Journal of Econometrics 58 (1993) 71–120. North-Holland

Semiparametric least squares (SLS) and weighted SLS estimation of single-index models

Hidehiko Ichimura*

University of Minnesota, Minneapolis, MN 55455, USA

For the class of single-index models, I construct a semiparametric estimator of coefficients up to a multiplicative constant that exhibits $1/\sqrt{n}$-consistency and asymptotic normality. This class of models includes censored and truncated Tobit models, binary choice models, and duration models with unobserved individual heterogeneity and random censoring. I also investigate a weighting scheme that achieves the semiparametric efficiency bound.

1. Introduction

In this paper I define a new semiparametric least squares (SLS) estimator for the single-index model, establish its $1/\sqrt{n}$-consistency and asymptotic normality, where n here refers to the sample size, and present a consistent estimator of the covariance matrix. I also study a weighted SLS (WSLS) estimator and show that it achieves the semiparametric efficiency bound obtained by Newey (1990) for the single-index model.[1]

Correspondence to: Hidehiko Ichimura, Department of Economics, University of Minnesota, 271 19th Avenue South, Minneapolis, MN 55455, USA.

*I thank Professors Daniel McFadden and James Powell for advice and Professor Charles Manski for encouragement. I have benefited from many comments I received from two anonymous referees, Professors Jonanthan Feinstein, Robert Gertner, Peter Hall, Lung-Fei Lee, Daniel McFadden, and Kazumitsu Nawata. The points I owe to them are indicated in the paper. I also received extensive editorial suggestions from Dr. Steven Herzenberg on a very early version and professional editorial help from Patricia Haswell on this version. I am grateful to all of these individuals. This paper supersedes three earlier papers, 'Single Index Models and their Consistent Semiparametric Estimation' (October 1985), 'Estimation of Single Index Models' (September 1987) (Ph.D. thesis), and 'Semiparametric Least Squares Estimation of Single Index Models' (August 1988).

[1] The optimal weighting scheme involves knowing the conditional variance of y given x and hence is infeasible. In this paper I do not investigate the use of estimated weights.

We say that a probabilistic model is *semiparametric* if an index parameterizing the distribution consists of two parts, say θ and γ, where θ lies in a finite-dimensional space Θ and γ lies in an infinite-dimensional space Γ.

Let $(\theta_0, \gamma_0) \in \Theta \times \Gamma$ be the true value of the parameter in the model. We call an estimator of θ_0 semiparametric if the model is semiparametric and the definition of the estimator does not involve knowledge of γ_0. We call a model and an estimation method *parametric* if the model species γ_0 and thus restricts the distribution to a finite-dimensional space.

For example, the regression model

$$y = x'\beta_0 + \varepsilon,$$

with $\mathrm{E}(\varepsilon \mid x) = 0$ and independent and identically distributed (i.i.d.) sampling, is semiparametric, for β_0 corresponds to θ_0 and the joint distribution of x and ε corresponds to γ_0. An example of a semiparametric estimator is the ordinary least squares (OLS) estimator.

The regression model has played a prominent role in econometric analysis. Careful inspections of economic problems, however, have revealed limitations of the regression model and induced efforts to overcome the inadequacies of the basic model.[2] Recognition of the simultaneous equation problem, nonnegativity restrictions, probabilistic choices, disequilibria in markets, the selectivity bias problem, and the time dependence of economic decisions led to the construction of more appropriate econometric models.[3] All of these models were initially specified as parametric models.[4]

For the simultaneous equation model, Theil (1953a, b) and Basmann (1957) developed a semiparametric estimator, the two-stage least squares method. Analogous development has been ongoing for other models since Manski (1975) initiated research in semiparametric estimation for the discrete choice model.

Some of the models mentioned above are single-index models (see section 2). Under some regularity conditions, the SLS estimator for single-index models is consistent with rate $1/\sqrt{n}$, the typical rate achieved by parametric estimators under i.i.d. sampling. The $1/\sqrt{n}$ convergence rate of the SLS estimator implies that the estimator is not infinitely inefficient compared with conventional parametric approaches even though the model is not restricted within a finite-dimensional space, as it is for the parametric maximum likelihood (ML) estimation method. I call the estimator *semiparametric least squares* because the objective function resembles that of the nonlinear least squares (NLS) estimator.

[2] A statistical problem with the least squares estimator, nonrobustness, led Koenker and Basset (1978) to study the least absolute deviation estimator.

[3] Some of these models had more impact on empirical studies than others.

[4] To be precise, they are specified as parametric models up to an ancillary parameter, such as the marginal distribution of exogenous regressors.

In the next section, I define single-index model, and show that it includes censored Tobit models, binary choice models, and duration models, among others. Section 3 gives a geometric motivation for the proposed estimator. In the fourth section, I formally define the SLS estimator, discuss the identification of the parameters of the model, and show that under some regularity conditions, the estimation technique identifies the true parameters up to a multiplicative constant in the linear single-index model. Section 5 provides proofs of consistency and asymptotic normality of the SLS and the WSLS estimators. Section 6 addresses some efficiency issues, and section 7 presents a consistent estimator for the covariance matrix. Section 8 investigates the small-sample properties of the SLS estimator by analyzing a Monte Carlo experiment. The final section discusses some directions for future research.

2. Single-index models

Let L, M, and n be positive integers and let n denote the sample size.

Definition 2.1 (Single-Index Models). *The model*

$$y_i = \varphi(h(x_i;\theta_0)) + \varepsilon_i \quad \textit{for} \quad i = 1, \ldots, n,$$

where

(1) (x_i', y_i) *for* $i = 1, \ldots, n$ *is an i.i.d. sample*;

(2) $y_i \in R$ *and* $x_i \in R^L$ *are observed,* $\varepsilon_i \in R$ *is an unobserved disturbance, and* $\theta_0 \in R^M$ *is an unknown parameter to be estimated*;

(3) $\mathrm{E}(\varepsilon_i \mid x_i) = 0$;

(4) the function $h: X \times \Theta \rightarrow R$ *for some* $X \times \Theta \subset R^L \times R^M$ *is known up to a parameter* θ; *and*

(5) the function $\varphi: R \rightarrow R$ *is not known*;

is a single-index model.

The single-index model is semiparametric for two reasons: first, the function $\varphi: R \rightarrow R$ is not known, and second, the conditional probability of ε conditioned on x is not specified except for $\mathrm{E}(\varepsilon \mid x) = 0$. If the function φ is known, then it is well-known that the least squares method consistently estimates θ_0 under some general conditions. Thus the lack of knowledge of φ makes the estimation of θ_0 nontrivial.

Brillinger (1983) first proposed studying limited dependent variable models within this framework. He calls this model a generalized linear model, for he considers a case where $h(x;\theta) = x'\theta$. Since a class of models with the same name already exists in a different context [for example, see McCullagh and Nelder (1983)], 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^* but y, which is a transformation of y^*. Formally,

$$y^* = h(x;\theta_0) + v, \qquad y = \tau(y^*).$$

We assume that x and v are independent[5] and that the sampling of (x', y) is i.i.d. Furthermore, we assume that $x \in R^L$ and $\theta_0 \in R^M$.

The function $\tau: R \to R$ may or may not be known. If the function $\tau: R \to R$ takes the form

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

then the censored Tobit model results. If the function $\tau: R \to R$ takes the form

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

then the binary choice model results. In both cases the transformation function τ is known but $\mathrm{E}(y\,|\,x)$ is not known unless the distribution of the error term is specified. Knowing the transformation function τ alone does not allow least squares estimation.

The reduced form has the form of a single-index model, as the following calculation shows:

$$\mathrm{E}(y\,|\,x) = \int_{-\infty}^{+\infty} \tau(h(x;\theta_0) + v)\,\mathrm{d}F_v,$$

where F_v is the distribution function of the random variable v. By defining ε to be the difference between the observable y and the conditional expectation, or

$$\varepsilon = y - \mathrm{E}(y\,|\,x),$$

[5] In general, the distribution of v can depend on $h(x;\theta_0)$.

we can transform this class of latent variable models into the single-index model defined in Definition 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 also be regarded as an alternative to the errors-in-variable formulation of regression models. For example,[6] suppose y^* is unobserved true profit of a firm and we observe only reported profit y. We could assume that reported profit is true profit plus an error term. This is the errors-in-variable formulation. An alternative is to use the model $y = \tau(y^*)$. With this modeling strategy, we allow reported profit to differ systematically from true profit and yet we are able to consistently estimate the relative determinants of firm profitability.

Another model that can be regarded as a single-index model is the truncated Tobit model. Here, unlike in censored Tobit models, we do not observe x when the corresponding dependent variable y is censored. Let $I(A)$ be a function that takes value 1 when a logical statement A is true and value 0 when A is false. When the logical statement is indexed by i and no confusion ensues, $I(A_i)$ will be abbreviated as I_i. Then, truncated Tobit models can be written as

$$y = h(x;\theta_0) + u,$$

where u has density

$$I(u > -h(x;\theta_0))F_v(\mathrm{d}u)[1 - F_v(-h(x;\theta_0))]^{-1}.$$

Therefore, $\mathrm{E}(y \mid x)$ equals

$$h(x;\theta_0) + \int_{-h(x;\theta_0)}^{\infty} uF_v(\mathrm{d}u)[1 - F_v(-h(x;\theta_0))]^{-1},$$

and has the form $\varphi(h(x;\theta_0))$.

The class of single-index model also includes duration models as a special case if (1) censoring is random, (2) exogenous variables are time-independent, and (3) individual heterogeneity is independent of exogenous variables. To see this, let the conditional density function of a duration spell t and the unobserved heterogeneity α given x without censoring be

$$f(t,\alpha) = f_1(t;h(x;\theta_0))f_2(\alpha),$$

[6] I thank Professor Robert Gertner for suggesting this example.

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

$$f_C(t) - f_C(t)\int_{-\infty}^{t} f_1(s, h(x;\theta_0))\,\mathrm{d}s + [1 - F_C(t)]f_1(t, h(x;\theta_0)),$$

where F_C denotes the distribution function of the censoring point. Therefore the conditional expectation of the duration spell conditioned on the regressors has the form $\varphi(h(x;\theta_0))$ as desired.

Because single-index models abstract from the specific structures of each model, they do not take advantage of particular restrictions within each model other than those already exploited by the formulation of the single-index model. For binary choice models, the unknown function φ is the cumulative distribution function. Thus it is a nondecreasing function with range between 0 and 1. For censored Tobit models,

$$\varphi(h(x;\theta_0)) = \int_{-\infty}^{h(x;\theta_0)} F_v(s)\,\mathrm{d}s,$$

and thus φ is a nondecreasing function that is an integral of the cumulative distribution function of the error term. Nevertheless, in the single-index framework, both functions are specified only as measurable functions with some finite moments. We shall see that the abstraction results in a loss in identification for some models in the single-index class. Typically, Tobit models allow identification of all parameters, and binary choice models allow identification of parameters up to a multiplicative positive constant. In both cases we shall see that in the linear single-index model formulation, slope coefficients are identified only up to a multiplicative constant that is not necessarily positive. However, even in those cases analysis often simplifies considerably once unknown coefficients are estimated up to a scalar. SLS provides a convenient initial estimator in those cases. In other more complicated cases, such as the unknown transformation model or the duration model with unobserved heterogeneity, the single-index approach provides a convenient framework in which to consider estimation.

3. Geometric motivation for the SLS estimation method

Recall the definition of the single-index model:

$$y_i = \varphi(h(x_i;\theta_0)) + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n.$$

SLS estimation is based on the following three observations:

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

These three observations indicate a way to identify θ_0. The conditional variance,[7]

$$\operatorname{var}[y \mid h(x;\theta)=c] = \mathrm{E}\{[y-\mathrm{E}[y \mid h(x;\theta)=c]]^2 \mid h(x;\theta)=c\},$$

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

$$\mathrm{E}\{W(x)[y-\mathrm{E}[y \mid h(x;\theta)]]^2\}$$

as the objective function, where $W\colon R^L \to R$ is a weighting function, and then find θ that minimizes the objective function.

4. The estimator and the identification conditions

The heuristic argument in the previous section suggests minimizing an objective function

$$J(\theta)=\mathrm{E}\{[y-\mathrm{E}(y \mid h(x;\theta))]^2\}$$

with respect to θ to define an estimator of θ_0. Since $J(\theta)$ is unknown, it must be estimated in order to define a feasible estimator.

If $\mathrm{E}[y \mid h(x;\theta)]$ is known, then a sample analog of $J(\theta)$ is

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

[7] In this paper I use the conditional variance as a measure of variability, but that choice is not unique. One may be able to use other measures, such as absolute deviation, and derive other estimation methods. This possibility is pursued in Hall and Ichimura (1991).

Since $\mathrm{E}[y\,|\,h(x;\theta)]$ is not known, we replace it with a kernel estimator. [For discussion of the kernel estimator, see Parzen (1962), Nadaraja (1964), Watson (1964), and Prakasa Rao (1983).] Thus, we define the SLS estimator as follows:

Definition 4.1 (SLS Estimator). *The SLS estimator minimizes the square root of*

$$J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \in X)\,[y_i - \hat{\mathrm{E}}(x_i,\theta)]^2 + o_p(n^{-1}),$$

where

(1) $X \subset R^L$;

(2) $$\hat{\mathrm{E}}(x_i,\theta) = \frac{\sum_{j\neq i} y_j I(x_j \in X_n) K([h(x_i;\theta) - h(x_j;\theta))]/a_n)}{\sum_{j\neq i} I(x_j \in X_n) K([h(x_i;\theta) - h(x_j;\theta)]/a_n)},$$

if

$$\sum_{j\neq i} I(x_j \in X_n) K([h(x_i;\theta) - h(x_j;\theta)]/a_n) \neq 0,$$

and

$$\begin{aligned} \hat{\mathrm{E}}(x_i;\theta) &= y_{\max} \quad \textit{if} \quad y_i \leq (y_{\max} + y_{\min})/2, \\ &= y_{\min} \quad \textit{otherwise}, \end{aligned}$$

where $X_n = \{x \,|\, \|x - x'\| \leq 2a_n$ *for some* $x' \in X\}$, $y_{\max} = \max_i y_i$, *and* $y_{\min} = \min_i y_i$;

(3) K: $R \to R$ *is a one-dimensional density function; and*

(4) $a_n > 0$ *and* $a_n \to 0$.

When all the denominators of kernel regression estimators are 0, *we assume that* $\hat{\theta} = 0$.

Some remarks on this definition are in order. First, clearly any objective function that is strictly a monotonic transformation of $J_n(\theta)$ defines the SLS estimator. The current version is convenient for proving consistency.

Second, for each finite n, there is a positive probability that $\hat{\theta} = 0$. As I show later, the probability that $\hat{\theta} = 0$ goes to 0 if $na_n^2 \to \infty$ and the density of $h(x;\theta)$ is uniformly bounded and bounded away from 0 on X.

