ESTIMATION OF SINGLE INDEX MODELS

by

HIDEHIKO ICHIMURA

B.A. Economics
Osaka University
(1981)

Submitted to the Department of Economics
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy in Economics

at the

Massachusetts Institute of Technology

September, 1987

Signature of Author____________________

Department of Economics
September 1, 1987

Certified by____________________

Daniel McFadden
Thesis Supervisor

Accepted by____________________

Richard Eckaus
Departmental Graduate Committee

ACKNOWLEDGEMENTS

I thank my advisor professor Daniel McFadden for providing me with the ideal environment for a graduate study and introducing me to nonparametric literature. His advices were always constructive and I benefitted grately from them. Trying ideas on the board in front of him are thrills only those who experienced can appreciate.

I also thank professor James Powell for advising me for a year before he took a leave to University of Wisconsin. My interest in semiparametric estimation is cultivated by Jim and stimulated by a paper by Manski (1984) and the presidential address given by professor McFadden (1985). I thank professor Manski for encouragements at the initial stage of this thesis. In addition I would like to thank professor Takeshi Amemiya for turning my interest to econometrics and professor Kiyoshi Kuga for introducing me to rigorous thinking.

I have met many stimulating individuals in Cambridge. My classmates Bill English, Steve Herzenberg, and David Wilcox are serious about social problems in general, stimulated me in their "liberal" ways, and at the same time, were sources of good times. Moreover, Steve edited and corrected my English in earlier drafts of the thesis, Bill and David gave comments and corrected my English on virtually everything I wrote in English. I am grateful to them.

Furthermore, I learned from and had fun with Jonathan Feinstein, Rob Gertner, Keith Hylton, Nobuhiro Kiyotaki, Anne Maasland, Kiminori Matsuyama, Bruce Meyer, Danny Quah, Brad Reiff, Jennifer Solomon, Glenn Sueyoshi, and Toshiki Yotsuzuka, among others. Thank you.

I dedicate this thesis to my parents. They supported me in pursuing my interest as I like, encouraged me, and always believed in me.

Surrounded by all these people and many more, I am a lucky man.

ESTIMATION OF SINGLE INDEX MODELS

by

HIDEHIKO ICHIMURA

Submitted to the Department of Economics
on September 1, 1987 in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy in Economics

ABSTRACT

A semiparametric estimator for a class of single index models is proposed. Consistency and asymptotic normality of the estimator are proven and a consistent estimator of the covariance matrix is given. The estimation method is implemented using a simulated sample of size 250. Through a Monte Carlo experiment of 1000 repetitions, small sample properties are examined. The experiment shows that under the circumstances where conventional maximum likelihood estimator is not consistent, the proposed estimator performs the best in terms of the estimated mean square error among existing semiparametric estimators.

Thesis Supervisor : Daniel McFadden

Title : Professor of Economics

1 Introduction.

The regression model

$$y_i = x_i'\beta_0 + \epsilon_i$$

with $\mathrm{E}(\epsilon_i|x_i) = 0$, has played a prominent role in econometric analysis. At the same time careful inspections of economic problems revealed the limit of the regression model and induced efforts to overcome the limitations of the basic model. Recognition of the simultaneous equations problem, nonnegativity restrictions, probabilistic choices, disequilibra in markets, the selectivity bias problem, or the time dependence of economic decisions all lead to constructing appropriate econometric models.

While most models are carefully specified, estimations almost always involve casual specification of error distributions, except for the simultaneous equations model. At first sight specifying distributions seem inevitable, because error terms enter structural models nonlinearly and they are entangled with the conditional expectation. Thus in models within which error terms enter nonlinearly, the reduced forms embody the specification of the error distribution. This is in contrast with the basic regression model where $\mathrm{E}(y_i|x_i) = x_i'\beta_0$ so long as $\mathrm{E}(\epsilon_i|x_i) = 0$, regardless of the distribution of ϵ_i.

Yet, the necessity of the appended distributional assumption is not at all obvious for identifying the parameters in nonlinear error models. In fact not only is it unnecessary, a misspecification of error distributions in general lead to an inconsistent estimator. Therefore it is desirable to construct an estimator which treats distributions as infinite dimensional nuisance parameters: a semiparametric estimator.

Manski (1975) initiated research to relax distributional assumptions studying probabilistic choice models.[1] Since then much effort has been directed towards this goal, but proposed semiparametric estimators rarely have known asymptotic distributions. Even when an asymptotic distribution is known the asymptotic covariance matrix is not easily estimable. An exception is Powell's (1985) least absolute deviation (LAD) estimator for

[1] Manski (1985) proves consistency of Maximum Score (MS) estimator for binary choice models. Amemiya (1985) (pp. 339-346) proves consistency of MS estimator for multinomial choice models.

censored Tobit models.[2] Without a measure of accuracy semiparametric estimators are not widely in use despite their potential usefulness.

This thesis proposes a semiparametric estimator for a class of single index models. Single index models are a strict subclass of the nonadditive error models which include limited dependent variable models, binary choice models, and censored duration models without necessarily assuming proportional hazards model. Under some regularity conditions the estimator is consistent with rate $\sqrt{n}$, the typical rate achieved by parametric estimators under i.i.d. sampling. The asymptotic distribution is normal. A consistent estimator of the covariance matrix is also presented. The result allows economists to focus in specifying structural models and frees them from distributional worries for broader class than before. In addition, even when the error distribution has a theoretical justification it is now possible to perform a specification test of the theoretical distribution with an alternative set of estimates ready when the test is rejected.

The second section defines the class of single index models. There, the class is shown to include censored Tobit models, binary choice models, and duration models among others. The third section motivates the proposed estimator geometrically. The fourth section formally defines the proposed estimator and discusses the identification of the parameters of the model. We show that under some regularity conditions the estimation technique identifies the true parameters up to a multiplicative constant in the linear single index models. We also show that that is the best we can do in the context of single index models. The following two sections provide proofs of consistency and asymptotic normality, respectively. The seventh section presents a consistent estimator for the covariance matrix. The eighth section discusses small sample properties of the estimator examining the result from a Monte Carlo experiment. The last section discusses some directions for future research.

[2]Powell also proposed an estimator for truncated Tobit models which requires the symmetricity of the error distribution.

2 Single Index Models.

This thesis studies estimation of single index models.

DEFINITION **2.1 (Single Index Models)**

$$y_i = \varphi\left(h(x_i;\theta_0)\right) + \epsilon_i \qquad \text{for } i = 1, \ldots, n,$$

where

1. *(y_i, x_i', ϵ_i) for $i = 1, \ldots, n$ are i.i.d. sample of size n,*
2. *$y_i \in R$ and $x_i \in R^K$ are observed, $\epsilon_i \in R$ is an unobserved disturbance, and $\theta_0 \in R^m$ is an unknown true parametor to be estimated,*
3. *The distribution of ϵ_i depends on x_i only through the index $h(x_i;\theta_0)$,*
4. *$E(\epsilon_i|x_i) = 0$,*
5. *The function $h : R^K \times R^m \to R$ is known up to a parameter θ, and*
6. *The function $\varphi : R \to R$ is not known.*

If the function $\varphi : R \to R$ is known, the LS method is feasible. Thus assumption six differentiates the single index model from the conventional additive error models. We shall see that the third assumption is redundunt for consistency but that the assumption plays a crucial role in proving asymptotic normality.

Brillinger (1983) apparently first proposed considering limited dependent variable models using this class of models. He calls this model a generalized linear model, for he considers a case where $h(x;\theta) = x'\theta$. Since a class of models defined by the same name already exists[3] we refer to this class as single index models, following Stoker (1986).

As an illustration of a single index model, consider the following latent dependent variable model. In this model we do not observe y_i^* but observe y_i, which is a transformation of y_i^*. Formally,

$$\begin{cases} y_i^* &= h(x_i;\theta_0) + \nu_i \\ y_i &= \tau(y_i^*) \end{cases} \qquad (i = 1, \ldots, n)$$

[3]See McCullagh and Nelder (1983)

where we do not observe y_i^*.

We assume that x_i and ν_i are independent, and (x_i', ν_i) for $i = 1, \ldots, n$ are i.i.d. Furthermore we assume that $x_i \in R^K$ and $\theta_0 \in R^m$. The function $\tau : R \rightarrow R$ may or may not be known.

If the function $\tau : R \rightarrow R$ takes

$$\tau(s) = \begin{cases} s & \text{if } s > 0 \\ 0 & \text{if } s \leq 0 \end{cases}$$

then censored Tobit model results.

If the function $\tau : R \rightarrow R$ takes

$$\tau(s) = \begin{cases} 1 & \text{if } s > 0 \\ 0 & \text{if } s \leq 0 \end{cases}$$

then binary choice model results.

In both cases the transformation function τ is known but $E(y \mid x)$ is not known unless the distribution of the error term is specified. Knowing the transformation function τ does not allow the LS estimation.

