-based estimators under assumption (ii).

EXAMPLE 3.3. Brown and Mariano (1984) and Brown and Newey (1992a) study residual-based predictors of a regression system and related efficiency properties. If x and ε are independent, the residual-based semiparametric predictor of Y conditional on X , $\int \pi(X, \varepsilon, \beta_0) f(\varepsilon) d\varepsilon$, is the sample average

$$\frac{1}{n} \sum_{i=1}^n \pi(X, \hat{\varepsilon}_i, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^n \pi(X, \rho(Y_i, X_i, \hat{\beta}), \hat{\beta}).$$

If $\hat{\beta}$ is semiparametrically efficient, the above predictor will be semiparametrically efficient. Without independence assumption, residual-based predictions given X is not feasible.

EXAMPLE 3.4. Another example of the sort of residual-based estimation is the joint distribution function estimation of regressors X and endogenous variable Y . Given the assumption that ε is independent of regressor X , the residual-based estimator of the empirical joint distribution of (X, Y)

$$\hat{F}(a, b) = \frac{1}{n} \sum_{i=1}^n 1(X_i \leq a) 1(Y_i \leq b)$$

is

$$\frac{1}{n} \sum_{j=1}^n \frac{1}{n} \sum_{i=1}^n 1(X_i \leq a) 1(\pi(X_i, \hat{\varepsilon}_j, \hat{\beta}) \leq b)$$

Brown and Newey (1992a) show that the residual-based estimator is semiparametric efficient if $\hat{\beta}$ is semiparametric efficient.

We know from the above examples that the independence of the error term ε and regressors X is crucial for residual-based semiparametric efficient estimation

and prediction. This test is to serve as means to test the independence of the error term and the regressors.

3.4 Test of Serial Independence Using Residuals

In time series econometrics, test of serial independence has its applications. For example, a test of random walk can be regarded as a test of serial independence of the first order difference sequence. If the data is observable, Blum *et al.*'s (1961) test can be utilized to test the first order serial independence by splitting the data into two groups, one is the odd index group and the other is the even index group. If data is not observable, for example, if we study the serial independence of a regression error term, the test, for the same reason discussed in previous section, is not feasible. This section demonstrates that our residual-based independence test can be extended to test serial independence.

We still use model (3.6) and (3.7). The null is that the error term $\{\varepsilon_i\}$ is serial independent. The alternative is that it has first order dependence⁸. The following definition is defined in the same way as in (3.8) and (3.24). Due to the regrouping of the sample, we assume that the sample size is $2n$. All the definitions and regularity assumptions in this section is parallel to that in Section 3.1. Thus we omit regularity assumptions, the corresponding lemmas and proof of the main theorem.

⁸ This test is to test the k -th order dependence for any fixed k . If we believe that there is k -th order dependence for several k s, we should conduct this test several times.

Definition 3.5. Our statistics is defined as

$$T_n^s(\beta) = n \iint [B_n(e_1, e_2)]^2 dG_{xe}(e_1, e_2) \quad (3.25)$$

where

$$\begin{aligned} B_n(e_1, e_2) &= G_{12}(e_1, e_2) - G_1(e_1)G_2(e_2) \\ G_1(e_1) &= \frac{1}{n} \sum_{i=1}^n 1(e_{2i-1} \leq e_1) \\ G_2(e_2) &= \frac{1}{n} \sum_{i=1}^n 1(e_{2i} \leq e_2) \\ G_{12}(e_1, e_2) &= \frac{1}{n} \sum_{i=1}^n 1(e_{2i-1} \leq e_1, e_{2i} \leq e_2) \end{aligned}$$

and $e_i = \rho_i(\beta) = \rho(z_i, \beta)$ for $i = 1, 2, \dots, 2n$. Under the independence assumption, the limiting distribution of $T_n^s(\beta_0)$ is that of T in equation (3.2). If $\hat{\beta}$ is an estimator of β_0 , $T_n^s(\hat{\beta})$ will be called the residual-based serial independence test statistics.

The main task in this section is to derive the asymptotic distribution of $T_n^s(\hat{\beta})$.

For any β , we define

$$\begin{aligned} m(z_{2j-1}, \beta) &= E_z [1(\rho(z, \beta) \leq \rho(z_{2j-1}, \beta))] \\ m(z_{2j}, \beta) &= E_z [1(\rho(z, \beta) \leq \rho(z_{2j}, \beta))] \\ m_{12}^j(\beta) &= m_{12}(z_{2j-1}, z_{2j}, \beta) \\ &= E_{12} [1(\rho(z_1, \beta) \leq \rho(z_{2j-1}, \beta), \rho(z_2, \beta) \leq \rho(z_{2j}, \beta))] \end{aligned}$$

where the last expectation is with respect to z_1, z_2 and z_1 and z_2 are stochastically independent.

Under regularity assumptions similar to Assumptions 3.1, 3.2(a) and 3.3(a),

for fixed j , the following expansions hold

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n 1 \left(\rho_{2i-1}(\hat{\beta}) \leq \rho_{2j-1}(\hat{\beta}) \right) \\
&= \frac{1}{n} \sum_{i=1}^n 1 \left(\varepsilon_{2i-1} \leq \varepsilon_{2j-1} \right) + \frac{\partial m(z_{2j-1}, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) + o_p \left(\frac{1}{\sqrt{n}} \right) \\
& \frac{1}{n} \sum_{i=1}^n 1 \left(\rho_{2i}(\hat{\beta}) \leq \rho_{2j}(\hat{\beta}) \right) \\
&= \frac{1}{n} \sum_{i=1}^n 1 \left(\varepsilon_{2i} \leq \varepsilon_{2j} \right) + \frac{\partial m(z_{2j}, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) + o_p \left(\frac{1}{\sqrt{n}} \right) \\
& \frac{1}{n} \sum_{i=1}^n 1 \left(\rho_{2i-1}(\hat{\beta}) \leq \rho_{2j-1}(\hat{\beta}), \rho_{2i}(\hat{\beta}) \leq \rho_{2j}(\hat{\beta}) \right) \\
&= \frac{1}{n} \sum_{i=1}^n 1 \left(\varepsilon_{2i-1} \leq \varepsilon_{2j-1}, \varepsilon_{2i} \leq \varepsilon_{2j} \right) + \frac{\partial m_{12}(\beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) + o_p \left(\frac{1}{\sqrt{n}} \right)
\end{aligned}$$