Third, the trimming term $I(x \in X)$ is introduced to guarantee that the density of $h(x;\theta)$ is bounded away from 0 on X. This set X has to be chosen accordingly. The condition helps to establish uniform convergence of $\hat{\mathrm{E}}(x,\theta)$ and its first and second derivatives to their respective limits in probability. These convergence results will be used to establish consistency and asymptotic normality of the estimator.[8]

Fourth, $\hat{\mathrm{E}}(x,\theta)$ is defined as the usual kernel regression estimator using (x_j, y_j) such that $x_j \in X_n$. Because X_n is taken to be a set that includes X in such a way that all boundary points in X are interior to X_n, in a neighbourhood of x, with probability approaching 1, there are data in all directions to take a local average. The construction helps to prove uniform convergence of $\hat{\mathrm{E}}(x,\theta)$. We take a monotonically convergent sequence of sets $\{X_n\}_{n=1}^{\infty}$ rather than a fixed set that includes X in order to reduce bias.[9] This point is elaborated further in section 5.

Fifth, the boundary conditions could be defined differently, but the current version makes the objective function lower-semicontinuous.

Sixth, conceivably one could use any estimator of $\mathrm{E}[y \mid h(x;\theta)]$ in place of a kernel estimator. There are at least two advantages to using a kernel estimator:[10] (a) the objective function is differentiable with probability approaching 1 if a differentiable kernel function is used, and (b) when a kernel estimator is used, a derivative of the objective function converges to a derivative of the limiting function.

Lastly, if φ is known, minimizing

$$\frac{1}{n}\sum_{i=1}^{n} [y_i - \varphi(h(x_i;\theta))]^2$$

yields an estimator that is $1/\sqrt{n}$-consistent and asymptotically normal under some general conditions. This is the NLS approach. One might be tempted to regard SLS as the same as NLS except that φ is replaced by its estimate, say $\hat{\varphi}$. This is not the case. The estimator of θ_0 in SLS involves estimation of $\mathrm{E}[y \mid h(x;\theta)] = \mathrm{E}[\varphi \mid h(x;\theta)]$ but does not involve estimation of φ. In general, the two functions $\varphi(h(x;\theta))$ and $\mathrm{E}[\varphi \mid h(x;\theta)]$ are different.

[8] I thank Professor Lung-Fei Lee for pointing out the importance of the trimming in establishing asymptotic properties of the estimator.

[9] I thank Professor Peter Hall for this idea.

[10] I thank Professor Daniel McFadden for suggesting the use of a kernel estimator.

The single-index model does not preclude heteroskedasticity. To take heteroskedasticity into account, we define the WSLS estimator.

Definition 4.2 (WSLS Estimator). *The WSLS estimator minimizes the square root of*

$$J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \in X)\, W(x_i)[y_i - \hat{\mathrm{E}}_W(x_i, \theta)]^2 + \mathrm{o}_p(n^{-1}),$$

where

(1) $X \subset R^L$;

(2) $$\hat{\mathrm{E}}_W(x_i, \theta) = \frac{\sum_{j \neq i} y_j I(x_j \in X_n)\, W(x_j) K([h(x_i;\theta) - h(x_j;\theta)]/a_n)}{\sum_{j \neq i} I(x_j \in X_n)\, W(x_j) K([h(x_i;\theta) - h(x_j;\theta)]/a_n)},$$

if

$$\sum_{j \neq i} I(x_j \in X_n)\, W(x_j) K([h(x_i;\theta) - h(x_j;\theta)]/a_n) \neq 0,$$

and

$$\hat{\mathrm{E}}_W(x_i, \theta) = \begin{cases} y_{\max} & \textit{if} \;\; y_i \leq (y_{\max} + y_{\min})/2, \\ y_{\min} & \textit{otherwise}, \end{cases}$$

where $X_n = \{x \mid \|x - x'\| \leq 2a_n$ *for some* $x' \in X\}$, $y_{\max} = \max_i y_i$, *and* $y_{\min} = \min_i y_i$;

(3) *weights* $\{W(x_i)\}_{i=1}^{n}$ *satisfy* $0 < W(x_i) < \bar{W}$, for some constant $\bar{W}$, *for some constant* $\bar{W}$;

(4) $K: R \to R$ *is a one-dimensional density function*; *and*

(5) $a_n > 0$ *and* $a_n \to 0$.

We assume that $\hat{\theta} = 0$ *when* $\hat{\mathrm{E}}_W(x_i;\theta) = 0$ *for all* $i = 1, \ldots, n$.

Two remarks on the weighting scheme in this definition are in order. First in the NLS framework, heteroskedasticity is corrected by reweighting. That is, if

$\varphi(h(x_i;\theta))$ is known, then the NLS weighting scheme is

$$\frac{1}{n}\sum_{i=1}^{n} W(x_i)[y_i - \varphi(h(x_i;\theta))]^2.$$

In this model the scheme $W(x_i) = \sigma^{-2}(x_i)$, where $\sigma^2(x_i)$ gives the conditional variance of y_i given x_i, is the optimal weights. Since in the single-index model $\varphi(h(x_i;\theta))$ is not known, we introduce weighting also in the kernel regression estimator (see also section 6).

Second, because I treat $W(x_i)$ as a known function in this paper, an optimal estimator is infeasible. Rather, the point of introducing a weighting scheme is theoretical: it is intended to clarify the roles played by weights that are different from the roles played by weights in NLS estimation problems. As we shall see in section 6, the semiparametric efficiency bound for the single-index model is not achieved if we use the NLS weighting scheme only.

Clearly, as discussed in the previous section, the function

$$J(\theta) = \mathrm{E}\{W(x)[y - \mathrm{E}[y\,|\,h(x;\theta)]]^2\}$$

takes the minimum value $\mathrm{E}(W(x)\varepsilon^2)$ when $\theta = \theta_0$. For identification we should ask whether θ_0 is the only θ that achieves the minimum value. This holds only under further assumptions. The general condition is that for all $\theta \neq \theta_0$ there is a positive probability in x such that

$$\varphi(h(x;\theta_0)) \neq \mathrm{E}(\varphi\,|\,h(x;\theta)).$$

Since the condition involves as unknown function φ and the distribution of x given $h(x;\theta)$, it is not immediately obvious whether the condition holds under some general conditions. I provide sufficient conditions for identification in the linear single-index model.

The linear single-index specifies

$$h(x;\theta) = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_L x_L,$$

where $\theta = (\theta_1, \ldots, \theta_L)'$ and $\theta \in \Theta \subset R^L$.

We assume that the unknown function φ is smooth and not everywhere constant:

Assumption 4.1. The unknown function $\varphi: R \to R$ is (1) *differentiable and* (2) *not constant on the support of $x'\beta_0$.*

In the following analysis I make assumptions on regressors that are not necessary for usual regression analysis. To allow some of the regressors to be deterministically related to other regressors, I denote x^l, for $l = 1, \dots, L$, to be a function of the more fundamental regressors. That is, $x^l = x^l(z^1, \dots, z^{L'})$, for $l = 1, \dots, L$, is a known function from $R^{L'}$ into R. I call $x^1, \dots, x^L$ *nominal regressors* and $z^1, \dots, z^{L'}$ *underlying regressors.* To simplify the exposition, we assume that the underlying regressors are all either continuous or discrete, but we could allow regressors to be mixtures of continuous and discrete types by rewriting the assumptions all conditional on the continuous parts of mixed regressors.

Now we place some restrictions on nominal regressors that amount to placing joint restrictions on the functions $x^l: R^{L'} \to R$ for $l = 1, \dots, L$ and the random variables $z^1, \dots, z^{L'}$. We arrange regressors so that the first L_1 nominal regressors, $x^1, \dots, x^{L_1}$, and the first L_1' underlying regressors, $z^1, \dots, z^{L_1'}$, have continuous marginal distributions.

Assumption 4.2.

(1) *$x^l: R^{L'} \to R$ for $l = 1, \dots, L$ has a partial derivative almost everywhere in z with respect to all continuous underlying regressors $z^{l'}$ for $l' = 1, \dots, L_1'$.*

(2) *For discrete nominal regressors, $\partial x^l(z)/\partial \Sigma z^{l'} = 0$ for $l = L_1 + 1, \dots, L$, and $l' = 1, \dots, L_1'$ almost everywhere in z.*

(3) *For each $l' = 1, \dots, L'$, the set*

$$\{(s_{l'}^1, \dots, s_{l'}^{L_1})' \mid s_{l'}^l = \partial x^l/\partial z^{l'} \\ \text{for some } z \in Z \text{ for each } l = 1, \dots, L_1\}$$

has positive measure with respect to z and that the intersection of their linear orthogonal sets is a singleton of a zero vector in R^{L_1}; that is

$$\bigcap_{l'=1}^{L_1'} \{(s_{l'}^1, \dots, s_{l'}^{L_1})' \mid s_{l'}^l = \partial x^l/\partial z^{l'} \\ \text{for some } z \in Z \text{ for each } l = 1, \dots, L_1\}^{\perp} = \{0\},$$

where Z is the support of z and $A^{\perp}$ denotes the linear space that is orthogonal to a set A.

(4) *For each $\theta \in \Theta$ there exists an open interval $\mathcal{T}$ and at least $L - L_1 + 1$ constant vectors $c^l = (c_{L_1+1}^l, \dots, c_L^l)'$ for $l = 0, \dots, L - L_1$ such that (i) $c^l - c^0$ for $l = 1, \dots, L - L_1$ are linearly independent; (ii) $\mathcal{T}$ is included in*

the following set:

$$\bigcap_{l=0}^{L-L_1} \{t \mid t = \theta_1 x^1(z) + \cdots + \theta_{L_1} x^{L_1}(z) + \theta_{L_1+1} c^l_{L_1+1} + \cdots + \theta_L c^l_L$$

$$\textit{for some } z \in Z(c^l)\},$$

where

$$Z(c^l) = \{z \in Z \mid x_{L_1+1}(z) = c^l_{L_1+1}, \ldots, x_L(z) = c^l_L\};$$

and (iii) φ is not periodic on $\mathscr{T}$; that is, if $\varphi(t) = \varphi(t + p)$ for all $t \in \mathscr{T}$, then $p = 0$.

Assumption 4.2(4) places restrictions on the size of the parameter space Θ and on the support of continuous nominal regressors.

Theorem 4.1 (Identification of Linear Single-Index Models). *Let*

$$y = \varphi(x'\theta_0) + \varepsilon$$

be a linear single-index model.

(1) If there exists a continuous nominal regressor whose coefficient is not 0 *and if Assumptions 4.1 and 4.2(1–3) are satisfied, then the coefficients corresponding to all continuous nominal regressors are identified up to a scalar constant.*

(2) If furthermore Assumption 4.2(4) is satisfied, then θ_0 is identified up to a scalar constant.

Proof. Suppose $\theta^* = (1, \theta_2^*, \ldots, \theta_L^*)' \in \Theta$ minimizes the objective function $\mathrm{E}\{W(x)[y - \mathrm{E}(y \mid x'\theta^*)]^2\}$. Since θ_0 also minimizes the objective function,

$$\mathrm{E}\{W(x)[y - \mathrm{E}(y \mid x'\theta_0)]^2\} = \mathrm{E}\{W(x)[y - \mathrm{E}(y \mid x'\theta^*)]^2\}.$$

Since the left-hand side equals $\mathrm{E}[W(x)\varepsilon^2]$ and the right-hand side equals $\mathrm{E}[W(x)\varepsilon^2] + \mathrm{E}\{W(x)[\varphi - \mathrm{E}(\varphi \mid x'\theta^*)]^2\}$,

we have

$$\mathrm{E}\{W(x)[\varphi - \mathrm{E}(\varphi \mid x'\theta^*)]^2\} = 0.$$

Moreover $W(x) > 0$ implies

$$\varphi(x'\theta_0) = \mathrm{E}(\varphi \mid x'\theta^*) \quad \text{a.e. in } z.$$

Let $t = x'\theta^*$; then

$$\varphi(\theta_{01}t + \gamma_2 x^2 + \cdots + \gamma_L x^L) = \mathrm{E}(\varphi \mid t),$$

where $\gamma_l = \theta_{0l} - \theta_{01}\theta_l^*$. Taking partial derivatives with respect to $z^1, \ldots, z^{L_1}$, we have for almost all z

$$\begin{array}{l} \varphi'(x'\theta_0)[\gamma_2 \partial x^2/\partial z^1 + \cdots + \gamma_{L_1}\partial x^{L_1}/\partial z^1] = 0 \\ \vdots \\ \varphi'(x'\theta_0)[\gamma_2 \partial x^2/\partial z^{L_1} + \cdots + \gamma_{L_1}\partial x^{L_1}/\partial z^{L_1}] = 0. \end{array}$$

By Assumption 4.1, there is a positive probability that $\varphi' \neq 0$. Thus

$$\begin{array}{ccccc} [\gamma_2 \partial x^2/\partial z^1 + & \cdots & + \gamma_{L_1}\partial x^{L_1}/\partial z^1] & = & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [\gamma_2 \partial x^2/\partial z^{L_1} + & \cdots & + \gamma_{L_1}\partial x^{L_1}/\partial z^{L_1}] & = & 0 \end{array}$$

holds with positive probability. Assumption 4.3(3) then implies $\gamma_2 = \cdots = \gamma_{L_1} = 0$. This completes the proof of part (1).

Now we assume without loss of generality that the coefficients of the continuous random variables are proportional to the true coefficients; that is,

$$(\theta_1^*, \ldots, \theta_{L_1}^*)' = r(\theta_{01}, \ldots, \theta_{0L_1})' \quad \text{for some} \quad r \neq 0.$$

Let $t = r(x^1\theta_{01} + \cdots + x^{L_1}\theta_{0L_1}) + x^{L_1+1}\theta_{L_1+1}^* + \cdots + x^L\theta_L^*$. Then

$$\begin{aligned} &\varphi(x^1\theta_{01} + \cdots + x^L\theta_{0L}) \\ &\quad = \mathrm{E}(\varphi \mid t) = \varphi(t/r + (\theta_{0L_1+1} - \theta_{L_1+1}^*/r)x^{L_1+1} + \cdots + (\theta_{0L} - \theta_L^*/r)x^L). \end{aligned}$$

The last equality implies that the right-hand side is constant for all $(x^{L_1+1}, \ldots, x^L)$ values. Then Assumption 4.2(4) implies that

$$(\theta_{L_1+1}^*, \ldots, \theta_L^*)' = r(\theta_{0L_1+1}, \ldots, \theta_{0L})'. \quad ■$$

Note that if all regressors are discrete, then a special configuration of the discrete support is necessary in order to identify the parameter, as Example 4.1 shows.

