DENSITY WEIGHTED LINEAR LEAST SQUARES

Whitney K. Newey* and Paul A. Ruud†

August, 1993
Revised June, 2001

1 Introduction

Several semi-parametric methods for index models have been developed. In a single index model, the conditional expectation of a dependent variable y given a $r \times 1$ vector of explanatory variables x is

$$\mathrm{E}[y \mid x] = \tau(x'\beta_0), \tag{1}$$

for an unknown vector of parameters β_0 and an unknown univariate function $\tau(\cdot)$. This model is implied by many important limited dependent variable and regression models, as discussed in Ruud (1986) and Stoker (1986). Consistent estimators for β_0, up to an unknown scale factor, have been developed by Ruud (1986), Stoker (1986), Powell, Stock, and Stoker (1989), Ichimura (1993), and others.

In this paper, we return to a type of estimator developed by Ruud (1986). He proposed an inverse-density-weighted quasi-maximum likelihood estimator. We consider least squares estimation that is weighted by the ratio of an elliptically symmetric density with compact support to a kernel estimator of the true density. We give conditions for $\sqrt{n}$-consistency and asymptotic normality of the estimator, and derive a consistent estimator for the asymptotic variance. We also show that the first-order conditions for the scaled least squares coefficients has an analogous form to the efficient score for an index model. This form is used to suggest ways to choose weights that have high efficiency.

Among the semi-parametric index estimators, the inverse-density-weighted least squares estimator is unique because it permits discontinuities in the transformation τ. Discontinuities in the conditional expectation of dependent variables arise in such economic problems as optimization over nonlinear budget sets and production frontiers. In labor supply for example, nonconvexities in

*Department of Economics, Massachusetts Institute of Technology.
†Department of Economics, University of California at Berkeley.

the budget frontier caused by welfare programs imply discontinuities in the desired hours of work. If the optimization errors are small, then these discontinuities translate into discontinuities in the conditional expectation of hours given socio-economic covariates that control for observable heterogeneity. The estimators that we consider in this paper accommodate such breaks when the index model is linear. In contrast, the average derivative estimators of Stoker (1986) and Powell, Stock, and Stoker (1989), and the kernel regression estimators of Ichimura (1993) all require that τ be differentiable. Thus, the results of this paper provide a way of estimating index parameters in nonsmooth cases that have previously been ruled out.

2 The Estimator

Our estimator is based on the idea of Ruud (1986). Suppose that the density has the linear conditional expectation (LCE) property that the conditional expectation of x given any linear combination of x is linear in that combination. Ruud (1986) showed that in this case quasi-maximum likelihood estimation (QMLE) is consistent for β_0, up to scale. He exploited this property by multiplying the quasi-likelihood function by the ratio of a LCE density to a nonparametric estimator of the true density of x. The resulting QMLE will be consistent for slope coefficients, because the "reweighting" has the effect of making the limit be the same as if the regressor density were the LCE density.

In this paper we focus on weighted least squares estimators, because they are particularly simple to compute. To describe the estimator, let $f(x,\theta)$ be an elliptically symmetric pdf, that has compact support and is parameterized by a vector θ. This density will be appropriate for the numerator of the weight, because it is well known that elliptically symmetric pdf's have the LCE property (see also the appendix). Let $\hat{\theta}$ denote an estimator of some value θ_0 of the parameter vector. For a kernel $K(u)$, satisyfing properties to be specified below, and a bandwidth parameter λ, let

$$\hat{h}(x) = \frac{1}{N}\sum_{i=1}^{n} \mathcal{K}_\lambda(x - x_i), \quad \mathcal{K}_\lambda(u) = \lambda^{-r}\mathcal{K}(u/\lambda),$$

where r is the dimension of x. This $\hat{h}(x)$ is a kernel density estimator. For $X = (1, x')'$, an inverse density weighted least square estimator is obtained as

$$\hat{\gamma} = \left(\sum_{i=1}^{n} \hat{w}_i X_i X_i'\right)^{-1} \sum_{i=1}^{n} \hat{w}_i X_i y_i, \quad \hat{w}_i = \frac{f(x_i,\hat{\theta})}{\hat{h}(x_i)}, \tag{2}$$

where the data observations are indexed by $i = 1, \ldots, n$.

The limit of this estimator will behave as if x had density $f(x,\theta_0)$. Thus, by Ruud (1986), we know that the coefficients of x in $\hat{\gamma}$ are consistent for β_0, up to a common scale factor. The density $f(x,\theta)$ is required to have compact support in order to deal with the technical problem that $\hat{h}(x)^{-1}$ could be large

for outlying values of x. Also, the parameter estimates $\hat{\theta}$ are present in order to allow for centering the location and scale of the density. Furthermore, allowing for $\hat{\theta}$ can be important for efficiency, as discussed in Section 4.

The kernel $\mathcal{K}(u)$ will be assumed to satisfy $\int \mathcal{K}(u)du = 1$, have a compact support, and satisfy other regularity conditions given below. It will also be assumed that $\mathcal{K}(u)$ is nonrandom, although in practice one would often use a scale normalization, where $\mathcal{K}(u) = \det(\hat{\Sigma})^{-1/2} p(\hat{\Sigma}^{-1/2}u)$ for a pdf $p(u)$, and $\hat{\Sigma}$ equal to the sample variance of x_i.

The estimator that will be consistent for β_0 up to scale is the coefficients of x that appear in $\hat{\gamma}$. A convenient way to normalize the scale is to suppose that the first coefficient in β_0 is 1 (which is just a normalization as long as it is nonzero). Partition $\gamma = (\gamma_1, \delta')$ and $\hat{\gamma} = (\hat{\gamma}_1, \hat{\delta}')'$ conformably, where γ_1 is a scalar (coefficient of the constant) and δ is a $r \times 1$ vector (the coefficients of x). Also, partition $\beta = (\beta_1, \beta_2')'$ and $\delta = (\delta_1, \delta_2')'$ conformably, where β_1 is a scalar, so that the dimension of β_2 is $r - 1$. The true value of β_1 is 1, by our scale normalization. An estimator of β_2 that includes this scale normalization is then

$$\hat{\beta}_2 = \hat{\delta}_2 / \hat{\delta}_1. \tag{3}$$

That is, $\hat{\beta}_2$ is the ratio of the coefficients in $\hat{\gamma}$ of all the regressors except the first one to the first regressor coefficient.