Thus if Assumption 3.4 holds, $T_n^s(\hat{\beta})$ can be written as

$$\begin{aligned}
T_n^s(\hat{\beta}) &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1 \left(\rho_{2i-1}(\hat{\beta}) \leq \rho_{2j-1}(\hat{\beta}), \rho_{2i}(\hat{\beta}) \leq \rho_{2j}(\hat{\beta}) \right) \right. \\
&\quad \left. - \left(\frac{1}{n} \sum_{i=1}^n 1 \left(\rho_{2i-1}(\hat{\beta}) \leq \rho_{2j-1}(\hat{\beta}) \right) \right) \left(\frac{1}{n} \sum_{i=1}^n 1 \left(\rho_{2i}(\hat{\beta}) \leq \rho_{2j}(\hat{\beta}) \right) \right) \right]^2 \\
&= \frac{1}{n} \sum_{j=1}^n \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n 1 \left(\varepsilon_{2i-1} \leq \varepsilon_{2j-1}, \varepsilon_{2i} \leq \varepsilon_{2j} \right) \right. \\
&\quad \left. - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1 \left(\varepsilon_{2i-1} \leq \varepsilon_{2j-1} \right) \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1 \left(\varepsilon_{2i} \leq \varepsilon_{2j} \right) \right) \right. \\
&\quad \left. + G(X_{2j-1}, X_{2j}; \varepsilon_{2j-1}, \varepsilon_{2j}) \times \sqrt{n} (\hat{\beta} - \beta_0) + o_p \left(\frac{1}{\sqrt{n}} \right) \right]^2
\end{aligned}$$

where

$$\begin{aligned}
& G(X_{2j-1}, X_{2j}; \varepsilon_{2j-1}, \varepsilon_{2j}) \\
&= \left(\frac{\partial m_{12}(\beta_0)}{\partial \beta'} - \varepsilon_{2j-1} \cdot \frac{\partial m(z_{2j}, \beta_0)}{\partial \beta'} - \varepsilon_{2j} \cdot \frac{\partial m(z_{2j-1}, \beta_0)}{\partial \beta'} \right)
\end{aligned}$$

Using exactly the same method as being used in the proof of Theorem 3.4, we can prove the following theorem.

Theorem 3.6. Under similar regularity assumptions as that in Theorem 3.4, $T_n^s(\hat{\beta})$ has the following asymptotic distribution

$$\int_{[0,1]^{2k+2}} \left[B(s, t) + G(X, Y; s, t) \left(\int_{[0,1]^{k+1}} \psi(x, \varepsilon_1) dB_x + \int_{[0,1]^{k+1}} \psi(y, \varepsilon_2) dB_y \right) \right]^2 dS$$

where

$$B(s, t) = (B_1(s) - sB_1(1))(B_2(t) - tB_2(1))$$

$$dB_x = \prod_1^k dB_{xi}(x_i) \cdot dB_1(\varepsilon_1)$$

$$dB_y = \prod_1^k dB_{yi}(y_i) \cdot dB_2(\varepsilon_2)$$

$$dS = dXdYdsdt = \left(\prod_1^k dX_i dY_i \right) dsdt$$

and $B_1, B_2; B_{xi}, B_{yi}, i = 1, \dots, k$, are jointly independent standard Brownian motions on $[0, 1]$.

Remark 3.7. The asymptotic result of Theorem 3.6 is also not feasible. We still use the bootstrapping procedure specified in Section 3.2. to simulate its critical values. Our bootstrapping procedure guarantees that the error terms are randomly drawn, thus serially independent. Using the proof of Theorem 3.5, we can prove that bootstrapped statistics has the same asymptotic distribution as that stated in the above theorem. For brevity, we omit the statement of the theorem and its proof.

3.5 Comparisons of Residual-Based Test with Other Tests

There are already various nonparametric tests for independence available, including the most recent ones of Robinson (1991b) and Brock *et al.* (BDS) (1995).

We do not consider that our residual-based test for independence generally dominates rival ones. We do believe our test has the following appeals: Robinson's test is not justified when one or both of the variables are not observable. Because Brock *et al.*'s test is based on the difference between the expectation of the product of variables and the product of the expectations of the variables and this difference being zero is only a necessary, not a sufficient, condition for independence, BDS test may not have power in some cases, see Dechert (1988) for examples. And also, Brock *et al.*'s residual-based asymptotics is justified only under an *orthogonality condition*, which is satisfied when the regression error term is additive. In nonlinear regression models, the orthogonality condition can not be justified. On the contrary, our test is residual-based, consistent for the general nonlinear regression models. On the other hand, the asymptotics of the tests of Robinson and BDS is standard, ours is nonstandard. Thus bootstrapped simulation procedure is invoked to obtain reference quantiles for our test.

BDS test is not *really* nonparametric in that, contrary to ours, it does not have an invariant finite sample distribution. Also the finite sample distributions of their statistics, when samples are drawn from some population distributions, are significantly different from the asymptotic distribution. The Monte Carlo experiments of Brock *et al.* (1991) shows the size of BDS test under some null distributions are significantly different from each other. Thus, for reference purposes, first, we can not use the asymptotic critical values, because they are significantly different from the finite sample critical values; second, we can not use the simulated finite sample

critical values, because the finite sample distribution will vary significantly under some different null population assumptions. So when implementing the BDS test, one simply can not find an accurate critical value to make references on. Bootstrapping, as Brock *et al.* (1991) suggests, may be a remedy to this problem. On the contrary, the finite sample distribution of the test we promote is *invariant* to the population distributions of the variables under the null.

3.6 Computer Simulations

This Section presents some computer simulation results, power comparisons with BDS test and size bootstrapping.

3.6.1 Power Comparisons with BDS Test As we mentioned earlier the test's finite sample distribution is invariant to any kind of transformation of the variables under the null. It is also invariant to monotonic transformations of the variables under alternatives. Due to the fact that it is variable to non-monotonic transformations under alternatives, by transforming variables, we may increase the power of the test. This is exactly the case for testing the ARCH, GARCH and NLMA models. Table 3.4 presents power comparisons between the BDS test, the test using the original data (ORD) and the test using the square of the data (SQD).