Example 4.1. Consider a linear single-index model with $L = 2$; that is,

$$h(x;\theta) = \theta_1 x_1 + \theta_2 x_2,$$

where $\theta_0 = (\theta_{01}, \theta_{02})'$ *and* $\theta = (\theta_1, \theta_2)'$. *If*

(1) *the support of* x *and* X *consists of isolated points in* R^2,

(2) $\inf_{x,x' \in X} \| x - x' \| > 0$, *and*

(3) *the set of slopes defined by joining any two points in* X *is not dense in* R^2,

then there exists $\varepsilon_0 > 0$ *and* $r_0 \neq \theta_{02}/\theta_{01}$ *such that the* σ*-algebra generated by* x *coincides with that generated by the sets*

$$\{x \in X \mid -\varepsilon_0 < x_1 + r_0 x_2 < \varepsilon_0\}.$$

Proof. By assumption (3) in the statement of the example, there exist r_0 and an open neighborhood of r_0, $U(r_0)$, such that for any r in $U(r_0)$ there does not correspond any slope defined by connecting two points in X; that is, for each point $x^* = (x_1^*, x_2^*)' \in X$, defining

$$A(x^*) = \{(x_1, x_2) \in R^2 \mid x_1 - x_1^* + r(x_2 - x_2^*) = 0 \text{ for some } r \in U(r_0)\},$$

$$X \cap A(x^*) = \{(x_1^*, x_2^*)'\}.$$

Also, writing $c = \inf_{x,x' \in X} \| x - x' \|$ and

$$B(x^*) = \{(x_1, x_2) \in R^2 \mid (x_1 - x_1^*)^2 + (x_2 - x_2^*)^2 < c^2\},$$

$$X \cap B(x^*) = \{(x_1^*, x_2^*)'\}.$$

Since the set $A(x^*) \cup B(x^*)$ includes a rectangular set of the form

$$\{(x_1, x_2) \in R^2 \mid -\varepsilon_0 < (x_1 - x_1^*) + r_0(x_2 - x_2^*) < \varepsilon_0\},$$

for some small but positive ε_0, we have desired results. ■

As a direct corollary to this example, no $r \in U(r_0)$ is distinguishable by SLS estimation from the slope of interest, θ_{02}/θ_{01}. Clearly, θ_{02}/θ_{01} is not in $U(r_0)$ in general. Therefore we assume that there exists a continuous variable with a nonzero coefficient.

Note, however, that if all regressors are discrete, then for each point in X, $\mathrm{E}(y\,|\,x)$ can be estimated with rate $1/\sqrt{n}$, which is the usual convergence rate for parametric models. In that case, therefore, researchers might as well take a nonparametric approach rather than a semiparametric one.

Assumption 4.1(1) is assumed for convenience. One could assume the existence of one-sided derivatives instead, at the cost of a longer proof. Without Assumption 4.1(2), clearly one cannot identify θ_0.

Assumptions 4.2(1) and 4.2(2) are regularity conditions. Assumption 4.2(3) excludes exact multicollinearity problems. Example 4.2 shows that the assumption is easy to check.

Example 4.2. Consider the same model as in Example 4.1 except that $x_1 = z$, $x_2 = z^2$, *and* z *has support on the unit interval* $[0,1]$. *Thus, in this example* $L = L_1 = 2$, $L' = L_1' = 1$, *and the set specified in Assumption 4.2(3) is* $\{(1,z)'\,|\,z\in[0,1]\}^{\perp}$, *which is* $\{0\}$ *as required.*

Assumption 4.2(4) rules out the situation described by Examples 4.3 and 4.4.[11]

Example 4.3. Consider the same model as in Example 4.1 except that x_1 *has support on* $[0,r_0]$, $r_0 > 0$, *and* x_2 *has discrete support* $\{0,1\}$. *Consider any* $r = \theta_2/\theta_1$ *that is greater than* r_0. *Then the line defined by* $x_1 + rx_2$ *intersects with the support of* $(x_1,x_2)'$ *at most once, and therefore conditioning on the line is the same as conditioning on* $(x_1,x_2)'$. *That is,* $\mathrm{E}(y\,|\,\theta_1 x_1 + \theta_2 x_2) = \varphi(x_1\theta_{01} + x_2\theta_{02})$ *for all* $(\theta_1,\theta_2)'$ *such that* $\theta_2/\theta_1 > r_0$.

Example 4.4. Consider the same model as in Example 4.3 except that now the parameter space is restricted to $[0,r_0]$ *so that the problem raised in Example 4.3 is not an issue. Suppose* φ *is periodic with periodicity* p. *Since the support of* x_2 *is* $\{0,1\}$, *it is easy to verify that*

$$\mathrm{E}[W(x)[y - \mathrm{E}(y\,|\,\theta_1 x_1 + \theta_2 x_2)]^2] = \mathrm{E}[W(x)\varepsilon^2],$$

for all θ_2/θ_1, *such that*

$$\theta_2/\theta_1 = ip/\theta_{01} + \theta_{02}/\theta_{01},$$

for any $i = 1, 2, \ldots$.

[11] I thank Professors Jonathan Feinstein and Kazumitsu Nawata for the examples.

5. Asymptotic properties of the SLS and WSLS estimators

In this section I prove that the SLS and the WSLS estimators are consistent and asymptotically normal. Since WSLS includes SLS as a special case, we study the asymptotics of WSLS.

We assume that (x_i', y_i) for $i = 1, \dots, n$ are observed.

Assumption 5.1. Observations (x_i', y_i) *for* $i = 1, \dots, n$ *are i.i.d.*

We introduce a parameter space Θ.

Assumption 5.2. Θ *is a subset of* R^M *and is compact. Moreover,* θ_0 *is in the interior of* Θ.

Let a Lebesgue density of $t = h(x;\theta)$ be $f(t;\theta)$ and define

$$T_\theta(X) = \{t \in R \mid t = h(x;\theta) \quad \text{and} \quad x \in X\}.$$

We also define a subset X of the support of x that satisfies the following assumptions.

Assumption 5.3.

(1) X *is compact.*

(2) $\inf_{x \in X} f(h(x;\theta);\theta) > 0$.

(3) $f(t;\theta)$ *and* $\mathrm{E}[y \mid h(x;\theta) = t]$ *are three times continuously differentiable with respect to* t*, and the third derivatives satisfy Lipschitz conditions for all* $t \in T_\theta(X)$ *uniformly in* $\theta \in \Theta$.

Assumption 5.4. The dependent variable y *has the mth-order absolute moment, where* $m \geq 2$*, and the conditional variance of* y *given* x *is uniformly bounded and bounded away from* 0 *on the subset* X *chosen in Assumption 5.3.*

Assumption 5.5. The function $h(x;\theta)$ *satisfies a Lipschitz condition on* $X \times \Theta$.

We allow kernel function K that satisfy the following assumptions.

Assumption 5.6.

(1) $K(s)$ *is twice continuously differentiable, and the second derivative satisfies a Lipschitz condition.*

(2) $\int K(s)\,\mathrm{d}s = 1$.

(3) $\int sK(s)\,\mathrm{d}s = 0$.

(4) $K(s) = 0$ *for* $s < -1$ *and* $s > 1$.

Note in particular that Assumption 5.6(1) implies Lipschitz properties of $K(s)$ and $K'(s)$ and that Assumptions 5.6(1) and (4) imply boundedness of $K(s)$. We place restrictions on the bandwidth sequence as we need them.

Let $\mathrm{E}_W(x,\theta)$ be the probability limit of $\hat{\mathrm{E}}_W(x,\theta)$. The consistency proof reduces to showing uniform convergence of the kernel regression estimator, and thus the following lemma (proven in the appendix) is useful.

Lemma 5.1. Under Assumptions 5.1–5.6, if $a_n \to 0$ *and* $na_n^{1+1/(m-1)}/(-\log a_n) \to \infty$, *where* $m \geq 2$ *is the highest absolute moment of y, then for any* $\varepsilon > 0$,

$$\Pr\left\{\sup_{(x',\theta')'\in X\times\Theta} |\hat{\mathrm{E}}_W(x,\theta) - \mathrm{E}_W(x,\theta)| > \varepsilon\right\}$$

converges to 0 *as* $n \to \infty$.

Theorem 5.1 (Consistency). *The WSLS estimator is consistent if Assumptions 5.1–5.6 hold, the bandwidth sequence satisfies* $a_n \to 0$ *and* $na_n^{1+1/(m-1)}/(-\log a_n) \to \infty$, *where* $m \geq 2$ *is the highest absolute moment of y, and* θ_0 *is identified.*

Proof. By the definition of WSLS,

$$\Pr\{J_n^{1/2}(\hat{\theta}) \leq J_n^{1/2}(\theta_0)\} = 1.$$

On the other hand, note that for any open set $U(\theta_0)$ that includes θ_0,

$$\begin{aligned}\Pr\{J_n^{1/2}(\hat{\theta}) \leq J_n^{1/2}(\theta_0)\} &= \Pr\{J_n^{1/2}(\hat{\theta}) \leq J_n^{1/2}(\theta_0) \text{ and } \hat{\theta}\in U(\theta_0)\} \\ &\quad + \Pr\{J_n^{1/2}(\hat{\theta}) \leq J_n^{1/2}(\theta_0) \text{ and } \hat{\theta}\in \Theta\backslash U(\theta_0)\} \\ &\leq \Pr\{\hat{\theta}\in U(\theta_0)\} \\ &\quad + \Pr\left\{\inf_{\theta\in\Theta\backslash U(\theta_0)} J_n^{1/2}(\theta) \leq J_n^{1/2}(\theta_0)\right\}.\end{aligned}$$

Therefore,

$$\Pr\left\{\inf_{\theta\in\Theta\backslash U(\theta_0)} J_n^{1/2}(\theta) \leq J_n^{1/2}(\theta_0)\right\} \to 0$$

implies consistency. To see that it indeed converges to 0, note that

$$\begin{aligned}
&\Pr\left\{\inf_{\theta\in\Theta\setminus U(\theta_0)} J_n^{1/2}(\theta)\le J_n^{1/2}(\theta_0)\right\}\\
&=\Pr\left\{\inf_{\theta\in\Theta\setminus U(\theta_0)}\left[J_n^{1/2}(\theta)-\tilde{J}_n^{1/2}(\theta)+\tilde{J}_n^{1/2}(\theta)-J^{1/2}(\theta)+J^{1/2}(\theta)\right]\right.\\
&\qquad\left.\le J_n^{1/2}(\theta_0)\right\}\\
&\le\Pr\left\{\inf_{\theta\in\Theta\setminus U(\theta_0)}\left[J_n^{1/2}(\theta)-\tilde{J}_n^{1/2}(\theta)\right]\right.\\
&\qquad+\inf_{\theta\in\Theta\setminus U(\theta_0)}\left[\tilde{J}_n^{1/2}(\theta)-J^{1/2}(\theta)\right]+J^{1/2}(\theta_0)-J_n^{1/2}(\theta_0)\\
&\qquad\left.\le J^{1/2}(\theta_0)-\inf_{\theta\in\Theta\setminus U(\theta_0)}J^{1/2}(\theta)\right\}\\
&\le\Pr\left\{\sup_{\theta\in\Theta}|J_n^{1/2}(\theta)-\tilde{J}_n^{1/2}(\theta)|+\sup_{\theta\in\Theta}|\tilde{J}_n^{1/2}(\theta)-J^{1/2}(\theta)|\right.\\
&\qquad\left.+|J_n^{1/2}(\theta_0)-J^{1/2}(\theta_0)|\ge\inf_{\theta\in\Theta\setminus U(\theta_0)}J^{1/2}(\theta)-J^{1/2}(\theta_0)\right\},
\end{aligned}$$

where

$$\tilde{J}_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}W(x_i)I(x_i\in X)[y_i-\mathrm{E}_W(x_i,\theta)]^2,$$

$$J(\theta)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{W(x_i)I(x_i\in X)[y_i-\mathrm{E}_W(x_i,\theta)]^2\}.$$

For each open set $U(\theta_0)$, the identification condition guarantees that there exists $\varepsilon>0$ such that

$$\inf_{\theta\in\Theta\setminus U(\theta_0)}J^{1/2}(\theta)-J^{1/2}(\theta_0)>\varepsilon.$$

Therefore it is sufficient to show that for each $\varepsilon > 0$,

$$\Pr\left\{\sup_{\theta\in\Theta} |J_n^{1/2}(\theta) - \tilde{J}_n^{1/2}(\theta)| > \varepsilon\right\} \to 0, \tag{1}$$

$$\Pr\left\{\sup_{\theta\in\Theta} |\tilde{J}_n^{1/2}(\theta) - J^{1/2}(\theta)| > \varepsilon\right\} \to 0, \tag{2}$$

$$\Pr\{ J_n^{1/2}(\theta) - J^{1/2}(\theta_0)| > \varepsilon\} \to 0. \tag{3}$$

Clearly, result (3) follows from results (1) and (2). Lemma 5.2 establishes result (1), and Lemma 5.3 establishes result (2). ■

Note that the consistency proof holds even when we replace the kernel regression estimator with any nonparametric estimator that satisfies the uniform convergence result of Lemma 5.1. As we shall see later (see Theorem 5.2), Lemma 5.1 is not sufficient for the asymptotic normality result.

Lemma 5.2. Under Assumptions 5.1–5.6, if $a_n \to 0$ *and* $na_n^{1+1/(m-1)}/(-\log a_n) \to \infty$, *where* $m \geq 3$ *is the highest absolute moment of the dependent random variable* y, *then for any* $\varepsilon > 0$,

$$\Pr\left\{\sup_{\theta\in\Theta} |J_n^{1/2}(\theta) - \tilde{J}_n^{1/2}(\theta)| > \varepsilon\right\}$$

converges to 0 *as* $n \to \infty$.

Proof. Since

$$|J_n^{1/2}(\theta) - \tilde{J}_n^{1/2}(\theta)| \leq \left\{\frac{1}{n}\sum_{i=1}^{n} [\mathrm{E}_W(x_i,\theta) - \hat{\mathrm{E}}_W(x_i,\theta)]^2\right\}^{1/2},$$

Lemma 5.1 implies the result. ■

Lemma 5.3. Under Assumptions 5.1–5.5, for any $\varepsilon > 0$,

$$\Pr\left\{\sup_{\theta\in\Theta} |\tilde{J}_n^{1/2}(\theta) - J^{1/2}(\theta)| > \varepsilon\right\}$$

converges to 0 *as* $n \to \infty$.