An important practical problem is the choice of bandwidth λ. The regularity conditions given below for $\sqrt{n}$-consistency will require that λ be chosen to be smaller than the value that would minimize the asymptotic mean square error of $\hat{h}$, a feature that is often referred to as "undersmoothing." Thus, choosing the bandwidth from cross-validation, or any other method that minimizes the asymptotic mean square error is not appropriate. It is beyond the scope of this paper to say much more about the theory of how to choose λ, but a practical method might be to start at a value obtained by cross-validation and decrease λ until $\hat{\beta}_2$ does not change much relative to its estimated standard error.

3 Asymptotic Variance Estimation

The estimator is a weighted least squares estimator with an estimated weight. In our case, where the conditional expectation (1) is not linear, estimation of the weights will affect the limiting distribution, complicating asymptotic variance estimation. There are two sources of variability in the weights, the nonparametric density estimator in the denominator and the $\hat{\theta}$ estimator in the numerator. Both sources will affect the asymptotic variance of $\hat{\gamma}$, but the asymptotic variance of $\hat{\beta}_2$ will only be affected by estimation of the denominator (the true density). This simplification follows from Newey and McFadden (1993, Theorem 6.2), which says that the asymptotic variance of $\hat{\beta}_2$ is not affected by estimation of θ if the limit of $\hat{\theta}$ does not affect the limit of $\hat{\beta}_2$. Here, $\hat{\beta}_2$ will be consistent no matter what the limit of $\hat{\theta}$ is, because of elliptical symmetry of $f(x, \theta)$ for *all* θ.

In most cases the parameters of interest are β_2, so that estimation of $\hat{\theta}$ can be ignored in the asymptotic variance. To avoid additional complication, we will focus on this case, by giving a consistent estimator of the asymptotic variance of $\hat{\beta}_2$.

An estimator of the asymptotic variance of $\hat{\beta}_2$ can be constructed as follows. Let

$$\hat{g}(x) = \sum_{i=1}^{n} \frac{y_i \mathcal{K}_\lambda(x - x_i)}{\hat{h}(x)}$$

be a kernel estimator of $\mathrm{E}[y \,|\, x]$. Define

$$\begin{aligned}
\hat{J} &\equiv \hat{\delta}_1^{-1}[0_{r-1}, -\hat{\beta}_2, I_{r-1}] \\
\hat{Q} &\equiv \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i X_i X_i' \\
\hat{\Sigma} &\equiv \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^2 X_i X_i' [y_i - \hat{g}(x_i)]^2,
\end{aligned}$$

where 0_{r-1} is a $r-1$ dimensional column vector of zeros and I_{r-1} is an $r-1$ dimensional identity matrix. Then a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta}_2 - \beta_{20})$ will be

$$\hat{V} = \hat{J}' \hat{Q}^{-1} \hat{\Sigma} \hat{Q}^{-1} \hat{J} \tag{4}$$

This estimator can be interpreted as being obtained by combining the delta-method with an asymptotic variance estimator for $\hat{\gamma}$. Here $\hat{J}$ is the Jacobian of the transformation from $\hat{\gamma}$ to $\hat{\beta}_2$, while $\hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}$ is an estimator for the asymptotic variance of $\hat{\gamma}$ that ignores estimation of θ_0. Consistency of this estimator of the asymptotic variance will be shown in Section 5.

The form of this estimator can be motivated by deriving the asymptotic variance of $\hat{\gamma}$, assuming that $\hat{\theta} = \theta_0$. Let $w(x) = f(x, \theta_0)/h_0(x)$, and $\gamma_0 = Q^{-1}\,\mathrm{E}[w(x)Xy]$ be the limit of $\hat{\gamma}$, for $Q = \mathrm{E}[w(x)XX']$. Then for $u = y - X'\gamma_0$,

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}} \hat{Q}^{-1} \sum_{i=1}^{n} \hat{w}_i X_i u_i. \tag{5}$$

Under appropriate regularity conditions, the first term will have limit Q^{-1}, so the asymptotic variance of $\hat{\gamma}$ will be $Q^{-1}\Sigma Q^{-1}$, where Σ is the asymptotic variance of $\sum_{i=1}^{n} \hat{w}_i X_i u_i/\sqrt{n}$. We can derive Σ using the results of Newey (1993, Propostion 5), which gives a general asymptotic variance formula when non-parametric density estimators are present. Let

$$D(x) = \mathrm{E}\left[f(x, \theta_0) X u \left. \frac{\partial(1/h)}{\partial h} \right|_{h=h_0(x)} \,\middle|\, x \right] \tag{6}$$

$$= -\frac{w(x) X \{\mathrm{E}[y \,|\, x] - X'\gamma_0\}}{h_0(x)}. \tag{7}$$

Then by $\mathrm{E}[D(x)h_0(x)] = -\mathrm{E}[w(x)Xu] = 0$ and Newey (1993),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{w}_i X_i u_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(x_i)X_i u_i \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{D(x_i)h_0(x_i) - \mathrm{E}[D(x)h_0(x)]\} + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(x_i)X_i\{y_i - \mathrm{E}[y_i \,|\, x_i]\} + o_p(1).\end{aligned}$$

This equation is given precise justification in Lemma 1 of Section 5. From this equation and the central limit theorem, the asymptotic variance of the term $\sum_{i=1}^{n}\hat{w}_i X_i u_i/\sqrt{n}$ will be $\Sigma = \mathrm{E}[w(x)^2 XX'\{y - \tau(x'\beta_0)\}^2]$. The estimator $\hat{\Sigma}$ that appears in $\hat{V}$ is simply a sample analogue of Σ, where $w(x)$ and $\mathrm{E}[y \,|\, x]$ have been replaced by estimators.