In Table 3.4, all the simulations are done with sample size equaling to 250. The number of replications is 1000 and confidence level is 5%. The first three models are copied from Brock *et al.* (1995), in which all the innovations are *i.i.d.* standard normal. For the last three models, as in Brock *et al.* (1995), we take

		ρ	W_1	W_2	W_3	ORD	SQD
Normal	NLMA	NA	0.28	0.35	0.31	0.08	0.28
	GARCH	NA	0.22	0.22	0.21	0.09	0.23
	ARCH	0.5	0.95	0.89	0.75	0.13	0.82
$T(1)$	ARCH	0.2	1.00	0.98	0.88	0.45	1.00
	ARCH	0.5	1.00	0.99	0.76	0.96	1.00
$T(2)$	ARCH	0.2	0.78	0.66	0.51	0.10	0.46
	ARCH	0.5	1.00	0.98	0.92	0.15	0.91
$T(3)$	ARCH	0.2	0.93	0.82	0.65	0.07	0.18
	ARCH	0.5	1.00	0.99	0.96	0.10	0.48

Table 3.4: Power Comparisons with BDS Test

$[-0.5, 0.5]$ as the characteristic interval and use 5000 replications to simulate the 5% finite sample critical value. The numbers in the table are the percentage the 1000 replicated statistical values being greater than the simulated critical values.

The first model is the nonlinear moving average (NLMA). It has the following form.

$$x_t = 0.5\varepsilon_{t-1}\varepsilon_{t-2} + \varepsilon_t$$

The GARCH and the ARCH models are presented in the following form:

$$x_t \sim N(0, h_t)$$

For the second GARCH model,

$$h_t = 1 + 0.1x_{t-1}^2 + 0.8h_{t-1}$$

For the third ARCH model

$$h_t = 1 + \rho x_{t-1}^2$$

The rest of the ARCH models can be presented as follows. The conditional distribution of x_t , given x_{t-1} , is a t -distribution multiplied by $h_t^{1/2}$, i.e. $T_t(n) * h_t^{1/2}$, where $T_t(n)$ is *i.i.d.* and n is the degrees of freedom.

We observe from Table 3.4 the following results:

(a) BDS test is very powerful when innovations are thin-tailed. When the innovations are fat-tailed, such as the t -distribution with 1 degree of freedom⁹, the test we advocate is also very powerful.

(b) Using the square of the data our test is very powerful. Using the raw data themselves, the test is not powerful, except for the case where the innovations are t -distribution with freedom of degree 1, in which our test is very powerful even using the raw data. Thus we suggest practitioners use some non-monotonic transformations of the data to implement the test and compare the results.

(c) As we mentioned earlier, BDS test needs the *a priori* information that the underlying innovations are normally distributed, as is the basis to obtain the finite sample critical values of the test. But if this assumption is wrong, we are vulnerable to make significant mistakes. On contrary, the statistics we are using has exact finite sample distributions.

3.6.2 Bootstrapping Experiments

The model we simulate on is

$$y = x + e$$

⁹ The t -distribution with 1 degree of freedom is a special kind of Cauchy distribution.

n, p	0.90	0.95	0.99	0.90	0.95	0.99
25	0.9006	0.9435	0.9693	0.9114	0.9470	0.9879
50	0.9154	0.9555	0.9854	0.9094	0.9510	0.9812
100	0.9195	0.9581	0.9850	0.9082	0.9550	0.9840
200	0.9090	0.9552	0.9876	0.9067	0.9510	0.9868
400	0.9109	0.9594	0.9921	0.9050	0.9484	0.9885

Table 3.5: Simulation Results

For different sample size and different distributions of regressor x and error term e , we get Table 3.5. In Table 3.5, n is the sample size, p the reference probabilities. The numbers inside the table are the simulated rejection probabilities corresponding to different sample sizes. The right half of the table corresponds to the specification that both x and e are independent standard normal distributions. The left half of the table corresponds to the specification that both x and e are independent uniform distributions on $[0, 1]$.

Table 3.5 is generated as follows. We bootstrap 1,000,000 times to obtain the critical values. We then calculate the residual-based statistics 100,000 times. The numbers inside the table are the percentages that the residual-based statistics are greater than their corresponding bootstrapped critical values. The confidence intervals, calculated with formula (3.4), of the three quantiles 0.9, 0.95 and 0.99 are $(0.9 - 0.00588, 0.9 + 0.00588)$, $(0.95 - 0.00431, 0.95 + 0.00431)$ and $(0.99 - 0.00196, 0.99 + 0.00196)$ respectively.

3.7 Conclusions of Chapter 3

Semi- and nonparametric tests of independence of two random variables are considered. Among the contributions of this paper are the following. First, the exact finite-sample distributions of Blum *et al.*'s (1961) discretely distributed statistics, based on directly observable variables, are calculated for very small sample sizes, $n = 1, 2, \dots, 8$. Although, in theory, our approach can be used to compute the exact distribution of the statistics for any size of sample, the calculation becomes prohibitively expensive for $n > 10$. Second, Monte Carlo simulations are used to approximate the quantiles for $9 \leq n \leq 200$. On the basis of these simulations, we find that even for $n = 200$, the quantiles of the distribution are still significantly different from Blum *et al.*'s (1961) asymptotic quantiles. Thus, for inference purposes our table will be more accurate than that of Blum *et al.* Third, we obtain the asymptotic distribution of the statistics which is based on variables such as residuals or predictions from a regression model that are not directly observable but must be computed using estimated parameters. This latter test can be useful in specification testing for a class of models. Fourth, since the residual based asymptotics depends on regression function, distribution parameters of the regressors and error term, and the estimator, we prove that the bootstrapped statistics has the same asymptotic distribution as that of the residual and /or predicted value based statistics. Fifth, we extend our residual-based independence test to test serial independence of regression error terms. Sixth, we compare this test with other sorts of tests for independence and present some simulation results.

Also Monte Carlo evidence shows that bootstrap works well.

3.8 Proof of the Theorems in Chapter 3

Proof of Theorem 3.1. We prove that there is a 1-1 correspondence between the consistent sets and the factorials of n , *i.e.* all the combinations of $(1, 2, \dots, n)$.

First, for each combination of $(1, 2, \dots, n)$, there corresponds a consistent set $(12, 13, \dots, 1n; 23, \dots, 2n; \dots; (n-1)n)$. ij always denotes $y_i \geq y_j$ in the following proof. For clarity we give another example. Assume the combination is $(2, 1, 3, \dots, n)$, then the corresponding consistent set will be $(21, 23, \dots, 2n; 13, \dots, 1n; \dots; (n-1)n)$. The rule for constructing the consistent set is that if the first number in a combination is k , then y_k will be the largest in the definition of the consistent set. From this construction we know there are at least $n!$ consistent sets.

Second we prove that for any consistent set, the $(y_i)_{i=1}^n$ can be ranked, *i.e.* there must be some i such that $y_i \geq y_j$ for all $j \neq i$ and some k such that $y_i \geq y_k \geq y_j$ for all $j \neq i$ and $j \neq k$; \dots . We prove this claim using mathematical induction.