Regardless of a transformation function, the reduced form takes the specification of a single index model as the following calculation shows:

$$E(y \mid x) = \int_{-\infty}^{+\infty} \tau\left(h(x; \theta_0) + \nu\right) dF_\nu$$

where F_ν is the distribution function of a random variable ν_i. By defining the difference between the observable y_i and the conditional expectation to be ϵ_i,

$$\epsilon_i = y_i - E(y_i | x_i),$$

we can transform this class of latent variable models into the single index model defined in 2.1.

Applications of single index models are not restricted to censored Tobit models and binary choice models. Since the transformation function τ is completely unspecified, single index models can be regarded also as an alternative to the errors in variable formulation of regression models. For example,[4] suppose y_i^* is unobserved true profit of firm i and we only observe reported profit y_i. We could assume that reported profit is true profit plus an error term. This is the errors in variable formulation. An alternative is the model $y = \tau(y^*)$. With this modeling strategy we allow reported profit to systematically differ from true profit and yet we are able to consistently estimate the relative determinants of firm profitability.

Another model which can be regarded as a member of the single index class is truncated Tobit models. Here, unlike censored Tobit models, we do not observe x_i when the corresponding dependent variable y_i is censored. Truncated Tobit models can be written as

$$y_i = h(x_i; \theta_0) + u_i$$

where u_i has a density

$$1\left(u > -h(x_i; \theta_0)\right) F_\nu(du)[1 - F_\nu\left(-h(x_i; \theta_0)\right)]^{-1},$$

where $1(\cdot)$ is an indicator function. Therefore $E(\, y_i | x_i\,)$ is

$$h(x_i; \theta_0) + \int_{-h(x_i;\theta_0)}^{\infty} u F_\nu(du)[1 - F_\nu\left(-h(x_i; \theta_0)\right)]^{-1},$$

and thus it has the form $\varphi\left(h(x_i; \theta_0)\right)$. Thus the truncated Tobit model also belongs to the class of single index models.

Finally we show that as a special case, the class of single index models also includes duration models if

1. censoring is random,

2. exogenous variables are time independent, and

3. individual heterogeneity is independent from exogenous variables.

[4] I thank Rob Gertner for suggesting this example. A conversation with Jonathan Feinstein was also helpful.

To see this, let the conditional density function of a duration spell t_i given x_i for individual i without censoring be

$$f_{T_i}(t_i) = f_1\left(t_i; h(x_i;\theta_0)\right) f_2(\alpha_i),$$

where $f_2(\alpha_i)$ denotes a density for individual heterogeneity. Let f_C be the density for censoring. Then the density function of a duration spell t_i with censoring is

$$f_{T_i}(t_i) + f_C(t_i) - F_{T_i}(t_i) f_C(t_i) - F_C(t_i) f_{T_i}(t_i),$$

where F_{T_i} and F_C denote distribution functions of the duration spell and the censoring point, respectively. Therefore the conditional expectation of the duration spell conditioned on the regressors has the form $\varphi\left(h(x_i;\theta_0)\right)$ as desired.

The single index class abstracts specific structures of models included in the class. In consequence we do not take advantage of particular restrictions each model possesses other than those already exploited by the formulation of single index class.

For binary choice models the unknown function φ is the cumulative distribution function. Thus it is a nondecreasing function with range between zero and one. For censored Tobit models

$$\varphi(h(x_i;\theta_0)) = \int_{-\infty}^{h(x_i;\theta_0)} F_\nu(s) ds,$$

and thus φ is nondecreasing function which is an integral of the cumulative distribution function of the error term.

Nevertheless in single index framework both functions are specified only as a measurable function with some finite moments. We shall see that the abstraction costs a loss in identification for some models in single index class, but not all.

Typically, Tobit models allow identification of all parameters and binary choice models allow identification of parameters up to a multiplicative constant. In both cases we shall see slope coefficients are identified up to a multiplicative constant.

3 Idea Behind the Estimation Method.

Before defining the estimation method for single index class, I give a geometric explanation for why the proposed estimator works. Recall the definition of the model

$$y_i = \varphi\left(h(x_i;\theta_0)\right) + \epsilon_i \qquad (i = 1, \ldots, n).$$

The estimation method is based on the following three facts:

1. The variation in y results both from the variation in $h(x;\theta_0)$ and from the variation in ϵ.

2. Nevertheless, on the contour line $h(x;\theta_0) = c$ where c is a given constant, the variability in y results only from the variation in ϵ.

3. Fact 2 does not necessarily hold on a contour line defined by $h(x;\theta) = c$ for $\theta \neq \theta_0$. Along this contour line the value of $h(x;\theta_0)$ changes and therefore the variability in y on the contour line $h(x;\theta) = c$ again results from both the variations in $h(x;\theta_0)$ and in ϵ.

These three facts indicate a way to identify θ_0. Since the conditional variance

$$\mathrm{Var}\left(y \mid h(x;\theta) = c\right)$$

measures the variability in y on a contour line $h(x;\theta) = c$, a sensible way to estimate θ_0 is, first construct a sample analog of

$$\mathrm{E}\left[\mathrm{Var}\left(y \mid h(x;\theta)\right)\right]$$

as the objective function, and second, find θ which minimizes the objective function.

Here, a particular weighting implied by the expectation operator is not only a natural choice but also a necessity. Namely, objective function

$$\int W(c)\mathrm{Var}\left(y \mid h(x;\theta) = c\right) dc,$$

where $W(c)$ is some weighting function, does not yield a consistent estimator in general.

Rewriting the objective function obviates the reason . Using the identity

$$\mathrm{Var}\,(\,y \mid h(x;\theta)\,) = \mathrm{Var}\,[\,\varphi\,(h(x;\theta_0)) \mid h(x;\theta)\,] + \mathrm{Var}(\,\epsilon^2 \mid h(x;\theta)\,),$$

the objective function is

$$\int W(c)\mathrm{Var}\,[\,\varphi\,(h(x;\theta_0)) \mid h(x;\theta) = c\,]\,dc + \int W(c)\mathrm{Var}(\,\epsilon^2 \mid h(x;\theta) = c\,)dc.$$

Although $\theta = \theta_0$ makes the first term vanish, it does not necessarily minimize the sum. But for weighting implied by the expectation operator the second term is constant by the iterated expectation argument. That is $\mathrm{E}\,[\mathrm{Var}\,(\,\epsilon \mid h(x;\theta)\,)]$ is the unconditional variance of ϵ which is a constant, say σ_ϵ^2. Therefore $\theta = \theta_0$ does minimize the sum.

For identification we should ask whether θ_0 is the only θ which achieves the minimum value σ_ϵ^2. Unfortunately this is not the case in general. Since the function φ is not restricted, two θs, say θ_1 and θ_2, are not distinguishable if $\varphi\,(h(x;\theta_1)) = \tilde{\varphi}\,(h(x;\theta_2))$ almost surely in x for some $\tilde{\varphi}$. For example, consider a linear single index model

$$y_i = \varphi(\alpha_0 + x'\beta_0) + \epsilon_i.$$

Considering the identity

$$\varphi(\alpha_0 + x'\beta_0) = \tilde{\varphi}(\alpha_1 + x'(\gamma\beta_0)),$$

for any α_1 and $\gamma \neq 0$, where $\tilde{\varphi}(s) \equiv \varphi(\gamma^{-1}(s - \alpha_1) + \alpha_0)$, at best slope coefficients are identified up to a multiplicative constant, and the costant term is not identified.

Thus define equivalence classes in R^m where θs lie :

$$\Theta_{\tilde{\theta}} = \left\{\,\theta \in R^m \mid \tilde{\varphi}(h(x;\theta)) = \varphi(h(x;\tilde{\theta}))\ \text{a.s. in } x \text{ for some } \tilde{\varphi}\,\right\}.$$

$\Theta_{\tilde{\theta}}$ denotes an equivalence class represented by $\tilde{\theta}$. In single index class we identify θ_0 up to the equivalence class Θ_{θ_0} at best. In the following section I define the estimator formally and prove that in fact we do identify θ_0 up to the equivalence class Θ_{θ_0}.

4 The Estimator and the Identification Condition.

The heuristic argument in the previous section suggests

$$\mathrm{E}\left[\mathrm{Var}\left(\, y | h(x;\theta)\,\right)\right]$$

as the objective function. In this section we formally prove the identification theorem and define the estimator.

The following theorem shows that minimizing

$$\mathrm{E}\left[\mathrm{Var}\left(\, y | h(x;\theta)\,\right)\right]$$

does identify parameters as much as we can expect in single index class.

THEOREM **4.1 (Identification)** *Suppose we have a single index model defined in Definition 2.1. For each $\tilde{\theta} \in R^m$ define the equivalence class*

$$\Theta_{\tilde{\theta}} = \left\{ \theta \in R^m \mid \tilde{\varphi}\left(h(x;\theta)\right) = \varphi(h(x;\tilde{\theta})) \text{ a.s. in } x \text{ for some } \tilde{\varphi} \right\},$$

where $\Theta_{\tilde{\theta}}$ denotes an equivalence class represented by $\tilde{\theta}$. Let

$$H(\theta) \equiv \mathrm{E}\left[\mathrm{Var}\left(\, y \mid h(x;\theta)\,\right)\right].$$

Then $\theta \notin \Theta_{\theta_0}$ implies $H(\theta_0) < H(\theta)$.