It is interesting to note that estimation of the density has the effect of lowering the asymptotic variance of the estimator. If the estimated density in the denominator were replaced by the true density, then Σ in the asymptotic variance would be replaced by the variance of $w(x)Xu$. Because Σ is the variance of $w(x)Xu - \mathrm{E}[w(x)Xu \,|\, x]$, it is smaller in the positive semi-definite sense than the variance of $w(x)Xu$.

4 Asymptotic Efficiency

The asymptotic efficiency of the estimator can be evaluated by comparing its asymptotic variance with the semiparametric variance bound for the index model of equation (1). It follows from the analysis of Section 3 that the asymptotic variance of $\sqrt{n}(\hat{\beta}_2 - \beta_{20})$ is $V = J'Q^{-1}\Sigma Q^{-1}J$ for $J = \delta_{10}^{-1}[0, -\beta_{20}, I]$. It is straightforward to derive a more convenient expression, as in $V = \mathrm{E}[\psi\psi']$, where $v = x'\beta_0$,

$$\psi = \delta_{10}^{-1}\left\{\mathrm{E}_w[\mathrm{Var}_w(x_2 \,|\, v)]\right\}^{-1} w(x)\left[x_2 - \mathrm{E}_w(x_2 \,|\, v)\right]\left[y - \tau(v)\right] \tag{8}$$

and $\mathrm{E}_w[\cdot] \equiv \mathrm{E}[w(x)(\cdot)]$: Details of this derivation are given in Lemma 3 in the Appendix. By way of comparison, the semiparametric variance bound for estimators of $\hat{\beta}_2$, as given by Newey and Stoker (1993), is $V^* = \mathrm{E}[\psi^*\psi^{*\prime}]$, where

$$\psi^* = \left\{\mathrm{E}_\sigma[\mathrm{Var}_\sigma(\tau_v x_2 \,|\, v)]\right\}^{-1}\sigma(x)^{-2}\tau_v(v)\left[x_2 - \mathrm{E}_\sigma(x_2 \,|\, v)\right]\left[y - \tau(v)\right] \tag{9}$$

and $\sigma^2(x) = \mathrm{Var}(y \,|\, x)$,

$$\mathrm{E}_\sigma[\cdot] \equiv \frac{\mathrm{E}[(\cdot)/\sigma^2(x)]}{\mathrm{E}[1/\sigma^2(x)]},$$

and $\tau_v(v) = d\tau(v)/dv$ (assuming differentiability holds).

The formulas (8) and (9) are analogous but fundamentally different. First of all, the weight $w(x)$ in $\mathrm{E}_w[\cdot]$ is replaced by $1/\sigma^2(x)$. The weighting by $1/\sigma^2(x)$ in

the variance bound accounts for heteroskedasticity, while the weighting by $w(x)$ is necessary for consistency of the WLS estimator. In addition, the efficiency bound contains the Jacobian term $\tau_v(x'\beta_0)$, which is not present in the WLS case, effectively replacing x_2 with $\tau_v x_2$. It is possible to extend this analysis to a nonlinear least squares framework that would permit us to introduce analogous terms. A good choice of the nonlinear regression function would be likely to improve the efficiency of the WLS estimator.

5 Asymptotic Normality

This section presents regularity conditions for asymptotic normality and consistency of the asymptotic variance estimator. We first derive a useful intermediate result, on the asymptotic distribution of a sample average that is weighted by the inverse of a kernel density estimator. This result justifies the asymptotic variance calculation given in Section 3.

To obtain results it is useful to impose certain conditions on the kernel, the density, and the bandwidth.

Assumption 1 *$\mathcal{K}(u)$ is Lipschitz, zero outside a bounded set, $\int \mathcal{K}(u)du = 1$, and there is a positive integer s such that for all r-tuples of nonnegative integers $(j_1, \ldots, j_r)$ with $\sum_{\ell=1}^r j_\ell < s$,*

$$\int \mathcal{K}(u)[\sum_{\ell=1}^{r}(u_\ell)^{j_\ell}]du = 0.$$

The bounded support condition for the kernel is imposed here to keep the conditions relatively simple. The last condition requires that the kernel be a higher order (bias reducing) kernel of order s. It will be used here to guarantee that the bias of the kernel estimator is small relative to variance. The next condition imposes smoothness on the density $h_0(x)$.

Assumption 2 *There is a nonnegative integer $d \geq s$ and an extension of $h_0(x)$ to all of $\mathbf{R}^r$ that is continuously differentiable to order d with bounded derivatives on $\mathbf{R}^r$.*

This condition is used in conjunction with Assumption 1 to make sure the bias of the estimator is small. It rules out cases where the density of x and its derivatives are nonzero on the boundary of the support by requiring smoothness everywhere. The next condition imposes some conditions on the bandwidth.

Assumption 3 *$\lambda = \lambda(n)$ such that $\sqrt{n}\lambda^r / \ln(n) \to \infty$ and $\sqrt{n}\lambda^s \to \infty$.*

Note that this condition implies that $s > r$, so that the order of the kernel and the degree of differentiability of the density must be larger than the dimension of x.

These three conditions imply the following result.

Lemma 1 *If Assumptions 1–3 are satisfied, $a(z) = 0$ except on a compact set $\mathcal{X}$ where $h_0(x)$ is bounded away from zero, $\mathrm{E}\left[\|a(z)\|^4\right] < \infty$, $\mathrm{E}[a(z) \mid x]$ is bounded on $\mathcal{X}$ and continuous in x on a set of full Lebesgue measure, then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{a(z_i)}{\hat{h}(x_i)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{a(z_i)}{h_0(x_i)} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\mathrm{E}[a(z) \mid x_i]}{h_0(x_i)} - \mathrm{E}\left[\frac{a(z)}{h_0(x)}\right]\right\} + o_p(1).$$

For $a(z) = f(x,\theta_0)Xu$, the conclusion of this result implies equation (7). Also, this result may be useful for other semiparametric estimators that depend on averages which are weighted by an inverse kernel density.

Some additional conditions are useful for showing asymptotic normality of the estimator from Section 2. The next condition imposes some requirements on the spherically symmetric density $f(x,\theta)$. Let $C(\theta)$ denote the closure of $\{f(x,\theta) \neq 0\}$ and θ_0 the probability limit of $\hat{\theta}$.

