TWO STEP SERIES ESTIMATION OF SAMPLE SELECTION MODELS

by

Whitney K. Newey[1]
Department of Economics
MIT, E52-262D
Cambridge, MA 02139

April 1988

Latest Revision, January 1999

Abstract

Sample selection models are important for correcting for the effects of nonrandom sampling in microeconomic data. This note is about semiparametric estimation using a series approximation to the selection correction term. Regression spline and power series approximations are considered. Consistency and asymptotic normality are shown, as well as consistency of an asymptotic variance estimator.

JEL Classification: C14, C24

Keywords: Sample selection models, semiparametric estimation, series estimation, two-step estimation.

[1]Presented at the 1988 European meeting of the Econometric Society. Helpful comments were provided by D.W.K. Andrews, G. Chamberlain, A. Gregory, J. Ham, J. MacKinnon, D. McFadden, and J. Powell. The NSF and the Sloan Foundation provided financial support.

1. Introduction

Sample selection models provide an approach to correcting for nonrandom sampling that is important in econometrics. Pioneering work in this area includes Gronau (1973) and Heckman (1974). This paper is about two-step estimation of these models without restricting the functional form of the selection correction. The estimators are particularly simple, using polynomial or spline approximations to correct for selection. Asymptotic normality and consistency of an asymptotic variance estimator are shown.

Some of the estimators considered here are similar to two-step least squares estimators with flexible correction terms previously proposed by Lee (1982) and Heckman and Robb (1987). The theory here allows the functional form of the correction to be entirely unknown, with the number of approximating functions growing with the sample size to achieve $\sqrt{n}$-consistency and asymptotic normality. Also, this paper adds to the menu of approximations by considering new types of power series, along with regression splines that are important in statistical approximation theory (e.g. Stone, 1985).

Early work on semiparametric estimation of sample selection models includes Cosslett (1991) and Gallant and Nychka (1987). These papers do not have asymptotic normality results. Powell (1987) and Ahn and Powell (1993) give distribution theory for density weighted kernel estimators. The series estimators analyzed here have the virtue of being extremely easy to implement. Also, some of the estimators are new, including the regression splines. Practical experience with these estimators is given in Newey, Powell, and Walker (1990).

Section 2 of the paper presents the model and discusses identification. The estimators are described in Section 3, and Section 4 gives the asymptotic theory.

2. The Model and Identification

The selection model model considered here is

(2.1) $y = x'\beta_0 + \xi$, y only observed if $d = 1$, $d \in \{0, 1\}$.

$$E[\xi | w,d=1] = E[\xi | v(w,\alpha_0),d=1], \quad \text{Prob}(d = 1|w) = \pi(v(w,\alpha_0)), \quad x \subseteq w.$$

Here the conditional mean of the disturbance, given selection and w, depends only on the index $v = v(w,\alpha_0)$. This restriction is implied by other familiar conditions, such as independence of disturbances and regressors, see Powell (1994). A basic implication of this model is that

(2.2) $$E[y|w,d=1] = x'\beta_0 + h_0(v), \quad h_0(v) = E[\xi|w,d=1]$$

The function $h_0(v)$ is a selection correction that is familiar. For example if $d = 1(v+\tilde{\xi} \geq 0)$, $(\xi,\tilde{\xi})$ is independent of w, $\tilde{\xi}$ has a standard normal distribution, and $E[\xi|\tilde{\xi}]$ is linear in $\tilde{\xi}$, then $h_0(v) = \phi(v)/\Phi(v)$, where $\Phi(v)$ and $\phi(v)$ are the standard normal CDF and p.d.f. respectively. This term is the correction term considered by Heckman (1976). In this paper we allow $h_0(v)$ to have an unknown functional form.

Equation (2.2) is an additive semiparametric regression like that considered by Robinson (1988), except that the variable $v = v(w,\alpha_0)$ depends on unknown parameters. Making use of this information is important for identification. Ignoring the structure implied by equation (2.1), and regarding h_0 as an unknown function of variables in w, would mean that any component of x that is included in those variables would not be identified.

The identification condition for this paper is

Assumption 1: $M = E[d(x-E[x|v,d=1])(x-E[x|v,d=1])']$ is nonsingular, i.e. for any $\lambda \neq 0$ there is no measurable function $f(v)$ such that $x'\lambda = f(v)$ when $d = 1$.