If $n = 2$, all the consistent sets are (12) and (21) , y_1 and y_2 can be ranked for both sets.

Assume that when the sample size is less than and/or equal to $n-1$ our claim is valid. We prove that when the sample size is n our claim is still valid. Without loss of generality, let us consider the last $n-1$ observations out of a sample of size n . Suppose its corresponding consistent set is $\Upsilon_{n-1} = (23, \dots, 2n; \dots; (n-1)n)$. The ranking from this consistent set is $y_2 \geq y_3 \geq \dots \geq y_n$.

We claim that any consistent sets constructed from (y_1, y_2, \dots, y_n) being restricted by Υ_{n-1} are of the following forms: $A_k = (21, 31, \dots, k1, 1(k+1), \dots, 1n; \Upsilon_{n-1})$, $k = 1, 2, \dots, n$. It is easy to see the ranking from these sets are $y_2 \geq y_3 \geq \dots \geq y_k \geq y_1 \geq y_{k+1} \geq \dots \geq y_n$. Next, we pick up any set which is not of the form of A_k . Without loss of generality we assume that set $(12, 31, 14, \dots, 1n; \Upsilon_{n-1})$ is picked. This set is not consistent, because from $(y_1 \geq y_2)$ and Υ_{n-1} we deduce that $y_1 \geq y_3$, which is a contradiction to $y_3 \geq y_1$. This set has 0 measure.

Proof of Lemma 3.1. Without loss of generality we assume the consistent set is $(y_1 \geq y_2, y_1 \geq y_3, \dots, y_1 \geq y_n; y_2 \geq y_3, \dots, y_2 \geq y_n; \dots; y_{n-1} \geq y_n)$. The measure of it is

$$\begin{aligned}
 & \int_0^1 dy_1 \int_0^{y_1} dy_2 \cdots \int_0^{y_{n-2}} dy_{n-1} \int_0^{y_{n-1}} dy_n \\
 &= \int_0^1 dy_1 \int_0^{y_1} dy_2 \cdots \int_0^{y_{n-2}} y_{n-1} dy_{n-1} \\
 &= \int_0^1 dy_1 \int_0^{y_1} dy_2 \cdots \int_0^{y_{n-3}} \frac{1}{2!} y_{n-2}^2 dy_{n-2} \\
 &= \int_0^1 \frac{1}{(n-1)!} y_1^{n-1} dy_1 = \frac{1}{n!}
 \end{aligned}$$

Proof of Theorem 3.4. We prove this theorem for $k = 1$. When $k > 1$, it can be proved accordingly. By Blum *et al.* (1961), Kiefer (1959) and Rosenblatt (1952), it is enough to prove that for given s, t, u and v , the asymptotic covariance of $B_n(s, t) + G(s, t) \sum_i \psi(x_i, \varepsilon_i) / \sqrt{n}$ and $B_n(u, v) + G(u, v) \sum_i \psi(x_i, \varepsilon_i) / \sqrt{n}$ equals to the covariance of

$$B(s, t) + G(s, t) \int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1(x_1) dB_\varepsilon(\varepsilon)$$

and

$$B(u, v) + G(u, v) \int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1(x_1) dB_\varepsilon(\varepsilon)$$

where

$$B_n(s, t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i \leq s, \varepsilon_i \leq t) - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i \leq s) \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(\varepsilon_i \leq t) \right)$$

By Blum et al. (1961) the asymptotic covariance of $B_n(s, t)$ and $B_n(u, v)$ equals to the covariance of $B(s, t)$ and $B(u, v)$. By the classical result of stochastic integral (Harrison, 1985)

$$E \left[\int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1(x_1) dB_\varepsilon(\varepsilon) \right]^2 = \int_0^1 \int_0^1 [\psi(x_1, \varepsilon)]^2 dx_1 d\varepsilon$$

which is the variance of $\sum_i \psi(x_i, \varepsilon_i) / \sqrt{n}$. Thus we need only to prove that for any given s and t , the asymptotic covariance $B_n(s, t)$ and $\sum_i^n \psi(x_i, \varepsilon_i) / \sqrt{n}$ equals to the covariance of $B(s, t)$ and $\int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1(x_1) dB_\varepsilon(\varepsilon)$. In the following derivation we use the equality: $E[\psi(x, \varepsilon)] = \int_0^1 d\varepsilon \int_0^1 \psi(x, \varepsilon) dx = 0$. The covariance $B_n(s, t)$ and $\sum_i^n \psi(x_i, \varepsilon_i) / \sqrt{n}$ is

$$\begin{aligned} & E \left[B_n(s, t) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i, \varepsilon_i) \right) \right] \\ &= E \left[\frac{1}{n} \sum_{i,j=1}^n 1(x_i \leq s, \varepsilon_i \leq t) \psi(x_j, \varepsilon_j) - \frac{1}{n^2} \sum_{i,j,k=1}^n 1(x_i \leq s, \varepsilon_j \leq t) \psi(x_k, \varepsilon_k) \right] \\ &= E \left[\frac{1}{n} \sum_{i \neq j} 1(x_i \leq s, \varepsilon_i \leq t) \psi(x_j, \varepsilon_j) + \frac{1}{n} \sum_{i=j} 1(x_i \leq s, \varepsilon_i \leq t) \psi(x_j, \varepsilon_j) \right. \\ &\quad - \frac{1}{n^2} \sum_{i=j=k} 1(x_i \leq s, \varepsilon_j \leq t) \psi(x_k, \varepsilon_k) - \frac{1}{n^2} \sum_{i=j, j \neq k} 1(x_i \leq s, \varepsilon_j \leq t) \psi(x_k, \varepsilon_k) \\ &\quad - \frac{1}{n^2} \sum_{j=k, k \neq i} 1(x_i \leq s, \varepsilon_j \leq t) \psi(x_k, \varepsilon_k) - \frac{1}{n^2} \sum_{i=k, k \neq j} 1(x_i \leq s, \varepsilon_j \leq t) \psi(x_k, \varepsilon_k) \\ &\quad \left. - \frac{1}{n^2} \sum_{i \neq k, k \neq j, i \neq j} 1(x_i \leq s, \varepsilon_j \leq t) \psi(x_k, \varepsilon_k) \right] \end{aligned}$$

$$\begin{aligned}
&= \int_0^t d\varepsilon \int_0^s \psi(x, \varepsilon) dx - \frac{1}{n} \int_0^t d\varepsilon \int_0^s \psi(x, \varepsilon) dx \\
&\quad - \frac{n-1}{n} t \int_0^1 d\varepsilon \int_0^s \psi(x, \varepsilon) dx - \frac{n-1}{n} s \int_0^t d\varepsilon \int_0^1 \psi(x, \varepsilon) dx \\
&= \int_0^t d\varepsilon \int_0^s \psi(x, \varepsilon) dx - t \int_0^1 d\varepsilon \int_0^s \psi(x, \varepsilon) dx - s \int_0^t d\varepsilon \int_0^1 \psi(x, \varepsilon) dx + O_p\left(\frac{1}{n}\right)
\end{aligned}$$