Assumption 4 *$C(\theta_0)$ is bounded, $h_0(x) > 0$ for $x \notin C(\theta_0)$, $C(\theta)$ is a continuous correspondence for θ in a neighborhood Θ of θ_0, and $f(x,\theta)$ is twice differentiable in θ with derivatives continuous in (x,θ).*

This assumption, which restricts the density $h_0(x)$ to be bounded away from zero where the trimming function is positive (the set $C(\theta_0)$), is extremely useful. It negates the "denominator problem" that would be present if the density of x were allowed to approach zero. This type of fixed trimming is theoretically more convenient than trimming that is relaxed as the sample size grows. Also, it may have the practical advantage of reducing outlier probl‘

The final condition imposes conditions on y and $\mathrm{E}[y \,|\, x]$.

Assumption 5 *$\mathrm{E}[y^4] < \infty$, $\mathrm{E}[y \,|\, x]$ is continuous almost everywhere with respect to Lebesgue measure and bounded on any bounded set, and $Q = \mathrm{E}[w(x)XX']$ is nonsingular.*

These conditions lead to the following asymptotic representation for $\hat{\gamma}$.

Theorem 1 *If Assumptions 1– 5 are satisfied then*

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}}Q^{-1}\sum_{i=1}^{n} w(x_i)X_i\{y_i - \mathrm{E}[y_i \mid x_i]\} + Q^{-1}\,\mathrm{E}\left[\frac{Xu}{h_0(x)}\frac{\partial f(x,\theta_0)}{\partial \theta'}\right]\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1).$$

The asymptotic distribution of $\hat{\beta}_2$ now follows in a straightforward way.

Theorem 2 *If Assumptions 1– 5 are satisfied, $\delta_{10} \neq 0$, and $f(x, \theta)$ is a spherically symmetric for all θ in a neighborhood of θ_0, then*

$$\sqrt{n}(\hat{\beta}_2 - \beta_{20}) \xrightarrow{d} \mathcal{N}(0, J'Q^{-1}\Sigma Q^{-1}J).$$

The last result that remains to be proved is the consistency of the asymptotic variance estimator.

Theorem 3 *If Assumptions 1– 5 are satisfied and $\delta_{10} \neq 0$ then*

$$\hat{J}'\hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}\hat{J} \xrightarrow{p} J'Q^{-1}\Sigma Q^{-1}J.$$

6 Monte Carlo Experiments

Ruud (1986) performed a simple Monte Carlo experiment to illustrate the use of density WLS. We repeat that experiment here to examine the success of the asymptotic approximations and to make a comparison of these estimators with the average derivative estimators of Powell, Stock and Stoker (1989). Both of these estimators are marginal estimators in the sense that they exploit marginal moment conditions, rather than conditional (on x) moment conditions.

The data were generated as follows. Two explanatory variables were drawn from a mixture of normal distributions:

$$h(x_1, x_2) = \phi(x_1 - 1/2)\phi(2x_2) + \phi(x_2 + 1/2)\phi(2x_1), \tag{10}$$

where ϕ is the standard normal pdf. In this way, positive x_1 tend to coincide with small x_2 and negative x_2 tend to coincide with small x_1. The dependent variable was generated by

$$y = \exp(x_1 + x_2 + u) \tag{11}$$

where u had a uniform distribution on $[-1/2, 1/2]$. Because the exponential function is convex, the OLS estimator for the linear regression of y on x_1, x_2, and a constant will overstate the relative effect of x_1 compared to the effect of x_2. The weights of the feasible density WLS estimator were computed using a kernel estimator of the density h and centering a normal pdf in the numerator on the sample mean and using the sample covariance matrix for a dispersion matrix. This pdf was trimmed at a standardized deviation from the mean of $\sqrt{6}$.

The joint pdf for x_1 and x_2 is pictured in Figure 1. Despite the mixture of two normals, the joint density remains unimodal and does not appear to be strangely idiosyncratic. The conditional expectation of x_2 given $x_1 + x_2$ is pictured in Figure 2. This function has a slight convexity, but not a dramatic one. This convexity will cause the OLS estimator to be inconsistent for the ratio of the slope parameters. Figure 3 gives a plot of the p.d.f. for $x'\beta = x_1 + x_2$ and the bounds on y conditional on $x'\beta$ from the data generating process.