Proof. Since

$$|\tilde{J}_n^{1/2}(\theta) - J^{1/2}(\theta)| \leq |\tilde{J}_n(\theta) - J(\theta)|^{1/2},$$

it is sufficient to show that for any $\varepsilon > 0$,

$$\Pr\left\{\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} \psi(x_i, y_i, \theta)\right| > \varepsilon\right\} \to 0$$

as $n \to \infty$, where

$$\psi(x_i, y_i, \theta) = W(x_i)I(x_i\in X)[y_i - \mathrm{E}_W(x_i, \theta)]^2$$
$$- \mathrm{E}\{W(x_i)I(x_i\in X)[y_i - \mathrm{E}_W(x_i, \theta)]^2\}.$$

This type of uniform convergence result is well established in the literature. We can, for example, use the approach of Andrews (1987) to show the result. Under our assumptions, Assumptions A1, A2, and A4 of Andrews (1987) are satisfied, and therefore the result holds. ■

Next we turn to proving asymptotic normality of the WSLS estimator. Denote a sequence of random functions $f_n(\theta)$ by $o_p(\alpha_n)$ if $f_n(\theta)/\alpha_n(\theta)$ converges in probability to 0 for a sequence of random functions $\alpha_n(\theta)$ uniformly in $\theta\in\Theta$, and denote a sequence of random functions $f_n(\theta)$ by $O_P(\alpha_n)$ if $f_n(\theta)/\alpha_n(\theta)$ is stochastically bounded for a sequence of random functions $\alpha_n(\theta)$ uniformly in θ. Also, we use the sup-norm, denoted $|a|$, for any finite-dimensional vectors and matrices a; that is, $|a|$ is a maximum absolute element of a.

It is convenient to organize the proof around the following lemma.

Lemma 5.4. Let $J_n(\theta)$ be twice continuously differentiable with probability approaching 1 *uniformly in θ, and let the following conditions hold when $J_n(\theta)$ is differentiable.*

(1) $\hat{\theta}$ converges in probability to θ_0 and

$$\hat{\theta} = \arg\inf_{\theta\in\Theta} J_n(\theta).$$

(2) There exists a random vector Δ_n that converges in distribution to a normal random vector with mean 0 *and variance–covariance matrix Σ, such that*

$$\left|\sqrt{n}\,\frac{\partial J_n(\theta_0)}{\partial\theta} - \Delta_n\right| = o_p(1).$$

(3) There exists a positive-definite matrix V such that

$$\left|\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'} - V\right| = o_p(1).$$

(4) For any $\varepsilon > 0$, there exists a neighborhood of θ_0, U_0, such that

$$\Pr\left\{\sup_{\theta\in U_0}\left|\frac{\partial^2 J_n(\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'}\right| > \varepsilon\right\} \to 0$$

as $n \to \infty$.

Then $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to a normal random vector with mean 0 and variance–covariance matrix $V^{-1}\Sigma V^{-1}$.

Proof. The proof consists of two steps: the first step shows that the estimator is $1/\sqrt{n}$-consistent, and the second step shows that the estimator is asymptotically normal.

Step 1. With probability close to 1, by Taylor expansion, for $\bar{\theta}$ between $\hat{\theta}$ and θ_0,

$$J_n(\hat{\theta}) = J_n(\theta_0) + \frac{\partial J_n(\theta_0)}{\partial\theta'}(\hat{\theta} - \theta_0) + \tfrac{1}{2}(\hat{\theta} - \theta_0)'\frac{\partial^2 J_n(\bar{\theta})}{\partial\theta\partial\theta'}(\hat{\theta} - \theta_0).$$

By assumption (1) in the statement of the lemma,

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

Then the Taylor expansion formula and assumptions (3) and (4) imply that

$$(\hat{\theta} - \theta_0)'\frac{\partial J_n(\theta_0)}{\partial\theta} + \tfrac{1}{2}(\hat{\theta} - \theta_0)'V(\hat{\theta} - \theta_0) + o_p(|\hat{\theta} - \theta_0|^2) \leq 0.$$

Multiply both sides by $n(1 + \sqrt{n}|\hat{\theta} - \theta_0|)^{-2}$ and define

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

Then

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

If $\sqrt{n}|\hat{\theta}-\theta_0| \to \infty$, then the inequality implies $c_n'(\hat{\theta})Vc_n(\hat{\theta}) \leq o_p(1)$. Since V is positive-definite, $|c_n(\hat{\theta})| = o_p(1)$ or $\sqrt{n}|\hat{\theta}-\theta_0| = o_p(1)$. This is a contradiction. Therefore $\sqrt{n}|\hat{\theta}-\theta_0| = O_P(1)$.

Step 2. Step 1 and Taylor expansion imply

$$J_n(\hat{\theta}) = J_n(\theta_0) + (\hat{\theta}-\theta_0)' \frac{\partial J_n(\theta_0)}{\partial \theta'} + \tfrac{1}{2}(\hat{\theta}-\theta_0)' V(\hat{\theta}-\theta_0)' + o_p(n^{-1}).$$

Rewriting, we have

$$J_n(\hat{\theta}) = \tfrac{1}{2}\left[(\hat{\theta}-\theta_0) + V^{-1}\frac{\partial J_n(\theta_0)}{\partial \theta}\right]' V \left[(\hat{\theta}-\theta_0) + V^{-1}\frac{\partial J_n(\theta_0)}{\partial \theta}\right]$$

$$+ J_n(\theta_0) - \tfrac{1}{2}\frac{\partial J_n(\theta_0)}{\partial \theta'} V^{-1} \frac{\partial J_n(\theta_0)}{\partial \theta} + o_p(n^{-1}) \leq 0.$$

Next evaluate $J_n(\theta)$ at

$$\tilde{\theta} = \theta_0 - V^{-1}\frac{\partial J_n(\theta_0)}{\partial \theta}.$$

Applying the same argument as above we have

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

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

$$J_n(\hat{\theta}) \leq J_n(\tilde{\theta}).$$

Therefore,

$$\tfrac{1}{2}\left[(\hat{\theta}-\theta_0) + V^{-1}\frac{\partial J_n(\theta_0)}{\partial \theta}\right]' V \left[(\hat{\theta}-\theta_0) + V^{-1}\frac{\partial J_n(\theta_0)}{\partial \theta}\right] \leq o_p(n^{-1}).$$

Multiplying both sides by n, since V is positive-definite,

$$\sqrt{n}(\hat{\theta}-\theta_0) = -V^{-1}\sqrt{n}\,\frac{\partial J_n(\theta_0)}{\partial \theta} + o_p(1). \quad \blacksquare$$

Theorem 5.2 (Asymptotic Normality). *Under Assumptions 5.1–5.6, if dependent variable y has first m absolute moments, when $m \geq 3$, the parameter θ_0 is identified, and the bandwidth sequence satisfies $na_n^8 \to 0$ and $na_n^{3+3/(m-1)}/(-\log a_n) \to \infty$, then the WSLS estimator converges with rate $1/\sqrt{n}$ and the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ is normal with mean 0 and variance–covariance matrix $V^{-1}\Sigma V^{-1}$, where*

$$V = \mathrm{E}\left[W(x) \frac{\partial \mathrm{E}_W(x, \theta_0)}{\partial \theta} \frac{\partial \mathrm{E}_W(x, \theta_0)}{\partial \theta'} \middle| x \in X \right],$$

$$\Sigma = \mathrm{E}\left[W^2(x)\sigma^2(x) \frac{\partial \mathrm{E}_W(x, \theta_0)}{\partial \theta} \frac{\partial \mathrm{E}_W(x, \theta_0)}{\partial \theta'} \middle| x \in X \right],$$

$$\frac{\partial \mathrm{E}_W(x, \theta_0)}{\partial \theta} = \varphi'(h(x; \theta_0))$$

$$\times \left[\frac{\partial h(x; \theta_0)}{\partial \theta} - \mathrm{E}_W\left(\frac{\partial h(x; \theta_0)}{\partial \theta} \middle| h(x; \theta_0), x \in X \right) \right],$$

where $\sigma^2(x) = \mathrm{var}(y \mid x)$.

Note that the variance–covariance matrix is similar to that of the conventional NLS estimator when $\varphi(h(x;\theta))$ is used. They are not quite the same because φ is not known. The WSLS approach may utilize only the variation along the known function $h(x;\theta)$, and therefore conditional variance, instead of the second moment, of $\varphi'\partial h(x;\theta_0)/\partial\theta$ shows up in the formula. Thus we can identify two sources of inefficiency with the WSLS estimator: heteroskedasticity and the semiparametric nature of the problem. In fact, I show in the next section that the optimally weighted WSLS estimator achieves the semiparametric efficiency bound for the single-index model.

Following lemmas (proven in the appendix) are useful for the proof of asymptotic normality of WSLS estimator.

Lemma 5.5. Order $h(x_i;\theta)$ from smallest to largest, and call the ordered index, $h(x_i;\theta)$, $h_{(i)}$. Let $s_i = h_{(i+1)} - h_{(i)}$ for $i = 0, \ldots, n$ and let $h_{(0)}$ and $h_{(n+1)}$ be the lower and the upper endpoints respectively, of the support of $h(x_i;\theta)$. If assumption 5.3 holds, then the second moment of s_i is of order $\mathrm{O}(n^{-2})$ uniformly in i.

Lemma 5.6. Under Assumptions 5.1–5.6, if $a_n \to 0$ and $na_n^{2+2/(m-1)}/(-\log a_n) \to \infty$, where $m \geq 2$ is the highest absolute moment of y, then for *any* $\varepsilon > 0$,

$$\Pr\left\{\sup_{(x',\theta')'\in X\times\Theta}\left|\frac{\partial\hat{\mathrm{E}}_W(x,\theta)}{\partial\theta}-\frac{\partial\mathrm{E}_W(x,\theta)}{\partial\theta}\right|>\varepsilon\right\}$$

converges to 0 *as* $n \to\infty$.

Lemma 5.7. Under Assumptions 5.1–5.6, if $a_n \to 0$ *and* $na_n^{3+3/(m-1)}/(-\log a_n)\to\infty$, *where* $m \geq 2$ *is the highest absolute moment of y, then for any* $\varepsilon > 0$,

$$\Pr\left\{\sup_{(x',\theta')'\in X\times\Theta}\left|\frac{\partial^2\hat{\mathrm{E}}_W(x,\theta)}{\partial\theta\partial\theta'}-\frac{\partial^2\mathrm{E}_W(x,\theta)}{\partial\theta\partial\theta'}\right|>\varepsilon\right\}$$

converges to 0 *as* $n \to\infty$.

Lemma 5.8. Under Assumptions 5.1–5.6, if $a_n \to 0$ *and* $na_n^4 \to \infty$,

$$\Pr\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i\left|\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta}-\frac{\partial\mathrm{E}_W(x_i,\theta_0)}{\partial\theta}\right]\right|>\varepsilon\right\}$$

converges to 0 *as* $n \to\infty$.

Lemma 5.9. Under Assumptions 5.1–5.6, if $a_n \to 0$ *and* $na_n^4 \to \infty$,

$$\Pr\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i[\varphi(h(x_i;\theta_0))-\hat{\mathrm{E}}_W(x_i,\theta_0)]\frac{\partial\mathrm{E}_W(x_i,\theta_0)}{\partial\theta}\right|>\varepsilon\right\}$$

converges to 0 *as* $n \to\infty$.

Lemma 5.10. Under Assumptions 5.1–5.6, if $na_n^8 \to 0$ *and* $na_n^{3+3/(m-1)}/(-\log a_n)\to\infty$, *and dependent variable y has first m absolute moments, where* $m \geq 3$, *then*

$$\Pr\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i[\varphi(h(x_i;\theta_0))-\hat{\mathrm{E}}_W(x_i,\theta_0)]\right.\right.$$
$$\left.\left.\times\left[\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta}-\frac{\partial\mathrm{E}_W(x_i,\theta_0)}{\partial\theta}\right]\right|>\varepsilon\right\}$$

converges to 0 *as* $n \to\infty$.

Proof. We first show that the objective function is twice continuously differentiable with probability approaching 1 as $n \to \infty$. We then verify the three conditions in the lemma above.

Step 1. Since none of the denominators of the kernel regression estimators are zero if the maximum spacing of $h(x_i, \theta)$ is less than a_n, it is sufficient for our purpose to prove that

$$\sup_{\theta \in \Theta} \Pr\left\{\max_i s_i > a_n\right\},$$

where s_i is defined in Lemma 5.5, goes to 0 as $n \to \infty$. Since

$$\Pr\left\{\max_i s_i > a_n\right\} \le \sum_{i=1}^{n} \Pr\{s_i > a_n\},$$

and, under our assumption, $na_n^2 \to \infty$, by Markov's inequality, it is sufficient to show that the second moments of s_i are $\mathrm{O}(n^{-2})$ uniformly in i, which is shown in Lemma 5.5.

Step 2. To see that

$$\sqrt{n}\,\frac{\partial J_n(\theta_0)}{\partial \theta} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma),$$

note that

$$\sqrt{n}\,\frac{\partial J_n(\theta_0)}{\partial \theta} = -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} W(x_i) I(x_i \in X)[y_i - \hat{\mathrm{E}}_W(x_i, \theta_0)]\, \frac{\partial \hat{\mathrm{E}}_W(x_i, \theta_0)}{\partial \theta}$$

$$= -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} W_i I_i \varepsilon_i \frac{\partial \mathrm{E}_W(x_i, \theta_0)}{\partial \theta} \tag{4}$$

$$-\frac{2}{\sqrt{n}} \sum_{i=1}^{n} W_i I_i \varepsilon_i \left[\frac{\partial \hat{\mathrm{E}}_W(x_i, \theta_0)}{\partial \theta} - \frac{\partial \mathrm{E}_W(x_i, \theta_0)}{\partial \theta}\right] \tag{5}$$

$$-\frac{2}{\sqrt{n}} \sum_{i=1}^{n} W_i I_i [\varphi(h(x_i; \theta_0)) - \hat{\mathrm{E}}_W(x_i, \theta_0)]\, \frac{\partial \mathrm{E}_W(x_i, \theta_0)}{\partial \theta} \tag{6}$$

$$-\frac{2}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i[\varphi(h(x_i;\theta_0)) - \hat{\mathrm{E}}_W(x_i,\theta_0)]$$

$$\times\left[\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta} - \frac{\partial\mathrm{E}_W(x_i,\theta_0)}{\partial\theta}\right]. \tag{7}$$

Clearly, term (4) converges in distribution to N$(0,\Sigma)$. Lemmas 5.8–5.10 show that terms (5), (6), and (7) all converge in probability to 0 under our assumptions.

Step 3. To see that

$$\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'} \xrightarrow{\mathrm{p}} V,$$

note that

$$\frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n} W(x_i)I(x_i\in X)\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta}\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta'} \tag{8}$$

$$-\frac{1}{n}\sum_{i=1}^{N} W(x_i)I(x_i\in X)[y_i - \hat{\mathrm{E}}_W(x_i,\theta)]\frac{\partial^2\hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta\partial\theta'}. \tag{9}$$

Clearly, Lemma 5.6 implies that the right-hand side of line (8) converges in probability to V and Lemma 5.1 and Lemma 5.7 imply that the expression on line (9) converges in probability to the zero matrix.