The covariance of $B(s, t)$ and $\int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1(x_1) dB_\varepsilon(\varepsilon)$ is

$$\begin{aligned}
&E \left[(B_1(s) - sB_1(1)) (B_\varepsilon(t) - tB_\varepsilon(1)) \left(\int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1(x_1) dB_\varepsilon(\varepsilon) \right) \right] \\
&= \int_0^t d\varepsilon \int_0^s \psi(x, \varepsilon) dx - t \int_0^1 d\varepsilon \int_0^s \psi(x, \varepsilon) dx \\
&\quad - s \int_0^t d\varepsilon \int_0^1 \psi(x, \varepsilon) dx + st \int_0^1 d\varepsilon \int_0^1 \psi(x, \varepsilon) dx \\
&= \int_0^t d\varepsilon \int_0^s \psi(x, \varepsilon) dx - t \int_0^1 d\varepsilon \int_0^s \psi(x, \varepsilon) dx - s \int_0^t d\varepsilon \int_0^1 \psi(x, \varepsilon) dx
\end{aligned}$$

The last equation holds because $\int_0^1 d\varepsilon \int_0^1 \psi(x, \varepsilon) dx = 0$. Asymptotically both covariances are equal.

Proof of Lemma 3.3. We use mathematical induction to prove that we can construct a transformation such that the claim holds.

If $k = 2$, we do the following transformation

$$\bar{x}_1 = f_1(x_1, x_2) = \int_0^{x_1} f(s, x_2) ds \text{ and } \bar{x}_2 = x_2.$$

The joint density of (\bar{x}_1, \bar{x}_2) is

$$g(\bar{x}_1, \bar{x}_2) = f(x_1, x_2) / \left| \frac{\partial(\bar{x}_1, \bar{x}_2)}{\partial(x_1, x_2)} \right| = 1.$$

The marginal distributions of \bar{x}_1 and \bar{x}_2 are the uniform distribution on $[0, 1]$. Thus (x_1, x_2) is independently uniformly distributed on $[0, 1]^2$. And also the transformation has unique inverse solution.

We assume this Lemma holds for (x_2, \dots, x_k) , i.e. we assume that (x_2, \dots, x_k) are independently uniformly distributed on $[0, 1]^{k-1}$. If (x_1, x_2, \dots, x_k) are independently uniformly distributed on $[0, 1]^k$, no transformation is needed. If not, we do the following transformation

$$\begin{aligned}\bar{x}_1 &= f_1(x_1, x_2, \dots, x_k) = \int_0^{x_1} f(s, x_2, \dots, x_k) ds \\ \bar{x}_j &= x_j \text{ for } j = 2, \dots, k\end{aligned}$$

Then the joint density of $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ is 1 and the marginal density of \bar{x}_1 is also 1. By the assumption about $(\bar{x}_2, \dots, \bar{x}_k)$, we know $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ are independently uniformly distributed on $[0, 1]^k$. Due to the monotonicity of $f_1(x_1, x_2, \dots, x_k)$ given (x_2, \dots, x_k) , this transformation is reversible.

Proof of Lemma 3.4: By Taylor's series expansion

$$\begin{aligned}& \frac{1}{n^2} \sum_{i,j=1}^n \psi(X_i, e_j) \\ &= \frac{1}{n^2} \sum_{i,j=1}^n \left[\psi(X_i, \varepsilon_j) + \psi'_{k+1}(X_i, \rho(X_j, Y_j, \bar{\beta})) \cdot \frac{\partial \rho(X_j, Y_j, \bar{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right]\end{aligned}$$

where $\bar{\beta}$ is between β_0 and $\hat{\beta}$. The first part of the sum is a V-statistics with mean zero. It is $O_p(1/\sqrt{n})$ by V-statistics Central Limit Theorem (see Serfling, 1980). The second part of the sum is also $O_p(1/\sqrt{n})$ by Assumption 3 (a) and condition (3.19).

Proof Lemma 3.5: We only prove the first expansion, the second one can be proved accordingly.

$$\frac{1}{n} \sum_{i=1}^n 1(\rho_i^*(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta}))$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq \varepsilon_j^*) \\
&\quad + \left[\frac{1}{n} \sum_{i=1}^n 1(\rho_i(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) - \frac{1}{n} \sum_{i=1}^n 1(\rho(z_i, \hat{\beta}) \leq \rho(z_j^*, \hat{\beta})) \right] \\
&\quad + \frac{1}{n} \sum_{i=1}^n \{ [1(\rho_i^*(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) - m_{\varepsilon}^{*j}(\tilde{\beta})] \\
&\quad - [1(\rho_i^*(\hat{\beta}) \leq \rho_j^*(\hat{\beta})) - m_{\varepsilon}^{*j}(\hat{\beta})] \}
\end{aligned}$$

By Assumption 3.1', the last 2 lines is $o_p(1/\sqrt{n})$. We will prove that the second line is

$$\frac{\partial m_{\varepsilon}(z_j^*, \beta_0)}{\partial \beta'} (\tilde{\beta} - \hat{\beta}) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

By Lemma 3.2, we know

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n 1(\rho_i(\hat{\beta}) \leq \rho_j^*(\hat{\beta})) \\
&= \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j^*) + \frac{\partial m_{\varepsilon}(z_j^*, \beta_0)}{\partial \beta'} (\hat{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \quad (3.26)
\end{aligned}$$

Because $\tilde{\beta}$ is also a \sqrt{n} -consistent estimator of β , this expansion still holds for $\tilde{\beta}$, i.e.

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n 1(\rho_i(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) \\
&= \frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i \leq \varepsilon_j) + \frac{\partial m_{\varepsilon}(z_j^*, \beta_0)}{\partial \beta'} (\tilde{\beta} - \beta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \quad (3.27)
\end{aligned}$$

(3.26) - (3.27), we get the results.