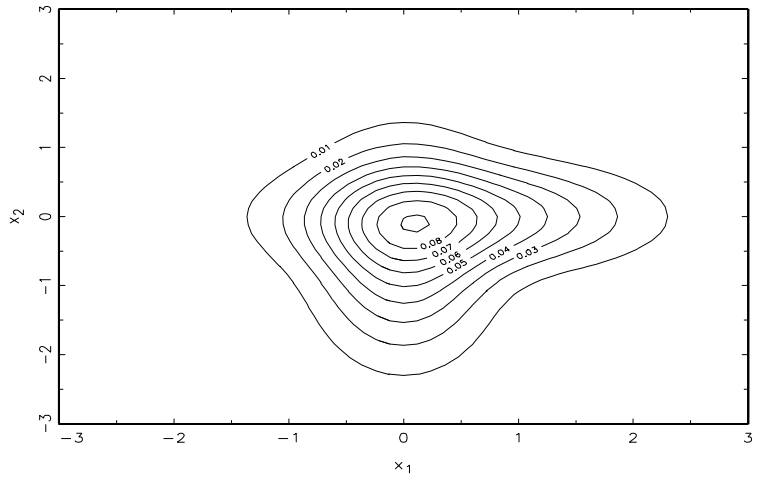

Figure 1: Contour Plot of Joint Density of x_1 and x_2

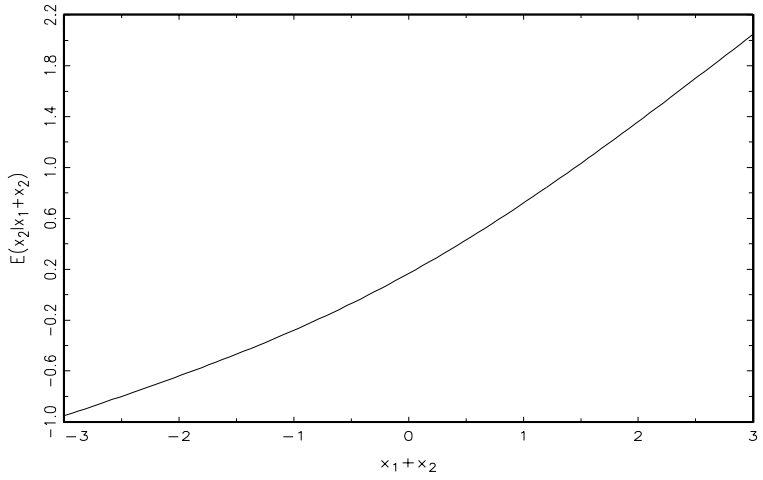

Figure 2: Conditional Expectation of x_2 Given $x_1 + x_2$

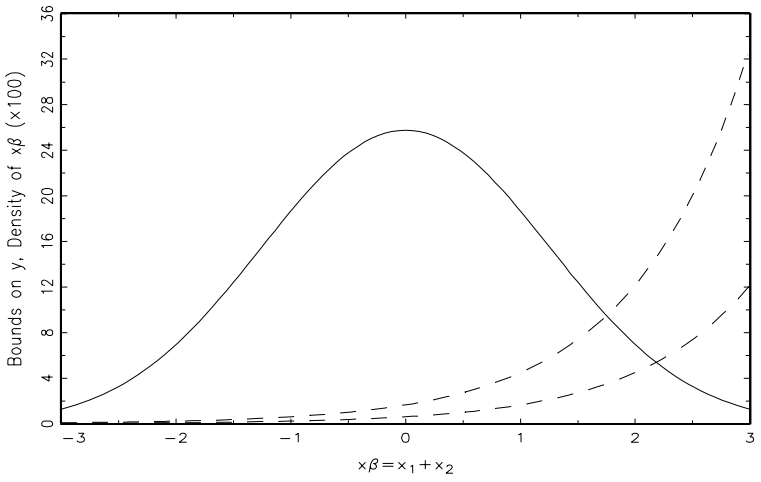

Figure 3: y Versus $x'\beta = x_1 + x_2$

There is substantial heteroskedasticity, with the variance increasing in the most informative region of the $x'\beta$ domain.

The extent of the inconsistency of OLS is shown in the first row of Table 1. For 100 observations, and 500 Monte Carlo replications, the average ratio of β_2/β_1 is 0.62. As expected, the relative importance of x_2 is diminished by its association with small values of $x'\beta$. The second line gives the feasible density WLS estimator and the third line the same estimator with the estimated density replaced by the actual density. The prediction of asymptotic approximations that the former would have smaller dispersion holds, but there is some bias in the feasible estimator. The fourth line of Table 1 lists a local version of the feasible estimator that divides the sample up into four orthants using the sample medians of x_1 and x_2, pooling the four estimators that can be computed for each orthant in a minimum chi- square estimator. This estimator exhibits none of the bias of the simple feasible estimator and also has a smaller variance than the exact density WLS estimator.

The remaining lines of the Monte Carlo results give the summary statistics for various average derivative estimators. The first average derivative estimator uses the exact density; the other four estimators are the four estimators simulated in Powell, Stock and Stoker (1989). The infeasible estimator also has no bias, but the feasible estimators exhibit strong bias relative to the density WLS estimators. The feasible estimators also exhibit less variation than the infeasible one, but on a root mean-squared error basis their performance is comparable.

At the bottom of Table 1, the efficiency bound and the asymptotic approximation to the variance of the feasible WLS estimator are given. The asymptotic

approximation works very well. But the efficiency bound is much smaller than the variance of the feasible density WLS estimator. In further research, we plan to investigate the possibility of attaining this bound using a technique like the local feasible WLS estimator just described.

The WLS estimators apply to discontinuous τ functions, whereas the average derivative estimators do not. We ran a second experiment to investigate the success of WLS with such functions. Using the same explanatory variables as in the first experiment, we changed (11) to

$$y = 1\{x_1 + x_2 > 1\} + u$$

where $u \sim \mathcal{N}(0, 0.01)$. In words, the data generating process of y is a mixture of $\mathcal{N}(0, 0.01)$ and $\mathcal{N}(1, 0.01)$ distributions, with the mean determined discretely by $x_1 + x_2$. Using 500 Monte Carlo replications of data sets with 100 observations, the OLS estimator (regressing y on a consant and the two x's) averaged 0.66 for the true ratio $\beta_2/\beta_1 = 1$, with a standard deviation of 0.17. The feasible density WLS estimator averaged 0.79 with a standard deviation of 0.32 and the (infeasible) exact density WLS estimator averaged 1.02 with a standard deviation of 0.44. In finite sample, the estimation of the density h clearly introduces some bias in the estimator that is not present when the exact density is used.

METHOD	**TRUE**	**MEAN**	**SD**	**RMSE**	**LQ**	**MEDIAN**	**UQ**	**MAE**
OLS	1.00	0.62	0.23	0.45	0.47	0.58	0.71	0.42
Weight Least Squares								
Feasible WLS	1.00	0.91	0.13	0.16	0.82	0.90	0.97	0.12
Exact WLS	1.00	0.99	0.33	0.33	0.78	0.95	1.16	0.19
Local WLS	1.00	0.98	0.26	0.26	0.82	0.95	1.11	0.15
Average Derivatives								
Exact IV	1.00	1.04	0.25	0.26	0.88	1.01	1.15	0.13
Un-Jackknifed Delta	1.00	0.78	0.14	0.26	0.68	0.76	0.87	0.24
Jackknifed Delta	1.00	0.81	0.14	0.23	0.72	0.80	0.89	0.20
Un-Jackknifed IV	1.00	0.76	0.12	0.27	0.67	0.75	0.83	0.25
Jackknifed IV	1.00	0.80	0.11	0.23	0.71	0.80	0.87	0.20
Asymptotic Approximations								
Efficient Score	1.00	1.00	0.037					
WLS	1.00	1.00	0.117					

Table 1: Monte Carlo Experiment Results

7 Appendix

We first give the a result showing that the LCE property holds for a spherically symmetric density.