Step 4. To see that for any $\varepsilon > 0$ there exists a neighborhood of θ_0, U_0, such that

$$\Pr\left\{\sup_{\theta\in U_0}\left|\frac{\partial^2 J_n(\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'}\right| > \varepsilon\right\} \to 0$$

as $n \to \infty$, note that

$$\sup_{\theta\in U_0}\left|\frac{\partial^2 J_n(\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 J_n(\theta_0)}{\partial\theta\partial\theta'}\right|$$

$$= \sup_{\theta\in U_0}\left|\frac{1}{n}\sum_{i=1}^{n} W_i I_i\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta}\frac{\partial\hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta'}\right.$$

$$-\frac{1}{n}\sum_{i=1}^{n} W_i I_i[y_i - \hat{\mathrm{E}}_W(x_i,\theta)]\frac{\partial^2\hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta\partial\theta'}$$

$$-\frac{1}{n}\sum_{i=1}^{n} W_i I_i \frac{\partial \hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta}\,\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta'}$$

$$+\frac{1}{n}\sum_{i=1}^{n} W_i I_i[y_i - \hat{\mathrm{E}}_W(x_i,\theta)]\,\frac{\partial^2 \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta\,\partial\theta'}\Bigg|$$

$$\leq \sup_{\theta\in U_0}\left|\frac{1}{n}\sum_{i=1}^{n} W_i I_i\left[\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta}-\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta}\right]\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta'}\right| \tag{10}$$

$$+\sup_{\theta\in U_0}\left|\frac{1}{n}\sum_{i=1}^{n} W_i I_i\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta}\left[\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta'}-\frac{\partial \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta'}\right]\right| \tag{11}$$

$$+\sup_{\theta\in U_0}\left|\frac{1}{n}\sum_{i=1}^{n} W_i I_i y_i\left[\frac{\partial^2 \hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta\,\partial\theta'}-\frac{\partial^2 \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta\,\partial\theta'}\right]\right| \tag{12}$$

$$+\sup_{\theta\in U_0}\left|\frac{1}{n}\sum_{i=1}^{n} W_i I_i[\hat{\mathrm{E}}_W(x_i,\theta)-\hat{\mathrm{E}}_W(x_i,\theta_0)]\,\frac{\partial^2 \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta\,\partial\theta'}\right| \tag{13}$$

$$+\sup_{\theta\in U_0}\left|\frac{1}{n}\sum_{i=1}^{n} W_i I_i \hat{\mathrm{E}}_W(x_i,\theta)\left[\frac{\partial^2 \hat{\mathrm{E}}_W(x_i,\theta)}{\partial\theta\,\partial\theta'}-\frac{\partial^2 \hat{\mathrm{E}}_W(x_i,\theta_0)}{\partial\theta\,\partial\theta'}\right]\right|. \tag{14}$$

Clearly, Lemmas 5.1, 5.6, and 5.7 imply that (10)–(14) all converge in probability to 0 in a neighborhood of θ_0, $U(\theta_0)$. ■

6. Weighting in WSLS

I will distinguish between *outer weighting*, the weighting that corresponds to the NLS weighting scheme, and *inner weighting*, the weighting used in kernel regression estimators. Outer weighting is introduced for the same reason that weighting is introduced in parametric heteroskedastic models: to adjust the error terms to homoskedastic ones. Inner weighting has a similar role. Recall that a kernel regression estimator, like any other nonparametric estimator, is a local average of data. Thus if the conditional variance of y given x depends only on $h(x;\theta_0)$, then the conditional variances of y_j given x_j such that $h(x_j;\theta_0)$ is near $h(x;\theta_0)$ are almost constant. On the other hand, if the conditional variance depends on x other than $h(x;\theta_0)$, then even for the data near $h(x;\theta_0)$, the variance of y given x is heteroskedastic. Thus the regression may be estimated more efficiently using weighted kernel estimators.[12]

[12] This point is further pursued in Ichimura and Newey (1990).

But increasing efficiency is not the only role of the inner weighting; it also reduces bias. To see this, note that

$$W(x)\,\frac{\partial \hat{\mathrm{E}}_W(x;\theta_0)}{\partial\theta}$$

$$\to W(x)\varphi'(h(x;\theta_0))$$

$$\times\left[\frac{\partial h(x_i;\theta_0)}{\partial\theta}-\frac{\mathrm{E}(W(x)\partial h(x;\theta_0)/\partial\theta)\,|\,h(x;\theta_0),\,x\in X)}{\mathrm{E}(W\,|\,h(x;\theta_0),\,x\in X)}\right]$$

in probability and that the right-hand side has conditional mean 0 given $h(x;\theta_0)$ for $x\in X$. Note also that if only the outer weighting is used, then the right-hand side does not have mean 0 unless $W(x)$ is a function of x only through $h(x;\theta_0)$.[13]

Indeed, the asymptotic variance–covariance matrix of the WSLS estimator, when $W(x)=1/\sigma^2(x)$ is used, coincides with the efficiency bound obtained by Newey (1990) for the single-index model, but if only outer weighting is used, $W(x)=1/\sigma^2(x)$ does not lead to the efficiency bound unless $W(x)$ depends on x only through $h(x;\theta_0)$.

By the same reason the inner weighting reduces bias, trimmings used in kernel regression estimators reduce bias. To see this, note that if we take $\{X_n\}_{n=1}^{\infty}$ such that $\bigcap_{n=1}^{\infty} X_n=\tilde{X}$, where $\tilde{X}\neq X$, then

$$W(x)\,\frac{\partial \hat{\mathrm{E}}_W(x;\theta_0)}{\partial\theta}$$

$$\to W(x)\varphi'(h(x;\theta_0))$$

$$\times\left[\frac{\partial h(x_i;\theta_0)}{\partial\theta}-\frac{\mathrm{E}(W(x)\partial h(x;\theta_0)/\partial\theta)\,|\,h(x;\theta_0),\,x\in X)}{\mathrm{E}(W\,|\,h(x;\theta_0),\,x\in X)}\right]$$

in probability and the conditional mean of the right-hand side given $h(x;\theta_0)$ for $x\in X$ is not zero.

7. Estimation of the covariance matrix

In order to perform hypothesis tests and construct confidence intervals we need a consistent estimator of the covariance matrix. Recall that the asymptotic variance–covariance matrix is $V^{-1}\Sigma V^{-1}$, where V and Σ are defined in Theorem 5.2.

[13] Therefore when the conditional variance of y given x is a function only of $h(x;\theta_0)$, such as in binary choice models, there is no need for inner weighting.

Theorem 7.1. Under Assumptions 5.1–5.6, if $a_n \to 0$ *and* $na_n^{2+2/(m-1)}/(-\log a_n) \to \infty$, *where* $m \geq 3$ *is the highest absolute moment of dependent variable y, then*

(1) V can be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n} W_i I_i \frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta}\frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta'},$$

(2) Σ can be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n} W_i I_i [y_i - \hat{\mathrm{E}}_W(x_i,\hat{\theta})]^2 \frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta}\frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta'}.$$

Note that the bandwidth used in the variance–covariance estimator may differ from that used to construct the SLS or WSLS estimator.

Proof. To show part (1), it is sufficient to show that

$$\left|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta}\frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta'} - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathrm{E}_W(x_i,\theta_0)}{\partial \theta}\frac{\partial \mathrm{E}_W(x_i,\theta_0)}{\partial \theta'}\right| = \mathrm{o}_p(1).$$

But this follows from consistency of $\hat{\theta}$ and Lemma 5.6. To show part (2), it is sufficient to show that

$$\left|\frac{1}{n}\sum_{i=1}^{n} [y_i - \hat{\mathrm{E}}_W(x_i,\hat{\theta})]^2 \frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta}\frac{\partial \hat{\mathrm{E}}_W(x_i,\hat{\theta})}{\partial \theta'} - \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \frac{\partial \mathrm{E}_W(x_i,\theta_0)}{\partial \theta}\frac{\partial \mathrm{E}_W(x_i,\theta_0)}{\partial \theta'}\right| = \mathrm{o}_p(1).$$

But this also follows from consistency of $\hat{\theta}$ and Lemmas 5.1 and 5.6. ■

8. Monte Carlo results

In this section we look at the small-sample properties of the SLS estimator via a Monte Carlo experiment with the sample size of 250 and 1000 trials. The construction of the experiment is identical to that of Cosslett (1986). He considers a binary choice model with two regressors:

$$y_i^* = \alpha_0 + \beta_{10}x_{1i} + \beta_{20}x_{2i} + \varepsilon_i.$$

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

The true parameter values are $\alpha_0 = 0$, $\beta_{10} = -2$, and $\beta_{20} = 1$. In Cosslett's specification, exogenous variables take two distributions and the errors take three distributions, giving rise to six models. The exogenous variables x_1 and x_2 are independently distributed. The two distributions of the exogenous variables are standard normal and standard exponential. Three mixtures of normal distributions are considered for the error distributions: (1) standard normal, (2) $0.75 \cdot \mathrm{N}(0,1) + 0.25 \cdot \mathrm{N}(0.25)$, and (3) $0.75 \cdot \mathrm{N}(-0.5,\ 1) + 0.25 \cdot \mathrm{N}(1.5,\ 2.5)$. 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 Cosslett's models in order to facilitate comparison of the performance of the SLS estimator with that of other estimators presented in his paper, including the maximum score estimator,[14] maximum rank correlation estimator,[15] nonparametric maximum likelihood estimator,[16] and its smoothed version along with the conventional probit ML estimator. My results are not directly comparable with Cosslett's, however, because we used different random number generators and different optimization methods. Cosslett's method of optimization is grid search, initially between -2.5 and -1.5 and further when the objective function is still improving at the boundaries. This particular method of optimization may choose values close to the truth more often than we really would when we did not know the truth. My experiment employs a different grid search method that treats different parameter values identically.

The first-stage grid search is done between -50 and 50 with grid width 1, and the seven values that performed best are selected. In the second stage, the grid search is done around the seven selected values with grid width 0.1, and the five values that performed best are selected. In the third stage, the grid search is done around the five selected values with grid width 0.01. The final stage is performed around the five selected values with grid width 0.001.

For a given sample size and hence for a given positive bandwidth, there always is a positive probability that the objective function is not differentiable, although the probability goes to 0 rather quickly. Therefore when we repeat trials 1000 times as we do here, the probability of the objective function being nondifferentiable at least for one trial is rather high. This is the reason we resort to the grid search method in the Monte Carlo study, although for a particular estimation problem the grid search method is an inefficient way to compute the optimum.

[14] See Manski (1975, 1985).

[15] See Han (1987).

[16] See Cosslett (1983).

Table 1

x_1 and x_2 normal – Performance of parametric and semiparametric estimators; a Monte Carlo experiment (250 observations, 1000 trials).

	Error 1		Error 2		Error 3	
Estimator[a]	Bias	RMSE[b]	Bias	RMSE[b]	Bias	RMSE[b]
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

[a] SLS = semiparametric least squares, MS = maximum score, MRC = maximum rank correlation, SML = smoothed maximum likelihood.

[b] Root mean square error.

Table 2

x_1 and x_2 exponential – Performance of parametric and semiparametric estimators; a Monte Carlo experiment (250 observations, 1000 trials).

	Error 1		Error 2		Error 3	
Estimator[a]	Bias	RMSE[b]	Bias	RMSE[b]	Bias	RMSE[b]
SLS	0.187	0.53	0.228	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

[a] SLS = semiparametric least squares, MS = maximum score, MRC = maximum rank correlation, SML = smoothed maximum likelihood.

[b] Root mean square error.

An IBM 8760 was used for the computation. Each calculation took about 1.8 seconds of cpu time. Computational speed with this algorithm increases roughly with the square of the sample size. The results are presented in tables 1 and 2. The results for other estimators are from Cosslett (1986).

Ruud (1983) showed that probit ML estimators are consistent when regressors are jointly normal. Thus, MLE in table 1 is consistent and asymptotically normal. Furthermore, clearly the probit estimator in table 2 with error 1 is consistent and efficient. Therefore error 2 and 3 in table 2 are the only cases where the probit estimator is not consistent. In those two cases, the SLS estimator performs the best in terms of the estimated mean square error. Note that optimal weighting is not used in the experiment. Even when the SLS estimator does not perform the best, the ratios of its MSEs to those of the best cases do not go below 0.64. The MSEs lie between 0.45 and 0.70. Compared with those of the other estimators, the estimated mean square errors of the SLS estimator are less affected by the differences in distributions.

9. Concluding remarks

We have established $1/\sqrt{n}$-consistency and asymptotic normality of the SLS and WSLS estimators for the single-index model. A consistent estimator of the covariance matrix was also presented. Since SLS estimation does not require specifying a parametric error distribution, the method allows economists to focus on specifying systematic effects of an econometric model and frees them from distributional worries for a broader class of models than before. We also investigated a weighting scheme that achieves the semiparametric efficiency bound obtained by Newey (1990) for the single-index model.

While the results extend the applicability of semiparametric estimation, a number of related issues were not addressed. Some of these issues are analyzed in other papers. In this section I summarize the results of these other papers and point out some problems that have not been investigated.

Ichimura and Lee (1991) studied an extension to the multiple-index model. Hall and Ichimura (1991) analyzed an extension of the single-index model to a general moment condition other than conditional mean zero.

The estimators presented in this paper treat a particular sequence of bandwidths and a kernel function as given. A practical implication is that any sequence of bandwidths and kernel function that satisfy certain assumptions will give rise asymptotically to the same estimators. As we showed, the choice does not affect the asymptotic distribution, and hence we are left with an array of estimators.

One approach to the bandwidth selection problem is to define an estimator that is independent of the choice of bandwidth sequence. For example, one may choose the bandwidth, $\hat{a}$, and the estimator, $\hat{\theta}$, that minimize the objective function. This method is investigated by Haerdle, Hall, and Ichimura (1991), who show that if the optimization is restricted to the $1/\sqrt{n}$ neighborhood of θ_0 for θ and to $[C_1 n^{-1/5}, C_2 n^{-1/5}]$ for a, where C_1 and C_2 are any given real number that satisfy $C_1 < C_2$, then $\hat{a}/a_0 \rightarrow 1$ asymptotically as $n \rightarrow \infty$, where a_0 is the optimal bandwidth for estimating φ when θ_0 is known. Whether there is a way to choose a bandwidth sequence that is optimal for the estimation of θ_0 is an open question.