PROOF.
If $\theta \notin \Theta_{\theta_0}$, then $\mathrm{E}(\varphi | h(x;\theta)) \neq \varphi(h(x;\theta_0))$ with positive probability. Hence, for this θ, $\mathrm{E}[\mathrm{Var}(\varphi | h(x;\theta))]$ is strictly positive. Then, inequality

$$H(\theta_0) = \sigma_\epsilon^2 < \mathrm{E}[\mathrm{Var}(\varphi | h(x;\theta))] + \sigma_\epsilon^2 = H(\theta)$$

implies the claim.

∎

Usefulness of the estimator crucially depends on the size of the equivalence class Θ_{θ_0}. Suppose φ is almost surely constant. Then $\Theta_{\theta_0} = R^m$, and thus the estimator is not useful. Consider another example where all the regressors in a linear model are discrete. Then in general, once again we can not identify the parameters.[5] To see this consider the example illustrated by Figure 1. Black dots in Figure 1 denote the probability masses. They are located at the intersections of vertical and horizontal lines drawn through the points $x_1 = \{\ldots, -2d, -d, 0, d, 2d, \ldots\}$ and $x_2 = \{\ldots, -2d, -d, 0, d, 2d, \ldots\}$, where d is a positive number. Any line with a slope r where r is an irrational number intersects with black dots at most once. Therefore there is no variation in φ along any line with irrational slope.

[5]We may be able to bound the region in which the true parameters lie.

Nevertheless Lemma 4.1 does show the identification of the coefficients up to a multiplicative constant in linear single index models.

LEMMA **4.1 (Identification of the Linear SIM)** *Consider a linear single index model*

$$y_i = \varphi(\alpha_0 + x_i'\beta_0) + \epsilon_i.$$

Let $B \subset R^m$ be the parameter space and S_b be the support of $x'b$. Assume

1. *The coefficient of at least one continuous regressor is nonzero.*
2. *The range of the support of a continuous regressor is infinite.*
3. *The unknown function φ is differentiable.*
4. *φ is not periodic, namely for each $b \in B$ the following holds:*

$$\forall c \neq 0, \quad \exists t \in S_b \text{ such that } \varphi(t+c) \neq \varphi(t).$$

Then we identify β_0 up to a multiplicative constant.

PROOF.
If for some $b \in B$ which is not proportional to β_0,

$$\mathrm{E}\left[\mathrm{Var}\left(y|x'\beta_0\right)\right] = \mathrm{E}\left[\mathrm{Var}\left(y|x'b\right)\right].$$

Then using $y = \varphi(x'\beta_0) + \epsilon$ and $E(\epsilon|x) = 0$,

$$\varphi(x'\beta_0) = E\left(\varphi|x'b\right) \qquad \text{a.e. in } x.$$

Defining $x'b = t$,

$$\varphi\left(\gamma_1 t + \gamma_2 x_2 + \cdots + \gamma_m x_m\right) = E\left(\varphi|t\right),$$

where $\gamma_1 = \beta_1/b_1$ and $\gamma_i = \beta_i - \gamma_1 b_i$ for $i = 2, \ldots, m$. Suppose x_i is a continuous random variable. Take a partial derivative with respect to x_i. Then $\varphi'(x'\beta_0) \cdot \gamma_i = 0$. From assumption four φ is not constant. Therefore $\gamma_i = 0$. Substituting $\gamma_i = 0$ we have without loss of generality

$$\varphi\left(\gamma_1 t + \gamma_2 x_2 + \cdots + \gamma_m x_m\right) = E\left(\varphi|t\right),$$

where $x_2, \ldots, x_m$ are all discrete random variables. Furthermore since b is not proportional to β_0, there exits at least one i such that $\gamma_i \neq 0$. Rename such i as 2. Take two different values of x_2, say c_1 and c_2. Then for all t,

$$\varphi\left(\gamma_1 t + \gamma_2 c_1 + \cdots + \gamma_m x_m\right) = E\left(\varphi|t\right) = \varphi\left(\gamma_1 t + \gamma_2\right)c_2 + \cdots + \gamma_m x_m\right).$$

Defining $s = \gamma_1 t + \gamma_2 c_1 + \cdots + \gamma_m x_m$ and $c = \gamma_2(c_2 - c_1)$,

$$\varphi(s) = \varphi(s + c).$$

But this contradicts no periodicity.

∎

These identification conditions provide sufficient conditions for warrranting the function

$$H(\theta) \equiv \mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]$$

to attain the minimum only at the true value. Of course function $H(\theta)$ itself is not known without a distributional assumption, and hence we construct a sample analog of $H(\theta)$. The feasible estimator is then defined as the minimizer of the sample analog.

There are at least a few ways to estimate $H(\theta)$. The one which lead to the simplest expression is based on the identity

$$\mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right] = \mathrm{E}\{y\left[y - \mathrm{E}\left(y|h(x;\theta)\right)\right]\}.$$

Since $\mathrm{E}(y^2)$ is constant with respect to θ, the identity leads to an estimator which maximizes a sample analog of $\mathrm{E}[y\mathrm{E}\left(y|h(x;\theta)\right)]$. Unfortunately, I cannot show that the second derivative of a sample analog converges to a negative definite matrix.

Instead, we estimate $H(\theta)$ based on the identity

$$\mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right] = \mathrm{E}\{\left[y - \mathrm{E}\left(y|h(x;\theta)\right)\right]^2\}.$$

If $\mathrm{E}\left(y|h(x;\theta)\right)$ is known, then the sample analog of $H(\theta)$ is,

$$n^{-1}\sum_{i=1}^{n}\left[y_i - \mathrm{E}\left(y_i|h(x_i;\theta)\right)\right]^2.$$

Since $\mathrm{E}\left(y|h(x;\theta)\right)$ is not known we replace it with a kernel estimator. Thus:

Definition 4.1 (Objective Function) *Let*

$$J_n(\theta) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\mathrm{E}}_{ni}(\theta))^2,$$

where

1. *n is the sample size,*
2. $$\hat{E}_{ni}(\theta) = \frac{\sum_{j\neq i} y_j K\left(\frac{h(x_i;\theta) - h(x_j;\theta)}{a_n}\right)}{\sum_{l\neq i} K\left(\frac{h(x_i;\theta) - h(x_l;\theta)}{a_n}\right)},$$
3. *$K : R \to R$ is a one dimensional density function,*
4. *$a_n > 0$ and $a_n \to 0$.*

$\hat{\mathrm{E}}_{ni}(\theta)$ is a kernel estimator for $\mathrm{E}\,(\,y_i \mid h(x_i;\theta)\,)$. Conceivably one can use any esimator of $\mathrm{E}\,(\,y_i \mid h(x_i;\theta)\,)$ in place of a kernel estimator. At least there are two advantages in using a kernel estimator.[6] The first advantage is that the objective function is differentiable if a differentiable kernel function is used. The second advantage is that when a kernel estimator is used, a derivative of the objective function converges to a derivative of the limiting function.[7]

In order to identify the parameters the parameter space Θ needs to be restricted. A restriction depends on the function $h(x;\theta)$. For example if $h(x;\theta) = \alpha + x'\beta$ then we can set α to zero and set a coefficient, say β_1, to one. Assuming that the assumptions in Lemma 4.1 hold, the rest of the parameters are identified. Then β_i estimates β_{i0}/β_{10}. We assume that the identifying restrictions are already imposed in the parameter space Θ.

With these caveats in mind we can now define the estimator. I call the estimator semiparametric least squares (SLS).

Definition 4.2 (SLS) *The estimator $\hat{\theta}_n$ minimizes $J_n(\theta)$, where $\hat{\theta}_n \in \Theta$.*

[6] I thank professor McFadden for a suggestion to use a kernel estimator.

[7] See Appendix 1 for a brief explanation of a kernel estimator.

5 Consistency Proof.

Now we turn to the large sample properties of the estimator. In this section consistency of the estimator is proven. In the following section asymptotic normality shall be proven. The approach taken is first used by A. Wald (1949). We prove that the objective function $J_n(\theta)$ converges uniformly in θ to $\mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]$. The identification condition guarantees that the true value θ_0 minimizes the limiting function. For example Theorem 4. 1. 1. of Amemiya (1985) shows that this result together with the compactness of the parameter space and continuity of the objective function with respect to the parameters implies consistency.