Proof Lemma 3.6: By our bootstrapping procedure and the Central Limit

Theorem, for fixed j

$$\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*) = \frac{1}{n} \sum_{i=1}^n 1(x_i \leq x_j^*) + O_p\left(\frac{1}{\sqrt{n}}\right) = F_x(x_j^*) + O_p\left(\frac{1}{\sqrt{n}}\right)$$

By assumption $F_x(t) = t$, our claim follows.

The second equality can be proved accordingly using Theorem 1 of Brown and Newey (1992a).

Proof of Theorem 3.5: By Lemma 3.5 and Lemma 3.6, (3.22), $T_n^*(\tilde{\beta})$ can be written as

$$\begin{aligned}
T_n^*(\tilde{\beta}) &= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*, \rho_i^*(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) \right. \\
&\quad \left. - \left(\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\rho_i^*(\tilde{\beta}) \leq \rho_j^*(\tilde{\beta})) \right) \right]^2 \\
&= \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*, \varepsilon_i^* \leq \varepsilon_j^*) + \frac{\partial m_{x\varepsilon}(z_j^*, \beta_0)}{\partial \beta'} (\tilde{\beta} - \hat{\beta}) \right. \\
&\quad \left. - \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq \varepsilon_j^*) + \frac{\partial m_{\varepsilon}(z_j^*, \beta_0)}{\partial \beta'} (\tilde{\beta} - \hat{\beta}) \right) \right. \\
&\quad \left. \times \left(\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*) \right) + o_p\left(\frac{1}{\sqrt{n}}\right) \right]^2 \\
&= \frac{1}{n} \sum_{j=1}^n \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq x_j^*, \varepsilon_i^* \leq \varepsilon_j^*) - \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq x_j^*) \right) \right. \\
&\quad \left. \times \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq \varepsilon_j^*) \right) + G(X_j^*, \varepsilon_j^*) \frac{\sum_i^n \psi(X_i^*, \varepsilon_i^*)}{\sqrt{n}} + o_p(1) \right]^2
\end{aligned}$$

Where

$$G(X_j^*, \varepsilon_j^*) = \frac{\partial m_{x\varepsilon}^*(z_j^*, \beta_0)}{\partial \beta'} - x_j^* \times \frac{\partial m_{\varepsilon}^*(z_j^*, \beta_0)}{\partial \beta'}$$

The expression of $T_n^*(\tilde{\beta})$ is the same as that of $T_n(\hat{\beta})$ except z_j is replaced by z_j^* .

We will use the same technique to prove this theorem as that being used by the proof of Theorem 4. For simplicity we prove the case $k = 1$.

(i) Given s, t, u and v , the asymptotic covariance of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq s, \varepsilon_i^* \leq t) - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq s) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq t) \right)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq u, \varepsilon_i^* \leq v) - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq u) \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(\varepsilon_i^* \leq v) \right)$$

equals the covariance of $B(s, t)$ and $B(u, v)$, $\{\min(u, s) - us\} \times \{\min(v, t) - vt\}$.

(ii) Given s and t , the asymptotic covariance of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq s, \varepsilon_i^* \leq t) - \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n 1(x_i^* \leq s) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq t) \right)$$

and $\sum_i^n \psi(x_i^*, \varepsilon_i^*) / \sqrt{n}$ equals to the covariance of $B(s, t)$ and $\int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1 dB_\varepsilon$.

(iii) The variance of $\sum_i^n \psi(x_i^*, \varepsilon_i^*) / \sqrt{n}$ equals asymptotically to the variance of $\int_0^1 \int_0^1 \psi(x_1, \varepsilon) dB_1 dB_\varepsilon$, which is $\int_0^1 \int_0^1 \psi^2(x_1, \varepsilon) dx_1 d\varepsilon$.

We will use heavily the following facts about the conditional expectation:

$$\begin{aligned} E[1(x_i^* \leq s) | x_1, x_2, \dots, x_n] &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq s) \\ E[1(\varepsilon_i^* \leq t) | e_1, e_2, \dots, e_n] &= \frac{1}{n} \sum_{i=1}^n 1(e_i \leq t) \end{aligned}$$

and Lemma 3.4

$$E[\psi(x_i^*, \varepsilon_i^*) | x_1, \dots, x_n; e_1, \dots, e_n] = \frac{1}{n^2} \sum_{i,j=1}^n \psi(x_i, e_j) = O_p\left(\frac{1}{\sqrt{n}}\right)$$

For the purpose of simplicity, the conditional expectation will not be explicitly denoted. We assume $s \leq u$ and $t \leq v$. The covariance of (i) is

$$\begin{aligned} & \frac{1}{n} E \left\{ \left[\sum_{i=1}^n \left(1(x_i^* \leq s) - \frac{1}{n} \sum_{k=1}^n 1(x_k^* \leq s) \right) \left(1(\varepsilon_i^* \leq t) - \frac{1}{n} \sum_{k=1}^n 1(\varepsilon_k^* \leq t) \right) \right] \right. \\ & \quad \times \left. \left[\sum_{i=1}^n \left(1(x_i^* \leq u) - \frac{1}{n} \sum_{k=1}^n 1(x_k^* \leq u) \right) \left(1(\varepsilon_i^* \leq v) - \frac{1}{n} \sum_{k=1}^n 1(\varepsilon_k^* \leq v) \right) \right] \right\} \\ &= \frac{1}{n} \sum_{i,j=1}^n E \left\{ \left(1(x_i^* \leq s) - s + O_p\left(\frac{1}{\sqrt{n}}\right) \right) \left(1(\varepsilon_i^* \leq t) - t + O_p\left(\frac{1}{\sqrt{n}}\right) \right) \right. \\ & \quad \times \left. \left(1(x_j^* \leq u) - u + O_p\left(\frac{1}{\sqrt{n}}\right) \right) \left(1(\varepsilon_j^* \leq v) - v + O_p\left(\frac{1}{\sqrt{n}}\right) \right) \right\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n E \left\{ \left(1(x_i^* \leq s) - s \cdot 1(x_i^* \leq u) - u \cdot 1(x_i^* \leq s) + su + O_p\left(\frac{1}{n}\right) \right) \right. \\
&\quad \times \left. \left(1(\varepsilon_i^* \leq t) - t \cdot 1(\varepsilon_i^* \leq v) - v \cdot 1(\varepsilon_i^* \leq s) + tv + O_p\left(\frac{1}{n}\right) \right) \right\} \\
&\quad + \frac{1}{n} \sum_{i \neq j}^n O_p\left(\frac{1}{n^2}\right) \\
&= s(1-u)t(1-v) + o_p(1) \\
&= \{\min(u, s) - us\} \cdot \{\min(v, t) - vt\} + o_p(1)
\end{aligned}$$