Another potential problem with the method is computational burden. The computation time is roughly n times more than with smooth parametric nonlinear regression estimation, where n is the sample size. While the present method requires considerably fewer assumptions regarding the shape of the error distribution, the asymptotic properties are derived assuming that the systematic part is correctly specified, as usual. In practice, since empirical research requires trial and error before the final specification is reached, the computational burden implies more casual specification of the systematic part, which might lead to a larger bias than would the casual specification of the parametric error term that we successfully avoided. Therefore it would be desirable to have a method with less computational burden.[17]

Another approach can be taken to the basic problem studied in this paper, namely, that forms of the error distributions in econometric models are too often casually assumed without any justification. Rather than asking what the error terms are and how they might be distributed in a specific context, I proposed an estimator that does not require knowledge of the error distribution in a model; in other words, I took a semiparametric approach. We could instead have faced the problem directly and tried to model the 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. This alternative approach has produced the Gaussian distribution, exponential distribution, Wiener process, and Poisson process in other disciplines. Although the two approaches are very different in their attitude toward the error terms, ultimately they should be complementary, for the semiparametric approach offers a way to test the assumptions behind the derived distribution while holding an alternative set of estimates ready should the specification be rejected.

Appendix

In this appendix we prove Lemma 5.1, 5.6, 5.7, 5.10, 5.8, 5.9, and 5.5 in this order after establishing some preliminary lemmas.

Since $\hat{\mathrm{E}}_W(x_i,\theta)$ can be written in the ratio form, A_{ni}/B_{ni}, where

$$A_{ni}=\frac{1}{(n-1)a_n}\sum_{j\neq i} y_j W(x_j)I(x_j\in X_n)K([h(x_i;\theta)-h(x_j;\theta)]/a_n),$$

$$B_{ni}=\frac{1}{(n-1)a_n}\sum_{j\neq i} W(x_j)I(x_j\in X_n)K([h(x_i;\theta)-h(x_j;\theta)]/a_n),$$

[17] Of course this point may be moot given a fast computational technology, and it may be risky to stress the point too much; as Feller (1968) writes: 'Only yesterday the practical things of today were decried as impractical, and the theories which will be practical tomorrow will always be branded as valueless games by the practical men of today.'

uniform convergence of $\hat{\mathrm{E}}_W(x_i, \theta)$, $\partial \hat{\mathrm{E}}_W(x_i, \theta)/\partial\theta$, and $\partial^2 \hat{\mathrm{E}}_W(x_i, \theta)/\partial\theta\partial\theta'$ will be proven by showing uniform convergence of A_{ni} and B_{ni} and their first and second derivatives with respect to θ. Since B_{ni} is A_{ni} with $y_j = 1$, only A_{ni} and its first and second derivatives are considered. Let A_i be the probability limit of A_{ni} and note the inequality

$$\sup_{\theta\in\Theta}\left|A_{ni} - A_i\right| \le \sup_{\theta\in\Theta}\left|A_{ni} - \mathrm{E}(A_{ni})\right| + \sup_{\theta\in\Theta}\left|\mathrm{E}(A_{ni}) - A_i\right|.$$

We refer to the second term of the right-hand side and analogous expressions for the first and the second derivatives with respect to θ as *bias terms*. Note further that, for any sequence of positive numbers $\{M_n\}_{n=1}^{\infty}$,

$$\begin{aligned}\sup_{\theta\in\Theta}\left|A_{ni} - \mathrm{E}(A_{ni})\right| &\le \sup_{\theta\in\Theta}\left|\frac{1}{(n-1)a_n}\sum_{j\neq i} g_{nj}(x_i, \theta) - \mathrm{E}[g_{nj}(x_i, \theta)]\right| \\ &\quad + \sup_{\theta\in\Theta}\left|\frac{1}{(n-1)a_n}\sum_{j\neq i} \tilde{g}_{nj}(x_i, \theta) - \mathrm{E}[\tilde{g}_{nj}(x_i, \theta)]\right|,\end{aligned} \tag{15}$$

where

$$\begin{aligned} g_{nj}(x, \theta) &= \frac{1}{a_n} W(x_j) I(x_j \in X_n) y_i I(y_i \in [-M_n, M_n]) \\ &\quad \times K([h(x;\theta) - h(x_j;\theta)]/a_n), \\ g_{nj}(x, \theta) &= \frac{1}{a_n} W(x_j) I(x_j \in X_n) y_i I(y_i \notin [-M_n, M_n]) \\ &\quad \times K([h(x;\theta) - h(x_j;\theta)]/a_n). \end{aligned}$$

We refer to the first and the second term of the right-hand side of inequality (15) and their analogous terms for the first and the second derivatives with respect to θ as *centered terms* and *tail terms*, respectively.

Lemmas A.2–A.4 show that the bias terms converge to 0 with rate a_n^2, Lemmas A.5–A.7 establish the convergence rates of the tail terms, and Lemmas A.8–A.10 show the convergence rates of the centered terms.

The following lemma, which can be proven using integration by parts formula, is used in the proofs of Lemmas A.2–A.4.

Lemma A.1. *Assumption* 5.6 *implies*

(1) $\int K'(s)\mathrm{d}s = 0$;

(2) $\int sK'(s)\mathrm{d}s = -1$;

(3) $\int s^2K'(s)\mathrm{d}s = 0$;

(4) $\int K''(s)\mathrm{d}s = 0$;

(5) $\int sK''(s)\mathrm{d}s = 0$;

(6) $\int s^2K''(s)\mathrm{d}s = 2$;

(7) $\int s^3K''(s)\mathrm{d}s = 0$.

The following three lemmas can be proven directly by change of variable formula and Taylor expansions. We prove Lemma A.4 only, because the other two can be proven analogously.

Lemma A.2. *Let f be a Lebesgue density of a random variable x and g be a function $g: R \to R$. When* $\mathrm{E}[g(x)/a_nK[(t-x)/a_n]]$ *exists, if $\psi = gf$ is twice continuously differentiable, the second derivative satisfy Lipschitz condition, a kernel function satisfies Assumption 5.6, and t is an interior point of the support of x, then for $a_n > 0$ and $a_n \to 0$,*

$$|\mathrm{E}[g(x)/a_nK[(t-x)/a_n]] - g(t)f(t)| = \mathrm{O}(a_n^2).$$

Lemma A.3. *Let f be a Lebesgue density of a random variable x and g be a function $g: R \to R$. When* $\mathrm{E}[g(x)/a_n^2K'[(t-x)/a_n]]$ *exists, if $\psi = gf$ is twice continuously differentiable, the second derivative satisfy Lipschitz condition, a kernel function satisfies Assumption 5.6, and t is an interior point of the support of x, then for $a_n > 0$ and $a_n \to 0$,*

$$|\mathrm{E}[g(x)/a_n^2K'[(t-x)/a_n]] - [g(t)f(t)]'| = \mathrm{O}(a_n^2).$$

Lemma A.4. *Let f be a Lebesgue density of a random variable x and g be a function $g: R \to R$. When* $\mathrm{E}[g(x)/a_n^3K''[(t-x)/a_n]]$ *exists, if $\psi = gf$ is three times continuously differentiable, the third derivative satisfy Lipschitz condition, a kernel function satisfies Assumption 5.6, and t is an interior point of the support of x, then for $a_n > 0$ and $a_n \to 0$,*

$$|\mathrm{E}[g(x)/a_n^3K''[(t-x)/a_n]] - [g(t)f(t)]''| = \mathrm{O}(a_n^2).$$

Proof. Note that

$$\left|\int_{-\infty}^{\infty} g(x)a_n^{-3}K''[(t-x)/a_n]f(x)\,\mathrm{d}x - [g(t)f(t)]''\right|$$

$$= \left|\int_{-\infty}^{\infty} \psi(t-a_ns)a_n^{-2}K''(s)\,\mathrm{d}s - \psi''(t)\right|. \qquad (16)$$

By Taylor expansion, for some value $\bar{t}$ between t and $t - a_ns$,

$$\psi(t-a_ns) = \psi(t) - a_n\psi'(t)s + a_n^2/2\psi''(t)s^2 - a_n^3/6\psi'''(\bar{t})s^3.$$

Therefore eq. (16) equals

$$\left|\int_{-\infty}^{\infty} [\psi(t) - a_n\psi'(t)s + a_n^2/2\psi''(t)s^2 - a_n^3/6\psi'''(t)s^3\right.$$

$$\left. + a_n^3/6[\psi'''(\bar{t}) - \psi'''(t)s^3]a_n^{-2}K''(s)]\,\mathrm{d}s - \psi''(t)\right|.$$

Lemma A.1, a Lipschitz condition on ψ''', and the fact that $|\bar{t} - t| \leq a_n|s|$ imply that there exists a constant c such that the last expression is not greater than

$$ca_n^2\int_{-\infty}^{\infty} |s^4K''(s)| = \mathrm{O}(a_n^2). \quad ■$$

Lemma A.5. Let $m \geq 2$ be the highest absolute moment of y and suppose Assumptions 5.4 and 5.6 hold. If $\varepsilon_{0n}a_nM_n^{m-1} \to \infty$, then

$$\Pr\left\{\sup_{(x,\theta)\in X\times\Theta}\left|\frac{1}{na_n}\sum_{i=1}^{n}\tilde{g}_{ni}(x,\theta) - \mathrm{E}[\tilde{g}_{ni}(x,\theta)]\right| > \varepsilon_{0n}\right\} \to 0,$$

where

$$\tilde{g}_{ni}(x,\theta) = y_iW(x_i)I(x_i\in X_n)I(y_i \notin [-M_n, M_n])$$

$$\times K([h(x;\theta) - h(x_i,\theta)]/a_n).$$

Proof. Note that

$$\Pr\left\{\sup_{(x,\theta)\in X\times\Theta}\left|\frac{1}{na_n}\sum_{i=1}^{n}\tilde{g}_{ni}(x,\theta) - \mathrm{E}[\tilde{g}_{ni}(x,\theta)]\right| > \varepsilon_{0n}\right\}$$

$$\le \Pr\left\{\sum_{i=1}^{n}\sup_{(x,\theta)\in X\times\Theta}|\tilde{g}_{ni}(x,\theta)-\mathrm{E}[\tilde{g}_{ni}(x,\theta)]|>na_n\varepsilon_{0n}\right\}$$

$$\le 2\mathrm{E}\left[\sup_{(x,\theta)\in X\times\Theta}\tilde{g}_{ni}(x,\theta)\right]\Big/(a_n\varepsilon_{0n})$$

$$\le C\mathrm{E}[|y_i|I(y_i\notin[-M_n,M_n])]/(a_n\varepsilon_{0n}),$$

where C is some large constant. By Hölder's inequality and Chebyschev's inequality,

$$\begin{aligned}\mathrm{E}[y_iI(y_i\notin[-M_n,M_n])] &\le [\mathrm{E}(|y_i|^m)]^{1/m}[\Pr\{|y_i|\ge M_n\}]^{1-1/m}\\ &\le [\mathrm{E}(|y_i|^m)]/M_n^{m-1}.\end{aligned}$$

This implies the result. ■

Proofs of the next two lemmas are analogous to the previous one and hence omitted.

Lemma A.6. Let $m\ge$ *be the highest absolute moment of* y *and suppose Assumptions 5.4 and 5.6 hold. If* $H(x,x_i,\theta)$ *is uniformally bounded on* $X\times X\times\Theta$ *and* $\varepsilon_{1n}a_n^2M_n^{m-1}\to\infty$, *then*

$$\Pr\left\{\sup_{(x,\theta)\in X\times\Theta}\left|\frac{1}{na_n^2}\sum_{i=1}^{n}\tilde{g}^1_{ni}(x,\theta)-\mathrm{E}[\tilde{g}^1_{ni}(x,\theta)]\right|>\varepsilon_{1n}\right\}\to 0,$$

where

$$\begin{aligned}\tilde{g}^1_{ni}(x,\theta)=y_iW(x_i)I(x_i\in X_n)I(y_i\notin[-M_n,M_n])H(x,x_i,\theta)\\ \times K'([h(x;\theta)-h(x_i,\theta)]/a_n).\end{aligned}$$

Lemma A.7. Let $m\ge 2$ *be the highest absolute moment of* y *and suppose Assumptions 5.4–5.6 hold. If* $H(x,x_i,\theta)$ *is uniformally bounded on* $X\times X\times\Theta$ *and* $\varepsilon_{2n}a_n^3M_n^{m-1}\to\infty$, *then*

$$\Pr\left\{\sup_{(x,\theta)\in X\times\Theta}\left|\frac{1}{na_n^3}\sum_{i=1}^{n}\tilde{g}^2_{ni}(x,\theta)-\mathrm{E}[\tilde{g}^2_{ni}(x,\theta)]\right|>\varepsilon_{2n}\right\}\to 0,$$

where

$$\tilde{g}_{ni}^2(x,\theta) = y_i W(x_i) I(x_i \in X_n) I(y_i \notin [-M_n, M_n]) H(x, x_i, \theta)$$

$$\times K''([h(x;\theta) - h(x_i;\theta)]/a_n).$$

We use Bernstein's inequality to prove Lemmas A.8–A.10. For completeness we list the inequality here:

Bernstein's Inequality. *Let $Y_{1n}, \ldots, Y_{nn}$ be independent random variables with 0 means and bounded ranges, that is, $|Y_{in}| \le c_n$. Write σ_{in}^2 for the variance of Y_{in}. Suppose $V_n \ge \sigma_{1n}^2 + \cdots + \sigma_{nn}^2$. Then for each $\eta_n > 0$,*

$$\Pr\{|Y_{1n} + \cdots + Y_{nn}| > \eta_n\} \le \exp[\tfrac{1}{2}\eta_n^2/(V_n + \tfrac{1}{3}c_n\eta_n)].$$