Lemma 2 *Let $x \sim f[(x-\theta)'A^{-1}(x-\theta)]$ be a random variable with an elliptically symmetric (about θ) p.d.f. If $\mathrm{E}[\|x\|]$ exists, then $\mathrm{E}(x \mid \delta'x) = \alpha_0 + \alpha_1\delta'x$.*

Proof. Let $B = \delta(\delta'A\delta)^{-1}\delta'$ and $b = \delta'x$. According to the orthogonal decomposition

$$A^{-1} = (I - BA)(A - ABA)^{-}(I - AB) + B$$

where $(A - ABA)^{-}$ denotes a generalized inverse of $A - ABA$, we can write

$$(x-\theta)'A^{-1}(x-\theta) = (b - \delta'\theta)'(\delta'A\delta)^{-1}(b - \delta'\theta) + (x-\gamma)'(A - ABA)^{-}(x-\gamma)$$

where

$$\gamma \equiv \theta + A\delta(\delta'A\delta)^{-1}(b - \delta'\theta).$$

Therefore the conditional distribution of x given $\delta'x = b$ is symmetric around the point $\theta + A\delta(\delta'A\delta)^{-1}(b-\delta'\theta)$. Under existence of $\mathrm{E}[\|x\|]$, implying existence of the conditional expectation, the result follows with $\alpha_0 = \theta - AB\theta$ and $\alpha_1 = A\delta(\delta'A\delta)$. QED.

Lemma 3 *The asymptotic variance V of $\hat{\beta}_2$ is (8).*

Proof. Note that $X'\gamma = \gamma_1 + v\delta_1 + \delta_{10}x_2'\pi_2$, where $v = x'\beta_0$ and $\pi_2 = (\delta_2 - \beta_{20}\delta_1)/\delta_{10}$. Let $\hat{\pi}_2$ be the coefficient of $\delta_{10}x_2$ in the inverse density weighted least squares regression of y on $(1, v, \delta_{10}x_2')$. By the usual least squares property, $\hat{\pi}_2 = (\hat{\delta}_2 - \beta_{20}\hat{\delta}_1)/\delta_{10}$. Noting that $\hat{\pi}_2$ is just a linearization of $\hat{\beta}_2$, the delta method implies that the asymptotic variance of $\hat{\beta}_2$ is the same as $\hat{\nu}_2$. Let $\mathrm{E}_w[\cdot] = \mathrm{E}[w(x)(\cdot)]$ denote the expectation when the marginal distribution of x is $f(x, \theta_0)$. Then by elliptical symmetry of $f(x, \theta_0)$, the projection of $\delta_{10}x_2$ on $(1, v)$ equals $\delta_{10}\,\mathrm{E}_w[x_2 \mid v]$. Then equation (8) follows by the the usual partial least squares formula. QED.

Throughout the rest of the Appendix, C will denote a generic positive constant (not depending on N), that may be different in different uses, and $\sum_i = \sum_{i=1}^n$. The outline of the Appendix is that some useful Lemmas will first be given, and then the results in the body of the paper proven.

Proof of Lemma 1: The proof proceeds by verifying the conditions of Lemmas 5.2 and 5.4 of Newey (1992). Let $\mathcal{X}$ denote a compact set where $h_0(x)$ is bounded

away from zero and $a(z) = 0$ for x not in $\mathcal{X}$, and let $\|h\| = \sup_{x \notin \mathcal{X}} |h(x)|$. Also, let

$$\begin{aligned} m(z,h) &= \frac{a(z)}{h(x)}, \\ D(z,h) &= -\frac{a(z)h(x)}{h_0(x)^2}, \\ A(x) &= \mathrm{E}[a(z) \,|\, x], \\ m(h) &= \mathrm{E}[D(z,h)] \end{aligned}$$

Note that $m(h) = \int \nu(x)h(x)dx$ for $\nu(x) = -\mathrm{E}[a(z) \,|\, x]/h_0(x)$. Note that $\nu(x)$ is continuous almost everywhere (with respect to Lebesgue measure), zero outside the compact set $\mathcal{X}$, and bounded. Therefore, by Assumption 3, the conditions of Lemma 5.2 of Newey (1992) are satisfied, so by its conclusion,

$$\sqrt{n}[m(\hat{h}) - m(h_0)] = \frac{1}{\sqrt{n}} \sum_i \{\nu(x_i) - \mathrm{E}[\nu(x_i)]\} + o_p(1).$$

To check the hypotheses of Lemma 5.4 of Newey (1993), let $\Delta = \Delta_1 = \Delta_2 = 0$, so that the norm $\|h\|_\Delta$ of that result is $\|h\| = \sup_{x \notin \mathcal{X}} |h(x)|$. Note that

(i) $D(z,h)$ is linear in h on the set where $\|h\| < \infty$;

(ii) for $b(z) = \|a(z)\|$ and $\|h - h_0\| \leq \epsilon$ for small enough ϵ,

$$\begin{aligned} &\|m(z,h) - m(z,h_0) - D(z,h-h_0)\| \\ &\leq \|a(z)\| \left| \frac{1}{h(x)} - \frac{1}{h_0(x)} + \frac{h(x)}{h_0(x)^2} - \frac{1}{h_0(x)} \right| \\ &= b(z) \left| \frac{1}{h_0(x)^2 h(x)} \right| \left| h_0(x)^2 - 2h(x)h_0(x) + h(x)^2 \right| \\ &\leq Cb(z) \; |h_0(x) - h(x)|^2 \\ &\leq Cb(z) \, \|h_0 - h\|^2 ; \end{aligned}$$

(iii) $\|D(z,h)\| \leq C \, \|a(z)\| \, \|h\|$ and $\mathrm{E}[\|a(z)\|^4] < \infty$;

(iv) for $\eta_n = [\ln(n)/(n\lambda^r)]^{1/2} + \lambda^s$, $\sqrt{n}\eta_n^2 \leq C[\ln(n)/\sqrt{n} \; \lambda^r] + \sqrt{n}\lambda^{2s} \to 0$, and $\sqrt{n}\lambda^r \to 0$ by $r > s$. Then by the conclusion of Lemma 5.4 of Newey (1993),

$$\frac{1}{\sqrt{n}} \sum_i [m(z_i, \hat{h}) - m(z_i, h_0)] = \sqrt{n}[m(\hat{h}) - m(h_0)] + o_p(1).$$

The conclusion then follows by the triangle inequality. QED.