We prove uniform convergence of $J_n(\theta)$ to $\mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]$ in two steps. First we prove that $\tilde{J}_n(\theta)$ converges uniformly to $\mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]$, where

$$\tilde{J}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \mathrm{E}\left(y_i|h(x_i;\theta)\right)\right]^2.$$

Second we prove that $J_n(\theta) - \tilde{J}_n(\theta)$ converges uniformly to zero. The inequality

$$\begin{aligned} & \sup_\theta \left|J_n(\theta) - \mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]\right| \\ \leq\; & \sup_\theta \left|\tilde{J}_n(\theta) - \mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]\right| + \sup_\theta \left|J_n(\theta) - \tilde{J}_n(\theta)\right| \end{aligned}$$

then implies the uniform convergence of $J_n(\theta)$ to $\mathrm{E}\left[\mathrm{Var}\left(y|h(x;\theta)\right)\right]$. The first step is well established in literature. For example Mickey et.al. (1963) establish

LEMMA **5.1** *Let g be a function on $X \times \Theta$ where X is a Euclidean space and Θ is a compact subset of a Euclidean space. Let $g(x,\theta)$ be a continuous function of θ for each x and a measurable function of x for each θ. Assume also that $|g(x,\theta)| \leq h(x)$ for all x and θ, where h is integrable with respect to a probability distribution function F on X. If $x_1, x_2, \ldots$ is a random sample from F then for almost every sequence (x_i)*

$$n^{-1}\sum_{i=1}^{n} g(x_i,\theta) \rightarrow \int g(x,\theta)dF(x)$$

uniformly in θ.

PROOF.
See Jennrich (1969).

∎

The second step is intrinsic to our approach. It shows that approximating the conditional expectation does not alter the consistency of the estimator. We use the following lemma in proving the second step.

LEMMA **5.2** *Let K be a symmetric, bounded density function on $[-c, c]$ and suppose the parameter space $\Theta \times T$ is compact. Let $f(t;\theta)$ be a density function for a random variable $h(x;\theta)$. Assume also that $\mathrm{E}(y|h = t) f(t;\theta)$ is uniformly continuous, twice continuously differentiable, and the second derivative satisfies a Lipschitz condition. If $(x_1, y_1), (x_2, y_2), \ldots$ is a random sample, $|y| < M$ for some $M \in R$, $na_n \to \infty$, and $a_n \to 0$ then*

$$(na_n)^{-1} \sum_{i=1}^{n} y_i K\left[(t - h_i(\theta))/a_n\right] - \mathrm{E}(y|h = t) f(t;\theta)$$

converges uniformly to zero in t and in θ.

PROOF.
We first break

$$(na_n)^{-1} \sum_{i=1}^{n} y_i K\left[(t - h_i(\theta))/a_n\right] - \mathrm{E}(y|h = t) f(t;\theta)$$

into two parts:

$$(na_n)^{-1} \sum_{i=1}^{n} y_i K\left[(t - h_i(\theta))/a_n\right] - \mathrm{E}\left\{y/a_n K\left[(t - h(\theta))/a_n\right]\right\}$$

and

$$\mathrm{E}\left\{y/a_n K\left[(t - h(\theta))/a_n\right]\right\} - \mathrm{E}(y|h = t) f(t;\theta).$$

Since the second part is $O(a_n^2)$ we only need to show that the first part converges uniformly to zero.

For conciseness denote

$$K_n(t, \theta) = K[(t - h(x;\theta))/a_n]$$

and

$$\mathrm{P_n}g = n^{-1}\sum_{i=1}^{n} g(x_i).$$

Without loss of generality assume that $|\theta| \leq 1$ for all $\theta \in \Theta$ and $|t| \leq 1$ for all $t \in T$. Partition Θ into N_1 cubes with the length of a side $a_n\delta$ and T into N_2 intervals of length $a_n\delta$, where δ is a small and positive real number. Then $N_1 = O(a_n^{-m})$ and $N_2 = O(a_n^{-1})$ and space $\Theta \times T$ is partitioned into $N_1 \times N_2 = N$ of $(m+1)$-dimensional cubes, B_k^N, $\{k = 1, \ldots, N\}$ which become smaller and smaller as n gets larger. Now pick a point (θ_k^N, t_k^N) from each B_k^N for $k = 1, \ldots, N$. Then

$$\begin{aligned}
& \Pr\left\{\sup_{(\theta,t)\in\Theta\times T} \left|\mathrm{P_n}\left\{ya_n^{-1}K_n(t,\theta) - \mathrm{E}\left[ya_n^{-1}K_n(t,\theta)\right]\right\}\right| > \epsilon\right\} \\
\leq\ & \Pr\left\{\textstyle\bigcup_{k=1}^{N} \sup_{(\theta,t)\in B_k^N} \left|\mathrm{P_n}\left\{ya_n^{-1}K_n(t,\theta) - \mathrm{E}\left[ya_n^{-1}K_n(t,\theta)\right]\right\}\right| > \epsilon\right\} \\
\leq\ & \textstyle\sum_{k=1}^{N} \Pr\left\{\sup_{(\theta,t)\in B_k^N} \left|\mathrm{P_n}\left\{ya_n^{-1}K_n(t,\theta) - \mathrm{E}\left[ya_n^{-1}K_n(t,\theta)\right]\right\}\right| > \epsilon\right\} \\
\leq\ & \textstyle\sum_{k=1}^{N} \Pr\left\{\left|\mathrm{P_n}\{ya_n^{-1}K_n(t_k^N,\theta_k^N) - \mathrm{E}[ya_n^{-1}K_n(t_k^N,\theta_k^N)]\}\right| > \epsilon/2\right\} \\
+\ & \textstyle\sum_{k=1}^{N} \Pr\left\{\sup_{(\theta,t)\in B_k^N} \left|\mathrm{E}[ya_n^{-1}K_n(t_k^N,\theta_k^N)] - \mathrm{E}[ya_n^{-1}K_n(t,\theta)]\right| > \epsilon/4\right\} \\
+\ & \textstyle\sum_{k=1}^{N} \Pr\left\{\left|\mathrm{P_n}\{\sup_{(\theta,t)\in B_k^N}[ya_n^{-1}(K_n(t,\theta) - K_n(t_k^N,\theta_k^N))]\}\right| > \epsilon/4\right\}
\end{aligned}$$

From Bernstein's inequality each element in the summation of the first term of the last expression is bounded by $2\exp[-n^2a_n^2/(na_n + Mna_n/3)]$. Since $N = O(a_n^{-(m+1)})$ the expression converges to zero if na_n goes to infinity.

The second term of the last expression is zero for small enough $\delta > 0$, because by assumption $\mathrm{E}(y|h = t)f(t;\theta)$ is uniformly continuous.

We turn to the third term of the last expression. The expectation of

$$\sup_{(\theta,t)\in B_k^N} \{y[K_n(t,\theta) - K_n(t_k^N,\theta_k^N)]\}$$

is $o(1)$. Applying Bernstein's inequality as before the third term is also $o(1)$.

∎

The following is an immediate Corollary.

COROLLARY 5.1 *Under the same assumption as in Lemma 5.2 if in addition the density of h satisfies*

$$f(t;\theta) > \delta > 0,$$

then

$$\frac{\sum_{i=1}^{n} y_i K\left[(t - h_i(\theta))/a_n\right]}{\sum_{i=1}^{n} K\left[(t - h_i(\theta))/a_n\right]}$$

converges uniformly in t and in θ to $\mathrm{E}(y|h(\theta) = t)$.

PROOF.

Apply Lemma 5.2 to the denominator with $y_i = 1$.

∎

Now it is straight forward to show

LEMMA 5.3 *Under the assumptions of Lemma 5.2 and Corollary 5.1*

$$J_n(\theta) - \tilde{J}_n(\theta)$$

converges uniformly to zero.

PROOF.

The difference between $J_n(\theta)$ and $\tilde{J}_n(\theta)$ is

$$n^{-1}\sum_{i=1}^{n}\left(\hat{\mathrm{E}}_{ni}(\theta) - \mathrm{E}_i(\theta)\right)\left(\hat{\mathrm{E}}_{ni}(\theta) + \mathrm{E}_i(\theta)\right) - 2n^{-1}\sum_{i=1}^{n} y_i\left(\hat{\mathrm{E}}_{ni}(\theta) - \mathrm{E}_i(\theta)\right).$$

To prove that both terms converge in probability to zero uniformly in θ it is sufficient to invoke Corollary 5.1 and note that by assumption y and $\mathrm{E}_i(\theta)$ are uniformly bounded.

∎

As noted at the beginning of the section if we use continuous kernel, then Lemma 5.3 implies consistency of the estimator.

6 Asymptotic Normality Proof.

In this section we show that the estimator has normal distribution asymptotically. The following two theorems are useful for our purpose.

THEOREM **6.1** **($\sqrt{n}$-Consistency)** *Suppose*

1.
$$J_n(\theta) = J_n(\theta_0) + (\theta-\theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} + 1/2(\theta-\theta_0)'V_{n0}(\theta-\theta_0) + o_p(|\theta-\theta_0|^2),$$

2.
$$\frac{\partial J_n(\theta_0)}{\partial\theta} = \vec{O}_P(1/\sqrt{n}),$$

3. $|V_{n0} - V_0| = o_p(1)$ *for some* V_0,

4. V_0 *is positive definite,*

5. $\hat{\theta}_n$ *is consistent and minimizes* $J_n(\theta)$.