The covariance in (ii) is

$$\begin{aligned}
&\frac{1}{n} E \left\{ \left[\sum_{i=1}^n \left(1(x_i^* \leq s) - \frac{1}{n} \sum_{k=1}^n 1(x_k^* \leq s) \right) \left(1(\varepsilon_i^* \leq t) - \frac{1}{n} \sum_{k=1}^n 1(\varepsilon_k^* \leq t) \right) \right] \right. \\
&\quad \times \left. \sum_{i=1}^n \psi(x_i^*, \varepsilon_i^*) \right\} \\
&= \frac{1}{n} E \left\{ \sum_{i=j=1}^n \left[(1(x_i^* \leq s) - s)(1(\varepsilon_i^* \leq t) - t) + O_p\left(\frac{1}{n}\right) \right] [\psi(x_j^*, \varepsilon_j^*)] \right\} \\
&\quad + \frac{1}{n} E \left\{ \sum_{i \neq j}^n \left[(1(x_i^* \leq s) - s)(1(\varepsilon_i^* \leq t) - t) + O_p\left(\frac{1}{n}\right) \right] [\psi(x_j^*, \varepsilon_j^*)] \right\} \\
&= E \left\{ \left[(1(x_i^* \leq s) - s)(1(\varepsilon_i^* \leq t) - t) + O_p\left(\frac{1}{n}\right) \right] [\psi(x_i^*, \varepsilon_i^*)] \right\} \\
&= E \{ [1(x_i^* \leq s, \varepsilon_i^* \leq t) - s \cdot 1(\varepsilon_i^* \leq t) - t \cdot 1(x_i^* \leq s) + st] [\psi(x_i^*, \varepsilon_i^*)] \} \\
&\quad + O_p\left(\frac{1}{n}\right) \\
&= \frac{1}{n} \sum_{i,j=1}^n [1(x_i \leq s, \varepsilon_j \leq t) - s \cdot 1(\varepsilon_j \leq t) - t \cdot 1(x_i \leq s) + st] \psi(x_i, \varepsilon_j) + O_p\left(\frac{1}{n}\right) \\
&= \int_0^t d\varepsilon \int_0^s \psi(x, \varepsilon) dx - t \int_0^1 d\varepsilon \int_0^s \psi(x, \varepsilon) dx - s \int_0^t d\varepsilon \int_0^1 \psi(x, \varepsilon) dx + O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

The last equality is obtained by the Central Limit Theorem of V-statistics (see Serfling, 1980).

(iii) The variance of $\sum_i^n \psi(x_i^*, \varepsilon_i^*) / \sqrt{n}$ is

$$E[\psi^2(x_i^*, \varepsilon_i^*)] = \frac{1}{n^2} \sum_{i,j}^n \psi^2(x_i, \varepsilon_j) = \int_0^1 \int_0^1 \psi^2(x_1, \varepsilon) dx_1 d\varepsilon + O_p\left(\frac{1}{\sqrt{n}}\right)$$

The last equality is obtained also by the Central Limit Theorem of V-statistics (see Serfling, 1980).

Thus by the argument of Section 2 of Kiefer (1959)

$$\begin{aligned}
 & T_n^* (\tilde{\beta}) \\
 = & \int_0^1 \int_0^1 \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq s, \varepsilon_i^* \leq t) \right. \\
 & \left. - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n 1(x_i^* \leq s) \right) \left(\frac{1}{n} \sum_{i=1}^n 1(\varepsilon_i^* \leq t) \right) \right]^2 ds dt + o_p(1)
 \end{aligned}$$

By Invariance principle, it converges weakly to the distribution of the random variable specified in Theorem 3.4.

For $k > 1$, the only difference in the proof of the asymptotic results for the bootstrapped statistics from that for the residual based statistics is that the bootstrapped samples $X_i^* = (x_{1i}^*, \dots, x_{ki}^*)$ are no longer drawn from independent distributions on $[0, 1]^k$ but from $(x_{1i}, \dots, x_{ki})_{i=1}^n$. Thus they are not independent in finite samples. But (x_{1i}, \dots, x_{ki}) is drawn from the independent distribution on $[0, 1]^k$, the arguments in X_i^* will become independent asymptotically. The above proof can also go through.

REFERENCES

- Ahn, Hyungtaik and James L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3-29.
- Amemiya T. (1977): "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model," *Econometrica*, 45, 955-968.
- Andrews, Donald W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43-72.
- Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432-452.
- Billingsley, Patrick (1968): *Convergence of Probability Measures*, John Wiley & Sons, New York.
- Blum, J. R., J. Kiefer and M. Rosenblatt (1961): "Distribution Free Tests of Independence Based on the Sample Distribution Function," *Annals of Mathematical statistics*, 32, 485-498.
- Brock, W. A. and W. D. Dechert (1988): "A General Class of Specification Tests: The Scalar Case," in *Business and Economics statistics Section of the Proceedings of the American statistical Society*, 77-79.
- Brock W. A., W. D. Dechert, J. A. Scheinkman and B. LeBaron (1995): "A Test for Independence Based on the Correlation Dimension," *Econometric Review* (forthcoming).
- Brock W. A., D. A. Hsieh and B. LeBaron (1991): *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*, MIT Press.
- Brown, Bryan W. (1990): "Simulation-Based Semiparametric Estimation and Prediction in Nonlinear Systems," Discussion paper, University of Rochester.
- Brown, Bryan W. (1992): "Optimal Stochastic Instrumental Variables Estimation in Nonlinear Systems," memo, Rice University.
- Brown, Bryan W. (1993): "Optimal Endogenous Instrumental Variables Estimation in Nonlinear Systems," memo, Rice University.