The following Lemma is useful for proving Theorem 1.

Lemma 4 *If $h_0(x)$ is continuous and Assumption 4 is satisfied then there is $\epsilon > 0$ and a compact set $\mathcal{X}$ such that $h_0(x) > 0$ for all $x \notin \mathcal{X}$ and $f(x,\theta) = 0, \partial f(x,\theta)/\partial\theta = 0,$ and $\partial^2 f(x,\theta)/\partial\theta\partial\theta' = 0$ for all $x \notin \mathcal{X}$ and $\|\theta - \theta_0\| < \epsilon$.*

Proof. By continuity of $C(\theta)$ and $h_0(x)$, there is ϵ small enough that $h_0(x) > 0$ for all $x \notin \mathcal{X}$ where $\mathcal{X}$ is the closure of $\cup_{\|\theta-\theta_0\|<\epsilon} C(\theta)$. By continuity of $C(\theta)$, the set $\mathcal{X}$ is compact. Also, for any $x \notin \mathcal{X}$, $f(x,\theta) = 0$ for *all* θ with $\|\theta - \theta_0\| < \epsilon$, so differentiating this identity at any such θ implies $\partial f(x,\theta)/\partial\theta = 0$ and $\partial^2 f(x,\theta)/\partial\theta\partial\theta' = 0$. QED.

Proof of Theorem 1: For the compact set $\mathcal{X}$ of Lemma 4,

$$\sup_{x\notin\mathcal{X}} \left|\hat{h}(x) - h_0(x)\right| \xrightarrow{p} 0$$

by Lemma B.3 of Newey (1993). Then by $h_0(x)$ bounded away from zero on $\mathcal{X}, \hat{h}(x)$ is bounded away from zero on $\mathcal{X}$ with probability approaching one. Also, for the ϵ of Lemma 4, $\left\|\hat{\theta} - \theta_0\right\| < \epsilon$ with probability approaching one, so that for all $x \notin \mathcal{X}$, $f(x,\bar{\theta}) = 0$, $\partial f(x,\bar{\theta})/\partial\theta = 0$, and $\partial^2 f(x,\bar{\theta})/\partial\theta\partial\theta'$, for any $\bar{\theta}$ on the line joining $\hat{\theta}$ and θ_0 (e.g., for $\bar{\theta} = \hat{\theta}$). It then follows that with probability approaching one, by X bounded on $\mathcal{X}$ and $f(x,\theta)$ Lipschitz in θ,

$$\begin{aligned}
\left\|\hat{Q} - \sum_i w_i X_i X_i'/n\right\| &\leq C \sum_i |\hat{w}_i - w_i| /n \\
&\leq C \sup_{x\in\mathcal{X}} \left[\left|f(x,\hat{\theta})\right| \left|\frac{1}{\hat{h}(x)} - \frac{1}{h_0(x)}\right|\right] \\
&\quad + \sup_{x\in\mathcal{X}} \left[\frac{1}{h_0(x)} \left|f(x,\hat{\theta}) - f(x,\theta)\right|\right] \\
&\xrightarrow{p} 0.
\end{aligned}$$

Also, by the law of large numbers, $\sum_i w_i X_i X_i'/n \xrightarrow{p} Q$, so by the triangle inequality, $\hat{Q} \xrightarrow{p} Q$.

Next, by a mean value expansion, for $\tilde{w}_i = f(x_i,\theta_0)/\hat{h}(x_i)$,

$$\frac{1}{\sqrt{n}} \sum_i \hat{w}_i X_i u_i = \frac{1}{\sqrt{n}} \sum_i \tilde{w}_i X_i u_i + \left[\frac{1}{n} \sum_i \frac{X_i u_i}{\hat{h}(x_i)} \frac{\partial f(x_i,\bar{\theta})}{\partial\theta'}\right] \sqrt{n}(\hat{\theta} - \theta_0).$$

It follows similarly to the argument for $\hat{Q} \xrightarrow{p} Q$ that the matrix in square brackets converges in probability to $\mathrm{E}[Xu/h_0(x)\, \partial f(x,\theta_0)/\partial\theta']$. It also follows by Lemma 1 that

$$\frac{1}{\sqrt{n}} \sum_i \tilde{w}_i X_i u_i = \frac{1}{\sqrt{n}} \sum_i w_i X_i u_i - \frac{1}{\sqrt{n}} \sum_i w_i X_i \{\mathrm{E}[y_i \,|\, x_i] - X_i'\gamma_0\} + o_p(1).$$

The conclusion then follows by the triangle inequality. QED.

Proof of Theorem 2: By the Theorem 1, the delta method, and the central limit theorem it sufficies to show that $J'Q^{-1}\,\mathrm{E}[Xu/h_0(x)\,\partial f(x,\theta_0)/\partial\theta'] = 0$. Let $Q(\theta) = \int XX' f(x,\theta)dx$ and $m(\theta) = \int X \cdot \mathrm{E}[y \,|\, x] f(x,\theta)dx$. By boundedness of X, $\mathrm{E}[y \,|\, x]$, and $f(x,\theta)$ on the set $\mathcal{X}$ of the proof of Theorem 2, both $Q(\theta)$ and $m(\theta)$ are differentiable, and $\partial m(\theta_0)/\partial\theta = \mathrm{E}[Xu/h_0(x)\,\partial f(x,\theta_0)/\partial\theta']$. It follows by $Q(\theta_0) = Q$ nonsingular that $Q(\theta)$ is nonsingular for θ in a neighborhood of θ_0. On this neighborhood of $Q(\theta)$ let $\gamma(\theta) = (\gamma_1(\theta), \delta(\theta)')' = Q(\theta)^{-1}m(\theta)$. Note that $\delta(\theta)$ is continuous function of θ and $\delta(\theta_0) = \delta_0$. Then by $\delta_{10} \neq 0$, there is an even smaller neighborhood where $\delta_1(\theta) \neq 0$. Let $\beta_2(\theta) = \delta_2(\theta)/\delta_1(\theta)$. By spherical symmetry of $f(x,\theta)$, it follows as in Ruud (1986) that $\beta_2(\theta) = \beta_{20}$. Differentiating this identity gives $0 = J'\partial\gamma(\theta_0)/\partial\theta$. Furthermore, differentiating the identity $\int X\{\mathrm{E}[y \,|\, x] - X'\gamma(\theta)\}f(x,\theta)dx = 0$ with respect to θ gives $\partial\gamma(\theta_0)/\partial\theta = Q^{-1}\,\mathrm{E}[Xu/h_0(x)\,\partial f(x,\theta_0)/\partial\theta']$. QED.