Then

$$|\hat{\theta}_n - \theta_0| = O_P(1/\sqrt{n}).$$

PROOF.

From assumption five

$$J_n(\hat{\theta}_n) - J_n(\theta_0) \leq 0.$$

Assumption one then implies

$$(\hat{\theta}_n - \theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} + 1/2(\hat{\theta}_n - \theta_0)'V_{n0}(\hat{\theta}_n - \theta_0) + o_p(|\hat{\theta}_n - \theta_0|^2) \leq 0.$$

Multiply both sides by $n(1 + \sqrt{n}|\hat{\theta}_n - \theta_0|)^{-2}$. Then defining

$$c_n(\vartheta) = (1 + \sqrt{n}|\theta - \theta_0|)^{-1}\sqrt{n}(\theta - \theta_0),$$

$$c_n'(\hat{\theta}_n)\sqrt{n}\frac{\partial J_n(\theta_0)}{\partial\theta}(1 + \sqrt{n}|\hat{\theta}_n - \theta_0|)^{-1} + 1/2c_n'(\hat{\theta}_n)V_0c_n(\hat{\theta}_n) + o_p(1) \leq 0.$$

If $\sqrt{n}|\hat{\theta}_n - \theta_0| \to \infty$, then the inequality implies $c_n'(\theta_n)V_0c_n(\theta_n) \leq o_p(1)$. Because V_0 is positive definite $c_n(\hat{\theta}_n) = \vec{o}_p(1)$, or $\sqrt{n}|\hat{\theta}_n - \theta_0| = o_p(1)$. This is a contradiction.

∎

We next examine the sufficient conditions for asymptotic normality.

THEOREM **6.2 (Asymptotic Normality)** *Suppose*

1.

$$J_n(\theta) = J_n(\theta_0) + (\theta-\theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} + 1/2(\theta-\theta_0)'V_{n0}(\theta-\theta_0) + o_p(|\theta-\theta_0|^2),$$

2.

$$\sqrt{n}\frac{\partial J_n(\theta_0)}{\partial\theta}$$

has the asymptotic distribution $N(\vec{0}, \Omega_0)$,

3. $|V_{n0} - V_0| = o_p(1)$ *for some* V_0,
4. V_0 *is positive definite,*
5. $\hat{\theta}_n$ *is consistent and minimizes* $J_n(\theta)$.

Then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *has asymptotic distribution* $N(\vec{0}, V_0^{-1}\Omega_0(V_0')^{-1})$.

PROOF.

These assumptions imply those of the Lemma above and hence $\sqrt{n}|\hat{\theta}_n - \theta_0| = O_P(1)$.

Evaluate $J_n(\theta)$ at $\hat{\theta}_n$, then

$$J_n(\hat{\theta}_n) = J_n(\theta_0) + (\hat{\theta}_n-\theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} + 1/2(\hat{\theta}_n-\theta_0)'V_{n0}(\hat{\theta}_n-\theta_0) + o_p(|\hat{\theta}_n-\theta_0|^2).$$

Rewriting we have

$$\begin{aligned} J_n(\hat{\theta}_n) &= 1/2[(\hat{\theta}_n - \theta) + V_0^{-1}\frac{\partial J_n(\theta_0)}{\partial\theta}]'V_0[(\hat{\theta}_n - \theta) + V_0^{-1}\frac{\partial J_n(\theta_0)}{\partial\theta}] \\ &+ J_n(\theta_0) - 1/2\frac{\partial J_n(\theta_0)}{\partial\theta'}V_0^{-1}\frac{\partial J_n(\theta_0)}{\partial\theta} + o_p(n^{-1}) \leq 0. \end{aligned}$$

Next evaluate $J_n(\theta)$ at $\tilde{\theta}_n = \theta_0 - V_0^{-1}\partial J_n(\theta_0)/\partial\theta$, then[8]

$$J_n(\tilde{\theta}_n) = J_n(\theta_0) - 1/2\frac{\partial J_n(\theta_0)}{\partial\theta'}V_0^{-1}\frac{\partial J_n(\theta_0)}{\partial\theta} + o_p(n^{-1}).$$

Since $\hat{\theta}_n$ minimizes $J_n(\theta)$,

$$J_n(\hat{\theta}_n) \le J_n(\tilde{\theta}_n).$$

Therefore

$$[(\hat{\theta}_n - \theta) + V_0^{-1}\frac{\partial J_n(\theta_0)}{\partial\theta}]'V_0[(\hat{\theta}_n - \theta) + V_0^{-1}\frac{\partial J_n(\theta_0)}{\partial\theta}] \le o_p(n^{-1}).$$

Multiply both sides by n. Since V_0 is positive definite,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_0^{-1}\sqrt{n}\frac{\partial J_n(\theta_0)}{\partial\theta} + o_p(1).$$

Finally assumption two implies the result.

∎

To apply these theorems to the estimator under consideration we verify the sufficient conditions which guarantee the five assumptions made in Theorem 6.2.

Assuming that the objective function is twice continuously differentiable the following expantion is valid:[9]

LEMMA **6.1** *Suppose $J_n(\theta)$ is twice continuously differentiable. Then for θ near θ_0, $J_n(\theta)$ equals*

$$J_n(\theta_0) + (\theta - \theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} + 1/2(\theta - \theta_0)'\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'}(\theta - \theta_0) + o_p(|\theta - \theta_0|^2).$$

PROOF.

Define

$$\begin{aligned}\psi_n^+(\theta) &= J_n(\theta) - J_n(\theta_0) - (\theta - \theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} \\ &- 1/2(\theta - \theta_0)'\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'}(\theta - \theta_0) + 1/2(\theta - \theta_0)'(\theta - \theta_0)\delta,\end{aligned}$$

[8] In fact, $\hat{\theta}_n$ and $\tilde{\theta}_n$ have the same asymptotic distribution.

[9] See Hardy (1952) pp.289-290.

for $\delta > 0$. Then

$$\psi_n^+(\theta_0) = 0, \qquad \frac{\partial \psi_n^+(\theta_0)}{\partial \theta} = 0, \text{ and } \frac{\partial^2 \psi_n^+(\theta_0)}{\partial \theta \partial \theta'} = \delta \cdot I.$$

Therefore θ_0 is the local minimum for ψ_n^+. Thus

$$0 = \psi_n^+(\theta_0) \leq \psi_n^+(\theta),$$

near θ_0. Next define

$$\begin{aligned}\psi_n^-(\theta) &= J_n(\theta) - J_n(\theta_0) - (\theta - \theta_0)' \frac{\partial J_n(\theta_0)}{\partial \theta} \\ &- 1/2(\theta - \theta_0)' \frac{\partial^2 J_n(\theta_0)}{\partial \theta \partial \theta'} (\theta - \theta_0) + 1/2(\theta - \theta_0)'(\theta - \theta_0)\delta,\end{aligned}$$

Then

$$\psi_n^-(\theta_0) = 0, \qquad \frac{\partial \psi_n^-(\theta_0)}{\partial \theta} = 0, \text{ and } \frac{\partial^2 \psi_n^-(\theta_0)}{\partial \theta \partial \theta'} = -\delta \cdot I.$$

Therefore θ_0 is the local maximum for ψ_n^-. Thus

$$\psi_n^-(\theta) \leq \psi_n^-(\theta_0) = 0.$$

These two inequalities imply the result.

∎

Thus assumption one is satisfied. Sufficient conditions to satisfy the rest of the assumptions are verified in Lemmas 6.4 and 6.5, respectively. In turn, these lemmas make use of Lemmas 6.2 and 6.3.

We use the following notations:

$$m_{ij}(\theta) = \frac{\partial w_{ij}(\theta)}{\partial \theta}, \qquad w_{ij}(\theta) = K_{ij}(\theta)/\sum_{l \neq i} K_{il}(\theta), \qquad K_{ij}(\theta) = K[(x_i'\theta - x_j'\theta)/a_n],$$

$$\hat{\varphi}_i = \sum_{j \neq i} \varphi(x_j'\theta_0) w_{ij}(\theta_0), \qquad \hat{\varphi}_i' = \sum_{j \neq i} \varphi(x_j'\theta_0) m_{ij}(\theta_0),$$

$$\hat{f}_n(x_i'\theta_0) = \hat{f}_{ni} = (n-1)^{-1} \sum_{j \neq i} a_n^{-1} K_{ij}(\theta_0),$$

and

$$z_i = [x_i - \mathrm{E}(x_j | x_j'\theta_0 = x_i'\theta_0)] \varphi'(x_i'\theta_0).$$

LEMMA 6.2 *Let K be a symmetric, bounded density function on $[-c, c]$. Assume that φ is twice continuously differentiable and the second derivative satisfies a Lipschitz condition. Futhermore assume that*

$$\mathrm{E}[(\varphi_j - \varphi_i)^2 K_{ij}(\theta_0)^2 / f^2(x_i'\theta_0)] < \infty,$$

where f is the density function of $x'\theta_0$. If $x_1, x_2, \ldots, x_n$ is a random sample, then

$$\mathrm{E}[(\hat{\varphi}_i - \varphi_i)^2] \leq O(a_n/n) + O(a_n^4).$$

PROOF.