Lemma A.8. *Suppose $(\log a_n)(1 + M_n\varepsilon_{0n})/(na_n\varepsilon_{0n}^2) \to 0$, $a_n^v/\varepsilon_{0n} \to 0$ for large v and $X \subset R^K$ and $\Theta \subset R^M$ are compact, then*

$$\Pr\left\{\sup_{X\times\Theta}\left|\sum_{i=1}^n [g_{ni}(x,\theta) - \mathrm{E}[g_{ni}(x,\theta)]]\right| \ge \varepsilon_{0n}\right\} \to 0,$$

where

$$g_{ni}(x,\theta) = \frac{1}{a_n} W_i I(x_i \in X_n) y_i I(y_i \in [-M_n, M_n])$$

$$\times K([h(x;\theta) - h(x_i;\theta)]/a_n).$$

Proof. Without loss of generality assume that $|\theta| \le 1$ for all $\theta \in \Theta$ and $|x| \le 1$ for all $x \in X$. Partition Θ into N_1 cubes with the length of a side $a_n^v\delta$ and X into N_2 cubes with the length of a side $a_n^v\delta$, where δ is a small and positive number and v is a large constant. Then $N_1 = \delta^{-M}(a_n^{-Mv})$ and $N_2 = \delta^{-K}(a_n^{-Kv})$ and space $X \times \Theta$ is partitioned into $N = N_1 \times N_2 = \delta^{-(M+K)}a_n^{-(M+K)v}$ of $(K+M)$-dimensional cubes, B_k^N, for $k = 1, \ldots, N$, which become smaller and smaller as n becomes larger. Now pick a point (x_k^N, θ_k^N) from each B_k^N for $k = 1, \ldots, N$,

$$\Pr\left\{\sup_{X\times\Theta}\left|\sum_{i=1}^n [g_{ni}(x,\theta) - \mathrm{E}[g_{ni}(x,\theta)]]\right| > n\varepsilon_{0n}\right\}$$

$$\le \Pr\left\{\bigcup_{k=1}^N\left[\sup_{B_k^N}\left|\sum_{i=1}^n [g_{ni}(x,\theta) - \mathrm{E}[g_{ni}(x,\theta)]]\right| > n\varepsilon_{0n}\right]\right\}$$

$$\leq \sum_{k=1}^{N} \Pr\left\{\sup_{B_k^N}\left|\sum_{i=1}^{n} [g_{ni}(x,\theta) - \mathrm{E}[g_{ni}(x,\theta)]]\right| > n\varepsilon_{0n}\right]\Bigg\}$$

$$\leq \sum_{k=1}^{N} \Pr\left\{\left|\sum_{i=1}^{n} [g_{ni}(x_k^N,\theta_k^N) - \mathrm{E}[g_{ni}(x_k^N,\theta_k^N)]]\right| > \frac{n\varepsilon_{0n}}{2}\right\} \tag{17}$$

$$+ \sum_{k=1}^{N} \Pr\left\{\sup_{B_k^N}\left|\sum_{i=1}^{n} [g_{ni}(x,\theta) - g_{ni}(x_k^N,\theta_k^N)]\right| > \frac{n\varepsilon_{0n}}{4}\right\} \tag{18}$$

$$+ \sum_{k=1}^{N} \Pr\left\{\sup_{B_k^N}\left|\sum_{i=1}^{n} [\mathrm{E}[g_{ni}(x,\theta)] - \mathrm{E}[g_{ni}(x_k^N,\theta_k^N)]]\right| > \frac{n\varepsilon_{0n}}{4}\right\}. \tag{19}$$

We show that terms (17)–(19) all converge to 0. To show that term (17) converges to 0, note that

$$\Pr\left\{\left|\sum_{i=1}^{n} [g_{ni}(x_k^N,\theta_k^N) - \mathrm{E}[g_{ni}(x_k^N,\theta_k^N)]]\right| > \frac{n\varepsilon_{0n}}{2}\right\}$$

$$= \Pr\Bigg\{\Bigg|\sum_{i=1}^{n} [W_i I_{ni} y_i I(y_i \in [-M_n, M_n]) K([h(x_k^N,\theta_k^N) - h(x_i,\theta_k^N)]/a_n)$$

$$- \mathrm{E}[y_i W_i I_{ni} I(y_i \in [-M_n, M_n])$$

$$\times K([h(x_k^N,\theta_k^N) - h(x_i,\theta_k^N)]/a_n)]\Bigg]\Bigg| > \frac{na_n\varepsilon_{0n}}{2}\Bigg\}. \tag{20}$$

Applying Bernstein's inequality with

$$\eta_n = na_n\varepsilon_{0n}/2, \qquad c_n = 2M_nK_1, \qquad V_n = na_nK_2,$$

where K_1 and K_2 are some positive constants, the right-hand side of eq. (20) is bounded by

$$2\exp[-K_3na_n\varepsilon_{0n}^2/(1 + M_n\varepsilon_{0n})],$$

where K_3 is some positive constant. Therefore term 17 is bounded by

$$2N\exp[-(K_3na_n\varepsilon_n^2)/(1 + M_n\varepsilon_{0n})]$$

$$= 2\delta^{-(M+K)}a_n^{-(M+K)v}\exp[-(K_3na_n\varepsilon_{0n}^2)/(1 + M_n\varepsilon_{0n})]$$

$$= 2\delta^{-(M+K)}\exp[-(M + K)v\log a_n - (K_3na_n\varepsilon_{1n}^2)/(1 + M_n\varepsilon_{0n})]. \tag{21}$$

Hence term (17) converges to 0 if $-(\log a_n)(1+M_n\varepsilon_{0n})/(na_n\varepsilon_{0n}^2)=\mathrm{o}(1)$, which holds under our assumption.

To show that term (18) converges to 0, note that

$$\Pr\left\{\sup_{B_k^N}\left|\sum_{i=1}^{n}\left[g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)\right]\right|>\frac{n\varepsilon_{0n}}{4}\right\}$$

$$\leq \Pr\left\{\sum_{i=1}^{n}\sup_{B_k^N}\left|g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)]\right|>\frac{n\varepsilon_{0n}}{4}\right\}$$

$$\leq \Pr\left\{\left|\sum_{i=1}^{n}\sup_{B_k^N}|g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)|\right.\right.$$

$$\left.\left.-\,\mathrm{E}\left[\sup_{B_k^N}|g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)|\right]\right|>\frac{n\varepsilon_{0n}}{8}\right\} \tag{22}$$

$$+\Pr\left\{\mathrm{E}\left[\sup_{B_k^N}|g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)|\right]>\frac{\varepsilon_{0n}}{8}\right\}. \tag{23}$$

Note that by Markov's inequality, term (23) is less than $K_4a_n^{v-1}/\varepsilon_{0n}$, for some large constant number K_4. Thus when $a_n^{v-1}/\varepsilon_{0n}=\mathrm{o}(1)$, term (23) is 0 for large enough n.

To show that term (22) converges to 0, use Bernstein's inequality again, this time with

$$\eta_n=n\varepsilon_{0n}a_n/8,\qquad c_n=M_nc\delta\cdot a_n^{v-1},\qquad V_n=nc\delta a_n^{2(v-1)},$$

for large $c>0$. Then term (22) is less than

$$2N\exp\left[(-K_5na_n\varepsilon_{0n}^2)/(a_n^{2(v-3)}+M_na_n^{v-1}\varepsilon_{0n}\right], \tag{24}$$

where K_5 is some positive constant number. When $v>4$, term (24) is smaller than term (21) asymptotically, and hence it converges to 0.

To show that term (19) converges to 0 just note the inequality

$$\Pr\left\{\sup_{B_k^N}|\mathrm{E}\left[g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)\right]|>\varepsilon_{0n}\right\}$$

$$\leq \Pr\left\{\mathrm{E}\left[\sup_{B_k^N}|g_{ni}(x,\theta)-g_{ni}(x_k^N,\theta_k^N)|\right]>\varepsilon_{0n}\right\}. \tag{25}$$

By the same approach taken for proving that term (23) converges to 0, term (25) converges to 0. This completes the proof. ■

The proof of the following two lemmas follows the same arguments and is thus omitted.

Lemma A.9. Suppose $(\log a_n)(1 + M_n\varepsilon_{1n})/(na_n^2\varepsilon_{1n}^2) \to 0$, $a_n^v/\varepsilon_{1n} \to 0$ *for large* v, *and* $X \subset R^K$ *and* $\Theta \subset R^M$ *are compact, then*

$$\Pr\left\{\sup_{X\times\Theta}\left|\sum_{i=1}^{n} [g_{ni}(x,\theta) - \mathrm{E}[g_{ni}(x,\theta)]]\right| \geq \varepsilon_{1n}\right\} \to 0,$$

where

$$g_{ni}(x,\theta) = \frac{y_i}{a_n^2} I(x_i, \in X_n) I(y_i \in [-M_n, M_n]) H(x, x_i, \theta)$$

$$\times K'([h(x;\theta) - h(x_i;\theta)]/a_n).$$

Lemma A.10. Suppose $(\log a_n)(1 + M_n\varepsilon_{2n})/(na_n^3\varepsilon_{2n}^2) \to 0$, $a_n^v/\varepsilon_{2n} \to 0$ *for large* v, *and* $X \subset R^K$ *and* $\Theta \subset R^M$ *are compact, then*

$$\Pr\left\{\sup_{X\times\Theta}\left|\sum_{i=1}^{n} [g_{ni}(x,\theta) - \mathrm{E}[g_{ni}(x,\theta)]]\right| \geq \varepsilon_{2n}\right\} \to 0,$$

where

$$g_{ni}(x,\theta) = \frac{y_i}{a_n^3} I(x_i, \in X_n) I(y_i \in [-M_n, M_n]) H(x, x_i, \theta)$$

$$\times K''([h(x;\theta) - h(x_i;\theta)]/a_n).$$

We now turn to the proofs of the lemmas in the text.

Proof of Lemma 5.1

By Lemma A.2, A.5, and A.8, we need to show that there exists a positive sequence $\{M_n\}$ that satisfies $a_n M_n^{m-1} \to \infty$ and $(-\log a_n)M_n/(na_n) \to 0$. For any positive sequence $\{b_n\}$ which converges to 0, $M_n = na_nb_n/(-\log a_n)$ satisfies the second condition. For this sequence $\{M_n\}$, the first condition is satisfied if

$$na_n^{1+1/(m-1)}b_n/(-\log a_n) \to \infty.$$

But if we set $b_n = [(-\log a_n)/na_n^{1+1/(m-1)}]^{1/2}$, under our assumption $b_n \to 0$ and the first condition is also satisfied. This completes the proof.

Proof of Lemma 5.6

By Lemmas A.3, A.6, and A.9, we need to show that there exists a positive sequence $\{M_n\}$ that satisfies $a_n^2 M_n^{m-1} \to \infty$ and $(-\log a_n)M_n/(na_n^2) \to 0$. For any positive sequence $\{b_n\}$ which converges to 0, $M_n = na_n^2 b_n/(-\log a_n)$ satisfies the second condition. For this sequence $\{M_n\}$, the first condition is satisfied if

$$na_n^{2+2/(m-1)} b_n/(-\log a_n) \to \infty.$$

But if we set $b_n = [(-\log a_n)/na_n^{2+2/(m-1)}]^{1/2}$, under our assumption $b_n \to 0$ and the first condition is also satisfied. This completes the proof.

Proof of Lemma 5.7

By Lemmas A.4, A.7, and A.10, we need to show that there exists a positive sequence $\{M_n\}$ that satisfies $a_n^3 M_n^{m-1} \to \infty$ and $(-\log a_n)M_n/(na_n^3) \to 0$. For any positive sequence $\{b_n\}$ which converges to 0, $M_n = na_n^3 b_n/(-\log a_n)$ satisfies the second condition. For this sequence $\{M_n\}$, the first condition is satisfied if

$$na_n^{3+3/(m-1)} b_n/(-\log a_n) \to \infty.$$

But if we set $b_n = [(-\log a_n)/na_n^{3+3/(m-1)}]^{1/2}$, under our assumption $b_n \to 0$ and the first condition is also satisfied. This completes the proof.

Proof of Lemma 5.10

By Lemmas A.2 and A.3 we need to show that there exists positive sequences a_n, $\{\varepsilon_{0n}\}$, and $\{\varepsilon_{1n}\}$ that satisfy

(i) $\sqrt{n}a_n^4 \to 0$,

(ii) $\sqrt{n}\varepsilon_{0n}a_n^2 \to 0$,

(iii) $\sqrt{n}\varepsilon_{1n}a_n^2 \to 0$,

(iv) $\sqrt{n}\varepsilon_{0n}\varepsilon_{1n} \to 0$.

By Lemmas A.5, A.6, A.8, and A.9, this amounts to finding positive sequences $\{M_n\}$, $\{\varepsilon_{0n}\}$, and $\{\varepsilon_{1n}\}$ that satisfy the four conditions above and

(v) $\varepsilon_{0n} a_n M_n^{m-1} \to \infty$,

(vi) $\varepsilon_{1n} a_n^2 M_n^{m-1} \to \infty$,

(vii) $(-\log a_n)/(n a_n \varepsilon_{0n}^2) \to 0$,

(viii) $(-\log a_n) M_n/(n a_n \varepsilon_{0n}) \to 0$,

(ix) $(-\log a_n)/(n a_n^2 \varepsilon_{1n}^2) \to 0$,

(x) $(-\log a_n) M_n/(n a_n^2 \varepsilon_{1n}) \to 0$.

To see that there exist such sequences let

$$\varepsilon_{0n} = \sqrt{\frac{-\log a_n}{n a_n}}\, b_{1n}, \qquad \varepsilon_{1n} = \sqrt{\frac{-\log a_n}{n a_n^2}}\, b_{1n}, \qquad M_n = \sqrt{\frac{n a_n^2}{-\log a_n}}\, b_{1n}^u,$$

where $0 < u < 1$ and b_{1n} is a positive sequence that diverges to infinity. Then ε_{0n}, ε_{1n}, and M_n satisfy conditions (v)–(x) since $n a_n^{2+2/(m-2)}/(-\log a_n) \to \infty$. The first four conditions are satisfied by taking b_{1n} to diverge slower than $\mathrm{o}([a_n \sqrt{-\log a_n}]^{-1})$ and $\mathrm{o}([n a_n^3/(-\log a_n)]^{1/4})$. This completes the proof.