- Brown, Bryan W. and R. Mariano (1984): "Residual-Based Stochastic Prediction and Estimation in a Nonlinear Simultaneous System," *Econometrica*, 52, 321-343.
- Brown, Bryan W. and Whitney Newey (1992a): "Efficient Semiparametric Estimation of Expectations," memo, Rice U. and MIT.
- Brown, Bryan W. and Whitney Newey (1992b): "Bootstrapping for GMM," memo, Rice U. and MIT.
- Burguete, J. F., A. R. Gallant, and G. Souza (1982): "On Unification of the Asymptotic Theory of Nonlinear Econometric Models", *Econometric Reviews*, 1, 151-190.
- Cameron A. Collin and Pravin K. Trivedi (1993): "Tests of Independence in Parametric Models: with Applications and Illustrations," *Journal of Business & Economic Statistics*, 11, 29-43.
- Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305-334.
- Cosslett, S. R. (1983): "Distribution-Free Maximum Likelihood Estimation of the Binary Choice Model", *Econometrica*, 51, 765-782.
- Cosslett, S. R. (1987): "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and Censored Regression Models", *Econometrica*, 55, 559-586.
- Deaton, A. (1985): "Panel Data from Time Series of Cross-Sections," *Journal of Econometrics* 30, 109-126.
- Dechert, W. (1988): A Characterization of Independence for a Gaussian Process in Terms of the Correlation Dimension, SSRI Working Paper #8812, Department of Economics, University of Wisconsin, Madison.
- Devroye, L. and T. J. Wagner (1980): "Distribution-Free Consistency Results in Nonparametric Discrimination And Function Estimation," *Annals of Statistics*, 8, 231-239.
- Gallant, A. R. (1987): *Nonlinear Statistical Models*. New York, John Wiley and Sons.
- Gallant, A. R. and D. W. Nychka (1987): "Semi-NonParametric Maximum Likelihood Estimation", *Econometrica*, 55, 363-390.
- Garen, J. (1984): "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica*, 52, 1199-1218.
- Gourieroux, G., A. Monfort, E. Renault and A. Trognon (1987): "Generalized Residuals," *Journal of Econometrics*, 34, 5-32.

- Han, A. K. (1987): "Non-Parametric Analysis of a Generalized Regression Model", *Journal of Econometrics*, 35, 303-316.
- Harrison, J. M. (1985): *Brownian Motion and Stochastic Flow Systems*, John Wiley & Sons.
- Heckman, J. (1978): "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica* 46, 931-961.
- Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.
- Hoeffding, Wassily (1948): "A Non-Parametric Test of Independence," *Annals of Mathematical statistics*, 19, 546-557.
- Horowitz, Joel L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-532.
- Ichimura, H. (1987): "Consistent Estimation of Index Model Coefficients", memo, MIT.
- Kiefer, J. (1959): "K Sample Analogues of the Kolmogorov Smirnov and Cramér von Mises Tests," *Annals of Mathematical statistics*, 30, 420-447.
- Kiefer, J., and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", *Annals of Mathematical Statistics*, 27, 887-960.
- Klein R. W. and R. S. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387-421.
- Lee, L. (1982): "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, XLIX, 355-372.
- Lee, L. (1984): "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity," *Econometrica*, 52, 843-863.
- Little, Roderick (1985): "A Note about Models for Selectivity Bias," *Econometrica*, 53, 1469-1474.
- Maddala, G. S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.
- Manski, C. F. (1984): "The Maximum Score Estimator of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228.
- Manski, C. F. (1988): "Identification of Binary Response Models", *Journal of American Statistical Association*, 83, 729-738.

- Manski, C. F. and S. Thompson (1986): "Operational Characteristics of the Maximum Score Estimator", *Journal of Econometrics*, 32, 85-108.
- Matzkin, Rosa L. (1992): "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models", *Econometrica*, 60, 239-270.
- Ming, Xing and F. Vella (1994a): "Semiparametric Estimation of Sample Selection Models with Multiple Selection Rules," (in progress), Department of Economics, Rice University.
- Ming, Xing and F. Vella (1994b): "A Monte Carlo Comparison of Competing Estimators of the Ordinal Treatment Models," working paper, Department of Economics, Rice University.
- Newey, W. (1988): "Two-Step Series Estimation of Sample Selection Models," unpublished paper, Department of Economics, MIT.
- Newey, Whitney K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161-1168.
- Newey, Whitney K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5,
- Newey, Whitney K. (1989): "Locally Efficient, Residual-Based Estimation of Non-linear Simultaneous Equations," memo, Economics Department, MIT.
- Newey W., J. Powell and F. Vella (1994): "Non-Parametric Instrumental Variables Estimation via Additive Models" working paper.
- Olsen, R. (1980): "A Least Squares Correction for Selectivity Bias," *Econometrica*, 48, 1815-1820.
- Pagan, A. R. and Y. Jung (1993): "Understanding Some Failures of Instrumental Variable Estimators," Working Paper, Australian National University.
- Pagan, A. and A. Ullah (1992): *Non-Parametric Econometrics*, manuscript.
- Pagan, A. R. and F. Vella (1989): "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics*, 4, s29-s60.
- Parzen, E. (1962): "On Estimation of a Probability Density Function and Mode", *The Annals of Mathematical Statistics*, 33, 1065-1076.
- Phillips, Peter C. B. (1983): "ERA's: A New Approach to Small Sample Theory", *Econometrica*, 51, 1505-1527.
- Powell, J. L. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303-325.

- Powell, J. L. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 53, 1435-1460.
- Powell, J. L. (1989): "Semiparametric Estimation of Censored Selection Models," manuscript, Department of Economics, University of Wisconsin at Madison.
- Powell, J. L., J. H. Stock and T. M. Stoker (1989): "Semiparametric Estimation of Weighted Average Derivatives," *Econometrica*, 57, 1403-1430.
- Robinson, Peter J. (1988): "Root- n -Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.
- Robinson P. M. (1991a): "Best Nonlinear Three Stage Least Squares Estimation of Certain Econometric Models," *Econometrica*, 59, 731-754.
- Robinson P. M. (1991b): "Consistent Nonparametric Entropy-Based Testing," *Review of Economic Studies*, 58, 437-453.
- Rosenblatt, M. (1952): "Limit Theorems Associated with Variants of the von Mises statistics," *Annals of Mathematical statistics*, 23 617-623.
- Ruud, P. A. (1983): "Sufficient Conditions for the consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models", 51, *Econometrica*, 225-228.
- Ruud, P. A. (1993): "The Semi-Parametric Maximum Likelihood Estimator of Discrete Dependent Variable Models", memo, Department of Economics, UC Berkeley.
- Serfling R. J. (1980): *Approximation Theorems of Mathematical statistics* (John Wiley & Sons, Inc.).
- Silverman, B. W. (1978): "Weak and Strong Uniform Consistency of Kernel Estimate of a Density and Its Derivatives," *Annals of Statistics*, 6, 177-184.
- Vella, Francis (1993): "A Simple Estimator for Simultaneous Equation Models with Censored Endogenous Regressors," *International Economic Review*, 34, May, 441-457.