Since $\sum_{j \neq i} w_{ij}(\theta_0) = 1$,

$$\begin{aligned}\sum_{j \neq i} \varphi_j w_{ij}(\theta_0) - \varphi_i &= \sum_{j \neq i} [(\varphi_j - \varphi_i) w_{ij}(\theta_0)] \\ &= [(n-1)a_n]^{-1} \sum_{j \neq i} [(\varphi_j - \varphi_i) K_{ij}(\theta_0)] / \hat{f}_{ni}.\end{aligned}$$

Hence

$$\mathrm{E}\Big[\Big(\sum_{j \neq i} \varphi_j w_{ij} - \varphi_i\Big)^2\Big] = [(n-1)a_n]^{-2} \mathrm{E}\Big\{\Big[\sum_{j \neq i} [(\varphi_j - \varphi_i) K_{ij}(\theta_0)] / \hat{f}_{ni}\Big]^2\Big\}.$$

In view of Lemma 5.2,

$$\sup_t |\hat{f}_n(t) - f(t)| \to 0,$$

as n goes to infinity. Therefore the right-hand side of the above equation has the same order with

$$[(n-1)a_n]^{-2} \mathrm{E}\Big\{\Big[\sum_{j \neq i} [(\varphi(x_j'\theta_0) - \varphi(x_i'\theta_0)) K_{ij}(\theta_0)] / f(x_i'\theta_0)\Big]^2\Big\},$$

which in tern has the order of $O(a_n/n) + O(a_n^4)$.

∎

Next we prove the mean square convergence of a derivative of a kernel regression estimator.

LEMMA 6.3 *Let K be a symmetric, differentiable, and bounded density function on $[-c, c]$, which satisfies $\int K'(s)ds = 0$, $\int sK'(s)ds = -1$, and $\int s^2K'(s)ds = 0$. Assume that φ and z_i are three times continuously differentiable and the third derivative satisfies a Lipschitz condition. Futhermore assume that*

$$\mathrm{E}[\varphi_i^2 x_i x_i' K_{ik}'^2 K_{ij}^2 / f^4(x_i'\theta_0)] < \infty,$$

where f is the density function of $x'\theta_0$. If $x_1, x_2, \ldots, x_n$ is a random sample, a_n goes to zero, and na_n^3 goes to infinity, then

$$\mathrm{E}[(\hat{\varphi}_i' - z_i)(\hat{\varphi}_i' - z_i)'] = o(1).$$

PROOF.

Since

$$\begin{aligned} & (n-1)^2 a_n^3 \Big(\sum_{j \neq i} \varphi_j m_{ij} - z_i\Big) \\ = & \sum_{j \neq i} \sum_{k \neq i,j} [\varphi_j (x_i - x_j) K_{ij}' K_{ik} - \varphi_j (x_i - x_k) K_{ij} K_{ik}' - a_n z_i K_{ij} K_{ik}] / \hat{f}_{ni}^2 \\ + & \sum_{j \neq i} a_n K_{ij}^2 / \hat{f}_{ni}^2, \end{aligned}$$

it is sufficient to prove that the mean squares of the last two terms have at most order $o(n^4 a_n^6)$, respectively.

Rewriting the second term

$$\begin{aligned} & \sum_{j \neq i} a_n K_{ij}^2 / \hat{f}_{ni}^2 \\ = & \; a_n \sum_{j \neq i} [K_{ij}^2 - \mathrm{E}(K_{ij}^2 | x_i)] / \hat{f}_{ni}^2 + (n-1) a_n^2 \mathrm{E}(a_n^{-1} K_{ij}^2 | x_i) / \hat{f}_{ni}^2. \end{aligned}$$

The mean square of the first term is of order na_n^3 and the mean square of the second term is of order $n^2 a_n^4$. Thus the mean square of $(n-1)^{-2} a_n^{-3} \sum_{j \neq i} a_n K_{ij}^2 / \hat{f}_{ni}^2$ is of order $O(n^{-3} a_n^{-3}) + O(n^{-1} a_n^{-1})$.

We turn to the first term,

$$\sum_{j \neq i} \sum_{k \neq i,j} [\varphi_j (x_i - x_j) K_{ij}' K_{ik} - \varphi_j (x_i - x_k) K_{ij} K_{ik}' - a_n z_i K_{ij} K_{ik}] / \hat{f}_{ni}^2.$$

By the same argument as in Lemma 6.2, the mean is of order $O(a_n^2)$. The variance is of order $O(n^3 a_n^3)$.

∎

Although we assume i.i.d. sampling, $J_n(\theta)$ is not a sum of independent random functions because we estimate $\mathrm{E}(y|x'\theta)$. Therefore we cannot apply a central limit theorem directly to

$$\sqrt{n}\frac{\partial J_n(\theta_0)}{\partial \theta}.$$

Nevertheless Lemma 6.4 holds.

LEMMA **6.4** *Under the assumptions of Lemma 6.2 and 6.3,*

$$\sqrt{n}\frac{\partial J_n(\theta_0)}{\partial \theta}$$

has the asymptotic distribution of $N(0, 4\Omega_0)$, *where*

$$\Omega_0 = \mathrm{E}\{\epsilon_i^2(\varphi_i')^2[x_i - \mathrm{E}(x_j|x_j'\theta_0 = x_i'\theta_0)][x_i - \mathrm{E}(x_j|x_j'\theta_0 = x_i'\theta_0)]'\}$$

PROOF.

Since

$$\begin{aligned}
\sqrt{n}\frac{\partial J_n(\theta_0)}{\partial \theta} &= -\frac{2}{\sqrt{n}}\sum_{i=1}^{n}(y_i - \hat{\mathrm{E}}_i(\theta_0))\hat{z}_i \\
&= -\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i z_i \qquad (1)\\
&- \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\hat{z}_i - z_i) \qquad (2)\\
&- \frac{2}{\sqrt{n}}\sum_{i=1}^{n}(\mathrm{E}_i(\theta_0) - \hat{\mathrm{E}}_i(\theta_0))\hat{z}_i. \qquad (3)
\end{aligned}$$

it is sufficient to prove that (2) and (3) are both $o_p(1)$. Note that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(\hat{z}_i - z_i)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j\neq i}\epsilon_i\epsilon_j m_{ij} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\Big[\sum_{j\neq i}\varphi_j m_{ij} - z_i\Big].$$

To prove that (2) is $o_p(1)$ we show that the two terms on the right hand side are both $o_p(1)$. Since the means of two terms are both zero, it suffices to show that the variances converge to zero. The variance of the first term has the order

$$nE(\epsilon_1^2\epsilon_2^2 m_{12}^2),$$

which in turn is of order

$$(na_n^2)^{-1}E[\epsilon_1^2\epsilon_2^2 K_{12}^2 \hat{f}_n(x_1'\theta_0)^{-2}].$$

Applying the dominating convergence argument to $\hat{f}_n(x_1'\theta_0)$ as in the proof of Lemma 6.2, the variance has the order

$$(na_n^2)^{-1}E[\epsilon_1^2\epsilon_2^2 K_{12}^2 f(x_1'\theta_0)^{-2}] = O((na_n)^{-1}).$$

Therefore the first term of (2) is $o_p(1)$. The variance of the second term of (2) is

$$E\Big[\epsilon_1^2\Big(\sum_{j\neq 1}\varphi_j m_{1j} - z_1\Big)^2\Big].$$

Because $E(\epsilon_i^2|x_i)$ is bounded by assumption, mean square convergence of $\sum_{j\neq i}\varphi_j m_{ij}$ to z_i, which is verified in Lemma 6.3, implies the convergence of the variance. Hence (2) is $o_p(1)$.

Now we turn to (3). Rewriting (3) we have

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n}(E_i(\theta_0) - \hat{\varphi}_i)\hat{z}_i + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\epsilon_j w_{ij}\hat{z}_i.$$

By Lemma 6.2, 6.3, and by Minkowski's inequality the first term is $o_p(1)$. The second term is $o_p(1)$ by a similar argument as that for the first term in (2).

∎

LEMMA **6.5** *Under the assumptions of Lemma 6.2 and 6.3*

$$\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'}$$

converges in probability to

$$2E\{(\varphi_i')^2[x_i - E(x_j|x_j'\theta_0 = x_i'\theta_0)][x_i - E(x_j|x_j'\theta_0 = x_i'\theta_0)]'\}.$$

PROOF.

Note that
$$\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'} = \frac{2}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i' - \frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{\mathrm{E}}_i(\theta_0))\frac{\partial^2\hat{\mathrm{E}}_i(\theta_0)}{\partial\theta\partial\theta'}.$$
The second term is $o_p(1)$ by the similar argument as in Lemma 6.4. Repeated applications of Lemma 6.3 and Minkowski's inequality prove that
$$\frac{2}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i' - \frac{2}{n}\sum_{i=1}^{n}z_i z_i'$$
is $o_p(1)$. Finally an application of law of large number proves that
$$\frac{2}{n}\sum_{i=1}^{n}z_i z_i'$$
converges to the desired limit.