Proof of Lemma 5.8

Note first that

$$\frac{\partial \hat{\mathrm{E}}_W(x_i, \theta_0)}{\partial \theta} = \frac{\partial A_{ni}/\partial \theta}{B_{ni}} - \frac{A_{ni}}{B_{ni}} \frac{\partial B_{ni}/\partial \theta}{B_{ni}},$$

$$\frac{\partial \hat{\mathrm{E}}_W(x_i, \theta_0)}{\partial \theta} = \frac{\partial A_i/\partial \theta}{B_i} - \frac{A_i}{B_i} \frac{\partial B_i/\partial \theta}{B_i}.$$

Thus

$$\frac{\partial \hat{\mathrm{E}}_W(x_i, \theta_0)}{\partial \theta} - \frac{\partial \mathrm{E}_W(x_i, \theta_0)}{\partial \theta}$$

$$= \frac{\partial A_{ni}/\partial \theta}{B_{ni}} - \frac{\partial A_i/\partial \theta}{B_i} - \frac{A_{ni}}{B_{ni}} \frac{\partial B_{ni}/\partial \theta}{B_{ni}} + \frac{A_i}{B_i} \frac{\partial B_i/\partial \theta}{B_i}.$$

Therefore, to prove the lemma, it is sufficient to show that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i \left[\frac{\partial A_{ni}/\partial\theta}{B_{ni}} - \frac{\partial A_i/\partial\theta}{B_i}\right]\right| = \mathrm{o}_p(1), \tag{26}$$

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i \left[\frac{A_{ni}}{B_{ni}}\frac{\partial B_{ni}/\partial\theta}{B_{ni}} - \frac{A_i}{B_i}\frac{\partial B_i/\partial\theta}{B_i}\right]\right| = \mathrm{o}_p(1). \tag{27}$$

Eqs. (26) and (27) can be shown analogously and hence we show only (26). Note that by Taylor expansion, for some value $\bar{B}_i$ between B_i and B_{ni},

$$\begin{aligned}\frac{\partial A_{ni}/\partial\theta}{B_{ni}} - \frac{\partial A_i/\partial\theta}{B_i} &= \left[\frac{\partial A_{ni}}{\partial\theta} - \frac{\partial A_i}{\partial\theta}\right]\frac{1}{B_i} + \left[\frac{\partial A_{ni}}{\partial\theta} - \frac{\partial A_i}{\partial\theta}\right]\left[\frac{1}{B_{ni}} - \frac{1}{B_i}\right] \\ &\quad + \frac{\partial A_i}{\partial\theta}\left[\frac{1}{B_{ni}} - \frac{1}{B_i}\right] \\ &= \left[\frac{\partial A_{ni}}{\partial\theta} - \frac{\partial A_i}{\partial\theta}\right]\frac{1}{B_i} + \left[\frac{\partial A_{ni}}{\partial\theta} - \frac{\partial A_i}{\partial\theta}\right]\left[\frac{1}{B_{ni}} - \frac{1}{B_i}\right] \\ &\quad - \frac{\partial A_i}{\partial\theta}\frac{1}{B_i^2}[B_{ni} - B_i] + \frac{\partial A_i}{\partial\theta}\frac{1}{\bar{B}_i^3}[B_{ni} - B_i]^2.\end{aligned}$$

Hence, to show (26), it is sufficient to show

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i \left[\frac{\partial A_{ni}}{\partial\theta} - \frac{\partial A_i}{\partial\theta}\right]\frac{1}{B_i}\right| = \mathrm{o}_p(1), \tag{28}$$

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i \left[\frac{\partial A_{ni}}{\partial\theta} - \frac{\partial A_i}{\partial\theta}\right]\left[\frac{1}{B_{ni}} - \frac{1}{B_i}\right]\right| = \mathrm{o}_p(1), \tag{29}$$

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i \frac{(B_{ni} - B_i)}{B_i^2}\frac{\partial A_i}{\partial\theta}\right] = \mathrm{o}_p(1), \tag{30}$$

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i \frac{1}{\bar{B}_i^3}[B_{ni} - B_i]^2\frac{\partial A_i}{\partial\theta}\right| = \mathrm{o}_p(1). \tag{31}$$

Eqs. (29) and (31) can be proven in the same way we proved Lemma 5.10. To show eqs. (28) and (30), we calculate the means and the variances explicitly.

Clearly, the means of the terms inside the norms are 0. To calculate the variance of the left-hand side of eq. (28), note that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i I_i \varepsilon_i\left[\frac{\partial A_{ni}}{\partial\theta}-\frac{\partial A_i}{\partial\theta}\right]\frac{1}{B_i}=\frac{1}{(n-1)\sqrt{n}}\frac{1}{a_n^2}\sum_{i=1}^{n}\sum_{j\neq i}\varepsilon_i\psi_{ij}^{(n)},$$

where

$$\psi_{ij}^{(n)}=\frac{W_iI_i}{B_i}\left[y_jW_jI_{nj}\left[\frac{\partial h_i}{\partial\theta}-\frac{\partial h_j}{\partial\theta}\right]K'\left(\frac{h(x_i;\theta_0)-h(x_j;\theta_0)}{a_n}\right)-a_n^2\frac{\partial A_i}{\partial\theta}\right].$$

The variance of

$$\frac{1}{(n-1)\sqrt{n}a_n^2}\sum_{i=1}^{n}\sum_{j\neq i}\varepsilon_i\psi_{ij}^{(n)}$$

equals

$$\frac{1}{(n-1)^2na_n^4}\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l\neq k}\varepsilon_i\varepsilon_k\psi_{ij}^{(n)}\psi_{kl}^{(n)}\right\}$$

$$=\frac{n-2}{(n-1)a_n^4}\mathrm{E}[\varepsilon_i^2\psi_{ij}^{(n)}\psi_{ik}^{(n)}]+\frac{1}{(n-1)a_n^4}\mathrm{E}[\varepsilon_i^2(\psi_{ij}^{(n)})^2]$$

$$+\frac{1}{(n-1)a_n^4}\mathrm{E}[\varepsilon_i\varepsilon_j\psi_{ij}^{(n)}\psi_{ji}^{(n)}],$$

where i, j, and k are all different. By Lemma A.3, since i, j, and k are all different,

$$\frac{1}{a_n^4}\mathrm{E}(\varepsilon_i^2\psi_{ij}^{(n)}\psi_{ik}^{(n)})=\mathrm{E}\left\{\varepsilon_i^2\mathrm{E}\left\{\frac{1}{a_n^2}\psi_{ij}^{(n)}\Big|\,i\right\}\mathrm{E}\left\{\frac{1}{a_n^2}\psi_{ik}^{(n)}\Big|\,i\right\}\right\}=\mathrm{O}(a_n^4).$$

If $a_n \to 0$ and $na_n^4 \to \infty$, then the variance converges to 0. Chebyshev's inequality then implies eq. (28). Eq. (30) can be shown analogously. This completes the proof.

Proof of Lemma 5.9

The proof is analogous to that of Lemma 5.8 and hence omitted.

Proof of Lemma 5.5

The density of $(h_{(1)}, \dots, h_{(n)})$ is

$$n!\, f_\theta(h_{(1)}) \cdots f_\theta(h_{(n)}) \quad \text{if} \quad \underline{h} < h_{(1)} - h_{(2)} < \cdots < h_{(n)} < \bar{h},$$

and 0 otherwise, where f_θ is the density of $h(x;\theta)$. Thus the density of $(s_1, \dots, s_n)$ is

$$n!\, f_\theta(s_1 + \underline{h}) f_\theta(s_1 + s_2 + \underline{h}) \cdots f_\theta(s_1 + s_2 + \cdots + s_n + \underline{h}) \quad \text{if} \quad s_i > 0,$$

for $i = 1, \dots, n$ and $s_1 + s_2 + \cdots + s_n < \bar{h} - \underline{h}$, and 0 otherwise. Therefore,

$$\begin{aligned}
\mathrm{E}(s_i^2) = {} & \int_0^{\bar{h}-\underline{h}} n!\, f_\theta(s_1 + \underline{h}) \int_0^{\bar{h}-\underline{h}-s_1} f_\theta(s_1 + s_2 + \underline{h}) \\
& \times \int_0^{\bar{h}-\underline{h}-s_1-s_2} \cdots \int_0^{\bar{h}-\underline{h}-s_1-s_2-\cdots-s_{i-1}} f_\theta(s_1 + \cdots + s_i + \underline{h}) s_i^2 \\
& \times \int_0^{\bar{h}-\underline{h}-s_1-s_2-\cdots-s_i} \cdots \int_0^{\bar{h}-\underline{h}-s_1-\cdots-s_{n-1}} f_\theta(s_1 + \cdots + s_n + \underline{h}) \\
& \times \mathrm{d}s_n \cdots \mathrm{d}s_1 \\
= {} & \int_0^{\bar{h}-\underline{h}} n!\, f_\theta(s_1 + \underline{h}) \int_0^{\bar{h}-\underline{h}-s_1} f_\theta(s_1 + s_2 + \underline{h}) \\
& \times \int_0^{\bar{h}-\underline{h}-s_1-s_2} \cdots \int_0^{\bar{h}-\underline{h}-s_1-\cdots-s_{i-1}} f_\theta(s_1 + \cdots + s_i + \underline{h}) s_i^2 \\
& \times \frac{1}{(n-i)!} [1 - F_\theta(s_1 + \cdots + s_i + \underline{h})]^{n-i} \,\mathrm{d}s_i \cdots \mathrm{d}s_1 .
\end{aligned}$$

By integration by parts,

$$\begin{aligned}
& \int_0^{\bar{h}-\underline{h}-s_1-\cdots-s_{i-1}} f_\theta(s_1 + \cdots + s_i + \underline{h}) s_i^2 \\
& \quad \times [1 - F_\theta(s_1 + \cdots + s_i + \underline{h})]^{n-i} \,\mathrm{d}s_i \\
& = \frac{2}{n-i+1} \int_0^{\bar{h}-\underline{h}-s_1-\cdots-s_{i-1}} [1 - F_\theta(s_1 + \cdots + s_i + \underline{h})]^{n-i+1} \, s_i \,\mathrm{d}s_i .
\end{aligned}$$

Approximate F_θ from below on $[0, \bar{h} - \underline{h} - s_1 - \cdots - s_{i-1}]$ by a continuous piecewise linear functions $c_k s_i + c'_k$ for $k = 1, \ldots, K < \infty$. Let t_k denote knots for $k - 1, \ldots, K - 1$. We can take $c'_1 = F_\theta(s_1 + \cdots + s_{i-1} + \underline{h})$ and $c_K(\bar{h} - \underline{h} - s_i - \cdots - s_{i-1}) + c'_K = 1$, $0 < c_k$, $c'_k < \infty$, because f_θ is bounded away from 0 and bounded from above. Then

$$\int_0^{\bar{h} - \underline{h} - s_1 - \cdots - s_{i-1}} [1 - F_\theta(s_1 + \cdots + s_i + \underline{h})]^{n-i+2} s_i \, \mathrm{d}s_i$$

$$\leq \int_0^{t_1} [1 - c_1 s_i - c'_1]^{n-i+1} s_i \, \mathrm{d}s_i + \int_{t_1}^{t_2} [1 - c_2 s_i - c'_2]^{n-i+1} s_i \, \mathrm{d}s_i$$

$$+ \cdots + \int_{t_{K-1}}^{\bar{h} - \underline{h} - s_1 - \cdots - s_{i-1}} [1 - c_K s_i - c'_K]^{n-i+1} s_i \, \mathrm{d}s_i.$$

Since each term of the right-hand side of the equation can be bounded by

$$\int_{t_j}^{t_{j+1}} [1 - c_{j+1} s_i - c'_{j+1}]^{n-i+1} s_i \, \mathrm{d}s_i$$

$$= -\frac{1}{c_{j+1}(n-i+2)} [1 - c_{j+1} s_i - c'_{j+1}]^{n-i+2} s_i \Big|_{t_j}^{t_{j+1}}$$

$$- \frac{1}{c_{j+1}^2 (n-i+2)(n-i+3)} [1 - c_{j+1} s_i - c'_{j+1}]^{n-i+3} \Big|_{t_j}^{t_{j+1}}.$$

Let $c^* = \max c_j$ and $c_* = \min c_j$, then

$$\int_{t_j}^{t_{j+1}} [1 - c_{j+1} s_i - c'_{j+1}]^{n-i+2} s_i \, \mathrm{d}s_i$$

$$\leq -\frac{1}{c_*(n-i+2)} [1 - c_{j+1} s_i - c'_{j+1}]^{n-i+2} s_i \Big|_{t_j}^{t_{j+1}}$$

$$- \frac{1}{c_*^2 (n-i+2)(n-i+3)} [1 - c_{j+1} s_i - c'_{j+1}]^{n-i+3} \Big|_{t_j}^{t_{j+1}}.$$

Summing over all knots we have

$$\frac{1}{c_*^2(n-i+2)(n-i+3)}[1-F_\theta(s_1+\cdots+s_{i-1}+\underline{h})]^{n-i+3}.$$

Completing the rest of the integral we have

$$\mathrm{E}(s_i^2) \le \frac{2}{c_*^2(n+1)(n+2)}.$$

References

Andrews, Donald W.K., 1987, Consistency in nonlinear econometric models: A generic uniform law of large numbers, Econometrica 55, 1465–1471.

Basmann, Robert L., 1957, A generalized classical method of linear estimation of coefficients in a structural equation, Econometrica 25, 77–83.

Brillinger, David R., 1983, A generalized linear model with 'Gaussian' regressor variables, in: Peter J. Bickel, Kjell A. Doksum, and J.L. Hodges, eds., A festschrift for Erich L. Lehmann (Woodsworth International Group, Belmont, CA).

Cosslett, Stephen R., 1983, Distribution-free maximum likelihood estimator of the binary choice model, Econometrica 51, 765–782.

Feller, William, 1968, An introduction to probability theory and its applications, Vol. 1, 3rd ed. (Wiley, New York, NY).

Haerdle, Wolfgang, Peter Hall, and Hidehiko Ichimura, 1991, Optimal smoothing in single index models, Discussion paper no. 9107 (CORE, Louvain-la-Neuve).

Hall, Peter and Hidehiko Ichimura, 1991, Optimal semi-parametric estimation in single-index models (Center for Mathematics and its Applications, Australian National University, Canberra).

Han, Aaron K., 1987, Non-parametric analysis of a generalized regression model, Journal of Econometrics 35, 303–316.

Ichimura, Hidehiko and Lung-Fei Lee, 1991, Semiparametric least squares estimation of multiple index models: Single equation estimation, in: William A. Barnett, James L. Powell, and George E. Tauchen, eds., Nonparametric and semiparametric methods in econometrics and statistics (Cambridge University Press, New York, NY).

Ichimura, Hidehiko and Whitney K. Newey, 1990, Efficiency of conditional expectation estimators in index models, Mimeo. (University of Minnesota, Minneapolis, MN and M.I.T., Cambridge, MA).

Manski, Charles F., 1975, The maximum score estimation of the stochastic utility model of choice, Journal of Econometrics 3, 205–228.

Manski, Charles F., 1985, Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator, Journal of Econometrics 27, 313–333.

McCullagh, P. and J. Nelder, 1983, Generalized linear models (Chapman-Hall, New York, NY).

Nadaraja, N., 1964, On regression estimator, Theory of Probability and Its Applications 9, 157–159.

Newey, Whitney K., 1990, Efficient estimation of semiparametric models via moment restrictions (Department of Economics, M.I.T., Cambridge, MA).

Parzen, Emanuel, 1962, On estimation of a probability density function and mode, Annals of Mathematical Statistics 33, 1065–1076.

Prakasa Rao, B.L.S., 1983, Nonparametric functional estimation (Academic Press, Orlando, FL).

Ruud, Paul A., 1983, Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution, Econometrica 51, 225–228.

Stoker, Thomas M., 1986, Consistent estimation of scaled coefficients, Econometrica 54, 1461–1481.

Theil, Henri, 1953, Estimation and simultaneous correlation in complete equation systems, Mimeo. (Central Planning Bureau, The Hague).

Theil, Henri, 1953, Repeated least-squares applied to complete equation systems, Mimeo. (Central Planning Bureau, The Hague).

Watson, G.S., 1964, Smooth regression analysis, Sankha Series A 26, 101–116.