Proof of Theorem 3: $\hat{J} \xrightarrow{p} J$ follows by $\hat{\gamma} \xrightarrow{p} \gamma_0$ and $\delta_{10} \neq 0$. Also $\hat{Q} \xrightarrow{p} Q$ follows as in the proof of Theorem 1. Therefore, by continuity of matrix inversion and multiplication, it only remains to show that $\hat{\Sigma} \xrightarrow{p} \Sigma$. Let $\hat{d}(x) = \sum_{i=1}^n K_\lambda(x - x_i)y_i$ and $d(x) = h_0(x)\,\mathrm{E}[y \,|\, x]$. By a change of variables, $\mathrm{E}[\hat{d}(x)] = \int K(u)d(x + u\lambda)du = \bar{d}(x)$, which is bounded on any bounded set by $K(u)$ having bounded support and $d(x)$ bounded on any bounded set. Furthermore, at each x where $d(x)$ is continuous, $d(x + u\lambda) \to d(x)$ as $\lambda \to 0$, so by the dominated convergence theorem, $\bar{d}(x) \to d(x)$ at each such x. Since the set of such x values has full Lebesgue measure, the dominated convergence theorem implies that $\int_{\mathcal{X}}[\bar{d}(x) - d(x)]^2 h_0(x)dx \to 0$. By Lemma B.1 of Newey (1993), $\sup_{x\notin\mathcal{X}}\left|\hat{d}(x) - \bar{d}(x)\right| \xrightarrow{p} 0$. Let $1_i = 1(x_i \notin \mathcal{X})$. Then $\sum_i 1_i \left|\bar{d}(x_i) - d(x_i)\right|^2/n \xrightarrow{p} 0$ by the Markov inequality. Also, by $\hat{h}(x)$ bounded away from zero uniformly on $\mathcal{X}$, with probability approaching one,

$$\frac{1}{n}\sum_i 1_i \left|\hat{g}(x_i) - g(x_i)\right|^2 \leq \frac{C}{n}\sum_i 1_i \left|\hat{d}(x_i) - \bar{d}(x_i)\right|^2 + \frac{1}{n}\sum_i 1_i \left|\bar{d}(x_i) - d(x_i)\right|^2$$

$$+ \frac{1}{n}\sum_i 1_i \left|d(x_i)\right|^2 \left|\frac{1}{\hat{h}(x_i)} - \frac{1}{h_0(x_i)}\right|^2$$

$$\leq C \sup_{x\notin\mathcal{X}}\left|\hat{d}(x) - \bar{d}(x)\right| + o_p(1) + \sup_{x\notin\mathcal{X}}\left|\hat{h}(x) - h_0(x)\right| \xrightarrow{p} 0.$$

Let $\tilde{\Sigma} = \sum_i \hat{w}_i^2 X_i X_i'[y_i - g(x_i)]/n$. Then arguing as in the proof of Theorem 1, using Lemma 4, it follows by the Cauchy-Schwartz inequality that for $1_i =$

$1(x_i \notin \mathcal{X})$,

$$\begin{aligned}
\left\|\hat{\Sigma} - \tilde{\Sigma}\right\| &\leq \frac{C}{n}\sum_i 1_i \left[2\left|y_i\right|\left|\hat{g}(x_i) - g(x_i)\right| + \left|\hat{g}(x_i)^2 - g(x_i)^2\right|\right] \\
&\leq \frac{C}{n}\sum_i 1_i [\hat{g}(x_i) - g(x_i)]^2 \\
&+ \; C\left[\left(\frac{1}{n}\sum_i \left|y_i\right|^2\right)^{1/2} + \left(\frac{1}{n}\sum_i \left|g(x_i)\right|^2\right)^{1/2}\right] \\
&\quad \left[\frac{1}{n}\sum_i 1_i \left|\hat{g}(x_i) - g(x_i)\right|^2\right]^{1/2} \\
&\xrightarrow{p} 0.
\end{aligned}$$

It also follows similarly to the proof that $\hat{Q} \xrightarrow{p} Q$ that $\tilde{\Sigma} \xrightarrow{p} \Sigma$. The conclusion that $\hat{\Sigma} \xrightarrow{p} \Sigma$ then follows by the triangle inequality. QED.

8 References

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58(1–2), 71–120.

Newey, W.K. (1994), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory,* 10, 233–253.

Newey, W.K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica,* 62(6), 1349–82.

Newey, W.K. and T.M. Stoker (1993), "Efficiency of Weighted Average Derivative Estimators and Index Models," *Econometrica,* 61, 1199–1223.

Newey, W.K. and D. McFadden (1993), "Large Sample Estimation and Hypothesis Testing," Engle, R., and D. McFadden, *Handbook of Econometrics, Vol. 4*, Amsterdam: North-Holland, 2111–2245.

Powell, James L., James H. Stock and Thomas M. Stoker (1989), "Semiparametric Estimation of Index Model Coefficients," *Econometrica*, 57, 1403–1430.

Ruud, Paul A. (1981), "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models," *Econometrica*, 51(1), 225-228.

Ruud, Paul A. (1986), "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157–187.

Stoker, Thomas M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica,* 54, 1461–1481.