∎

Note that the covariance matrix $V_0^{-1}\Omega_0 V_0^{-1}$ is similar to the covariance matrix of the conventional nonlinear LS estimator with heteroscedasticity, but that the covariance matrices differ because the current approach only looks at the variation along the known function.

Thus one can identify two sources of inefficiency in this approach. First one arises from not adjusting for the heteroscedasticity. We may be able to improve the efficiency by using the present estimator as the first step estimator and correcting for heteroscedasticity in the second step. The second source of inefficiency seems to be intrinsic to the semiparametric approach. It comes from the fact that we are only able to look at the variation along a known function $x'\theta$. If we are to look at the total variation we would need the knowledge of φ.

7 Estimation of the Covariance Matrix.

In order to perform hypothesis tests and construct confidence interval we need a consistent estimator of the covariance matrix. The following Lemma is useful.

LEMMA 7.1 *In addition to the assumptions in Lemma 5.2, if K' is bounded and na_n^3 goes to infinity, then*

$$(na_n^2)^{-1}\sum_{j=1}^{n} y_j(x_i - x_j)K'[x_i'\theta - x_j'\theta/a_n] - \mathrm{E}\{a_n^{-2}y_j(x_i - x_j)K'[x_i'\theta - x_j'\theta/a_n]\}$$

converges uniformly to zero in θ.

PROOF.
Identical to the proof of Lemma 5.2.

∎

Since $\hat{\theta}_n$ converges in probability to θ_0, Lemma 7.1 and 5.2 imply

$$\hat{z}(\hat{\theta}_n) - z(\theta_0) = o_p(1).$$

Therefore

$$\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i(\hat{\theta}_n)\hat{z}_i(\hat{\theta}_n) - V_0 = o_p(1).$$

Next it is straight forward to show that

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mathrm{E}}_i(\hat{\theta}_n))^2\hat{z}_i(\hat{\theta}_n)\hat{z}_i(\hat{\theta}_n)$$

estimates Ω_0.

Thus

$$n\left[\sum_{i=1}^{n}\hat{z}_i(\hat{\theta}_n)\hat{z}_i(\hat{\theta}_n)\right]^{-1}\sum_{i=1}^{n}(y_i - \hat{\mathrm{E}}_i(\hat{\theta}_n))^2\hat{z}_i(\hat{\theta}_n)\hat{z}_i(\hat{\theta}_n)\left[\sum_{i=1}^{n}\hat{z}_i(\hat{\theta}_n)\hat{z}_i(\hat{\theta}_n)\right]^{-1}$$

estimates the covariance matrix. Note that this matrix is guaranteed to be positive semidefinite.

8 Monte Carlo Results.

In this section we look at the small sample properties of the estimator by a Monte Carlo experiment. The construction of the experiment is identical to that of Cosslett's (1986). He considers a binary choice model with two regressors.

$$y_i^* = \alpha_0 + \beta_{10}x_1 + \beta_{20}x_2 + \epsilon_i.$$

As usual, observed indicator y_i takes value 1 if latent variable y_i^* is positive and y_i takes value 0 if y_i^* is nonpositive.

The true parameter values are $\alpha_0 = 0$, $\beta_{10} = -2$, and $\beta_{20} = 1$. In his specification, exogenous variables take two different distributions and the error distributions take three different distributions, giving rise to six different models. The exogenous variables x_1 and x_2 are independently distributed. Two distribution of exogenous variables are:

(A) Standard normal.

(B) Standard exponential.

Three different mixtures of normal distributions are considered for the error distributions.

(1) Standard normal.

(2) $0.75 \cdot N(0,1) + 0.25 \cdot N(0,25)$

(3) $0.75 \cdot N(-0.5,1) + 0.25 \cdot N(1.5,25)$

According to Cosslett's calculation the second distribution has standard error 2.65, skewness 0, and kurtosis 6.61. Similarly the third distribution has standard error 2.78, skewness 1.29, and kurtosis 6.29.

We take his models in order to facilitate the comparison of the small sample performance of the estimator with others presented in his paper. He presents the results for Maximum Score Estimator (MSE), Maximum Rank Correlation Estimator (MRC), Semiparametric Maximum Likelihood Estimator (SML), and its smoothed version (SML-1) of it along with the conventional Probit MLE. Note that none of these semiparametric estimators has known asymptotic distribution.

Our results are not directly comparable with his beacause of two reasons. The first reason is that random number generators used are different. Second, the optimization methods employed are different.

Cosslett optimizes by a grid search method. Initial grid search between -2.5 and -1.5 is done and when the objective function is still improving at the boundaries further search was performed.

This particular method of optimization might have a risk of choosing the values closer to the truth more often than we really would when we did not know the truth. In our experiment we employ a different grid search method which treats different parameter values identically.

The first stage grid search is done between -50 and 50 with grid width 1. Pick seven values which performed best. Then move to the second stage. This time the grid search is done around the selected seven values with the grid width of 0.1. Pick five values which performed best. Then move to the third stage. This time the grid search is done around the selected five values with the grid width of 0.01. The final stage is performed around the five selected values with grid width 0.001. The FORTRAN code is in Appendix 2. IBM 3090-600E at Cornell is used for the computation. Each calculation took about 1.8 seconds of cpu. Computational speed increases with roughly square of the sample size using current algorithm. The results are presented in Table 8.1 and in Table 8.2. The results for other estimators are from Cosslett (1986).

Ruud (1983) showed that ML estimators are consistent in (A). Furthermore, obviously Probit estimator in case (B-1) is consistent. Therefore cases (B-2) and (B-3) are the only cases where Probit estimator is not consistent. In those two cases the SLS performs the best in terms of the estimated mean square error.

The MSEs lie between 0.45 and 0.70. Compared with other estimators MSEs are not affected so much by the differences in distributions either in the error term or in the regressors.

9 Concluding Remarks.

This thesis studied a semiparametric estimation of single index models. We established $\sqrt{n}$-consistency and asymptotic normality of the semiparametric least squares estimator. Consistent estimator of the covariance matrix is also given.

While the results extend the applicability of semiparametric estimation, there are a number of related issues I did not address in this thesis. First of all, I did not consider estimation of the unknown function φ explicitly. In each specific single index model there exists an infinite dimensional problem corresponding to the unknown function φ. In binary choice model this is particularly interesting for prediction purposes. In duration models if we are willing to postulate the proportional hazard specification, then it may be of interest to estimate the base line hazard.

Since

$$\mathrm{E}(y|h(x;\theta_0)) = \varphi(h(x;\theta_0)),$$

$\hat{\mathrm{E}}_n(\hat{\theta}_n)$ is a natural estimator for φ. We established uniform convergence of $\hat{\mathrm{E}}_n(\theta)$ to $\mathrm{E}(y|h(x;\theta))$ and convergence of $\hat{\theta}_n$ to θ_0 in probability. These are sufficient to imply that $\hat{\mathrm{E}}_n(\hat{\theta}_n)$ converges in probability to $\varphi(h(x;\theta_0))$. The asymptotic distribution is not yet known.

Also, I did not consider an extention of the estimation method to multiple index models. The extention will be a useful one, for the extention naturally includes multinomial choice models. At the same time, the extention is not a trivial one, because a straight forward extention of SLS involves nonparametric estimation of more than one dimension function.

Furthermore, I have not studied how we should choose the band width or a kernel function. At this point we have a spectrum of estimators corresponding to different choices of band width or kernel functions, and we have no way of choosing among them. As we showed, the choice does not affect the asymptotic distribution and hence the decision should be based on small sample properties, such as mean square error.

There is another approach to the basic problem studied in this thesis. The basic problem was that forms of the error terms are too often casually assumed without any justification. Rather than asking what the error terms are and how they might be distributed in specific context, we proposed an estimator which does not require knowing the error distribution at all, namely we took a semiparametric approach. We could have faced the problem directly and tried to model errors. Specifically, we could have tried to derive the error distribution within a specific model based on the uniform distribution, rather than casually assume it. The alternative approach produced the Gaussian distribution, the exponential distribution, Wiener process, or the Poisson process in other disciplines.

Although two approaches are very different in attitude toward the error terms, ultimately they should be complements. For the first approach offers a way of testing the assumptions behind the derived distribution with an alternative set of estimates ready when the specification is rejected.

These are a few of the problems which await our future research.

APPENDIX 1

As many other nonparametric estimators, a kernel estimator is also a local averaging estimator. A natural estimator for a conditional expectation $\mathrm{E}(y \mid x = t)$ is to take an average of those y_is whose corresponding x_is are close to t, say between $t - a_n$ and $t + a_n$, where $a_n > 0$. The estimator

$$\frac{\sum_{j=1}^{n} y_j \cdot 1(|(t - x_j)/a_n| < 1)}{\sum_{l=1}^{n} 1(|(t - x_l)/a_n| < 1)}$$

results. To estimate the conditional expectation at $x = t$ consistently we have to shrink a_n, usually referred to as the band width, to zero. But we cannot set $a_n = 0$ because then there is no sample to take an average over. Therefore we must shrink a_n to zero as the sample size gets large while insuring that a positive fraction of the whole sample falls within distance a_n from t.

The kernel estimator generalizes this natural estimator. Instead of using zero-or-one weights it uses a weight function which is smooth in t and in x. Restricting K to be a density function does not bind once deciding to use nonnegative weights because the denominator and the numerator are homogeneous of the same degree in K. Parzen (1962) first proposed kernel estimators for the estimation of density functions. Subsequently Nadaraja (1964) and Watson (1964) applied the method to the estimation of the conditional expectations. A useful survey is available in Bierens (1985).

APPENDIX 2

```
integer n, k
parameter(n=250, k=2)
real expdev, rmix2
integer i, ii, iii, m, ind, idum , y(n)
double precision x(n,k), b0(k), sumy, ymean, f(100),
    1     b(100), beta(1000)
double precision bb(70), f1(70), bbb(50), f2(50), bbbb(50),
    1     f3(50),sum
double precision bias
b0(1)=dble(-2.0)
b0(2)=dble(1.0)
idum=-68302619
do 1000 m=1,1000
c
c   generate the data
c
do 5 i=1,n
x(i,1)=dble(expdev(idum))
x(i,2)=dble(expdev(idum))
y(i)=ind(x(i,1)*b0(1)+x(i,2)+dble(rmix2(idum)))
5 continue
c
c   calculate the mean of y
c
sumy=dble(0.)
do 6 i=1,n
sumy=sumy+dble(y(i))
6 continue
ymean=sumy/dble(n)
b(1)=dble(-50.)
do 7 i=2,100
b(i)=b(i-1)+dble(1.)
7 continue
do 8 i=1,100
call J(b(i),y,x,ymean,f(i))
8 continue
call sort(100,f,b)
do 11 i=1,7
do 10 ii=1,10
   iii=ii+10*(i-1)
   bb(iii)=b(i)-dble(.5)+dble(.1)*dble(ii)
10 continue
11 continue
do 12 i=1,70
   call J(bb(i),y,x,ymean,f1(i))
12 continue
```

```
call sort(70,f1,bb)
do 14 i=1,5
do 13 ii=1,10
   iii=ii+10*(i-1)
   bbb(iii)=bb(i)-dble(.05)+dble(.01)*dble(ii)
13 continue
14 continue
do 15 i=1,50
   call J(bbb(i),y,x,ymean,f2(i))
15 continue
call sort(50,f2,bbb)
do 17 i=1,5
do 16 ii=1,10
   iii=ii+10*(i-1)
   bbbb(iii)=bbb(i)-dble(.005)+dble(.001)*dble(ii)
16 continue
17 continue
do 18 i=1,50
   call J(bbbb(i),y,x,ymean,f3(i))
18 continue
call sort(50,f3,bbbb)
beta(m)=bbbb(1)
print *, 'this is the',m,'th iteration'
print *, 'the estimate is', bbbb(1)
1000 continue
call sort(1000,beta,beta)
do 1010 i=1,9
print *, beta(100*i)
1010 continue
sum=dble(0.)
do 1020 i=1,1000
sum=sum+beta(i)
1020 continue
bias=sum/dble(1000)
print *, 'bias is', bias+2
sum=dble(0.)
do 1030 i=1,1000
sum=sum+(beta(i)-bias)**2
1030 continue
print *, 'variance is', sum/dble(1000)
end
c
c subroutine to calculate the objective function
c
subroutine J(b,y,x,ymean,f)
integer n, k, index, i, ii
parameter(n=250, k=2)
integer y(n)
```

```
double precision b, x(n,k), w, ymean, f
double precision kern(n,n), reg(n), sum1, sum2, sumk
c
c   set kern(i,i)=0.
c
do 2 i=1,n
kern(i,i)=dble(0.)
2 continue
w=dble(n)**dble(-0.297)
c
c   calculate kern here
c
do 110 i=1,n-1
do 100 ii=i+1,n
kern(i,ii)=((min(abs((x(i,1)-x(ii,1))*b+x(i,2)x(ii,2)),w)/w)
     1            **2-dble(1.))**2
100 continue
110 continue
c
c   estimate regression line
c
do 160 i=1,n
sum1=dble(0.)
sum2=dble(0.)
   do 120 index=1,i
   sum1=sum1+kern(index,i)
120    continue
   do 130 index=i,n
   sum2=sum2+kern(i,index)
130    continue
sumk=sum1+sum2
if(sumk.eq.dble(0.))then
   reg(i)=ymean
else
sum1=dble(0.)
sum2=dble(0.)
   do 140 index=1,i
   sum1=sum1+dble(y(index))*kern(index,i)
140    continue
   do 150 index=i,n
   sum2=sum2+dble(y(index))*kern(i,index)
150    continue
reg(i)=(sum1+sum2)/sumk
endif
160 continue
c
c   now calculate sum of squares
c
```

```
f=dble(0.)
do 170 i=1,n
f=f+(dble(y(i))-reg(i))**2
170 continue
return
end
c
c uniform random number generator of Knuth
c         as implemented by PFTV
c
function ran3(idum)

parameter (mbig=1000000000., mseed=1618033., mz=0.,
    1 fac=1./mbig)
dimension ma(55)
data iff /0/
if(idum.lt.0 .or. iff .eq. 0) then
iff=1
mj=mseed-iabs(idum)
mj=mod(mj,mbig)
ma(55)=mj
mk=1
do 10 i=1,54
   ii=mod(21*i,55)
   ma(ii)=mk
   mk=mj-mk
   if(mk .lt. mz) mk=mk+mbig
   mj=ma(ii)
10 continue
do 30 k=1,4
   do 20 i=1,55
ma(i)=ma(i)-ma(1+mod(i+30,55))
if(ma(i) .lt. mz) ma(i)=ma(i)+mbig
20    continue
30 continue
inext=0
inextp=31
idum=1
endif
inext=inext+1
if(inext .eq. 56) inext=1
inextp=inextp+1
if(inextp .eq. 56) inextp=1
mj=ma(inext)-ma(inextp)
if(mj .lt. mz) mj=mj+mbig
ma(inext)=mj
ran3=mj*fac
return
```

```
c
c mixture of normal random number generator using ran3
c    w/ 0.75 --- N(-0.5,1), w/ 0.25 --- N(1.5,25)
c
function rmix2(idum)

if(ran3(idum).lt.0.75)then
rmix2=gasdev(idum)-0.5
else
rmix2=5.*gasdev(idev)+1.5
endif
return
end

c
c indicator function: 1 if arg>0, 0 otherwise.
c
integer function ind(arg)

double precision arg
if (arg .gt. dble(.0)) then
   ind=1
        else
           ind=0
endif
return
end
c
c       subroutine sort
c
c
subroutine sort(n,ra,rb)
double precision ra(n), rb(n), rra, rrb
integer l, n, ir, i, j

l=n/2+1
ir=n
10 continue
   if(l.gt.1)then
l=l-1
rra=ra(l)
rrb=rb(l)
   else
rra=ra(ir)
rrb=rb(ir)
ra(ir)=ra(1)
rb(ir)=rb(1)
ir=ir-1
```

```
if(ir.eq.1)then
   ra(1)=rra
   rb(1)=rrb
   return
endif
   endif
   i=l
   j=l+l
20    if(j.le.ir)then
if(j.lt.ir)then
   if(ra(j).lt.ra(j+1))j=j+1
endif
if(rra.lt.ra(j))then
   ra(i)=ra(j)
   rb(i)=rb(j)
   i=j
   j=j+j
else
   j=ir+1
endif
   go to 20
   endif
   ra(i)=rra
   rb(i)=rrb
go to 10
end
```

Table 1: x_1 and x_2 normal

	Error (1)		Error (2)		Error (3)	
Estimator	Bias	RMSE	Bias	RMSE	Bias	RMSE
SLS	0.077	0.45	0.174	0.56	0.178	0.56
Probit	-0.04	0.29	-0.11	0.49	-0.11	0.50
MS	-0.22	0.76	-0.34	1.16	-0.36	1.27
MRC	-0.05	0.34	-0.11	0.49	-0.11	0.52
SML	-0.08	0.43	-0.20	0.67	-0.20	0.70
SML-1	-0.05	0.31	-0.11	0.48	-0.10	0.47

Table 2: x_1 and x_2 exponential

	Error (1)		Error (2)		Error (3)	
Estimator	Bias	RMSE	Bias	RMSE	Bias	RMSE
SLS	0.187	0.53	0.288	0.70	0.259	0.69
Probit	-0.03	0.35	-0.23	0.72	-0.69	1.24
MS	-0.37	1.29	-0.51	1.87	-0.55	1.64
MRC	-0.05	0.43	-0.13	0.71	-0.27	1.32
SML	-0.10	0.53	-0.23	0.84	-0.29	1.01
SML-1	-0.06	0.39	-0.23	0.73	-0.43	1.38