This condition was imposed by Cosslett (1991), and is the selection model version of Robinson's (1988) identification condition for additive semiparametric regression. As shown by Chamberlain (1986), this condition is not necessary for identification, but it is necessary for existence of a (regular) $\sqrt{n}$-consistent estimator. It is important to note that this condition does not allow for a constant term in x, because it is not separately identified from $h_0(v)$.

More primitive conditions for Assumption 1 are available in some cases. A simple sufficient condition is that $Var(x)$ is nonsingular and the conditional distribution of v given x has an absolutely continuous component with conditional density that is positive on the entire real line for almost all x. An obvious necessary condition is that v not be a linear combination of x, requiring that something in v be excluded from x. Such an exclusion restriction is implied by many economic models, where d is a choice variable and v includes a price variable for another choice.

Identification of β_0 from equation (2.2) also requires identification of α_0. Here no specific assumptions will be imposed, in order to allow flexibility in the choice of an estimator of α_0. Of course, consistency of $\hat{\alpha}$ will imply identification of α_0, but different consistent estimators $\hat{\alpha}$ may correspond to different identifying assumptions. For brevity, a menu of different assumptions is not discussed here.

3. Estimation

The type of estimator we consider is a two-step estimator, where the first step is a semiparametric estimator $\hat{\alpha}$ of the selection parameters α_0 and the second step is least squares regression on x and approximating functions of $\hat{v} = v(x,\hat{\alpha})$ in the selected data. These estimators are analogous to Heckman's (1976) two-step procedure for the Gaussian disturbances case. The difference is that α is estimated by a distribution-free method rather than by probit and a nonparametric approximation to $h(v)$

is used in the second step regression rather than the inverse Mills ratio.

There are many distribution free estimates that are available for the first step, including those of Manski (1975), Cosslett (1983), and Ruud (1986). The first step will need to be $\sqrt{n}$-consistent, like the estimator of Powell, Stock, and Stoker (1989), Ichimura (1993), and Cavanagh and Sherman (1997). Also, the asymptotic variance of $\hat{\beta}$ will be an increasing function of the asymptotic variance of $\hat{\alpha}$, so an efficient estimator like that of Klein and Spady (1993) may be useful.

The second step consists of a linear regression of y on x and functions of $\hat{v}$ that can approximate $h_0(v)$. To describe the estimator let $\tau(v,\eta)$ denote some strictly monotonic transformation of v, depending on parameters η. This transformation is useful for adjusting the location and scale of v, as discussed below. Let $p^K(\tau) = (p_{1K}(\tau),\ldots,p_{KK}(\tau))'$ be a vector of functions with the property that for large K a linear combination of $p^K(\tau)$ can approximate an unknown function of τ. Suppose that the data are $z_i = (d_i,w_i,d_iy_i)$, $(i = 1, \ldots, n)$, assumed throughout to be i.i.d.. Let $\hat{\eta}$ denote an estimator of η, $\hat{v}_i = v(w_i,\hat{\alpha})$, $\hat{\tau}_i = \tau(\hat{v}_i,\hat{\eta})$, and $\hat{p}_i = p^K(\hat{\tau}_i)$, where a K superscript for $\hat{p}_i$ is suppressed for notational convenience. For $x = [d_1x_1,\ldots,d_nx_n]'$, $y = (d_1y_1,\ldots,d_ny_n)'$, $\hat{P} = [d_1\hat{p}_1,\ldots,d_n\hat{p}_n]'$, and $\hat{Q} = \hat{P}(\hat{P}'\hat{P})^{-1}\hat{P}'$ the estimator is

$$(3.1) \qquad \hat{\beta} = \hat{M}^{-1}x'(I-\hat{Q})y/n, \quad \hat{M} = x'(I-\hat{Q})x/n,$$

where the inverses will exist in large samples under conditions discussed below. The estimator $\hat{\beta}$ is the coefficient of x_i from the regression of y_i on x_i and $\hat{p}_i$ in the selected data.

This estimator depends on the choice of approximating functions and transformation. Here we consider two kinds of approximating functions, power series and splines. For power series the approximating functions are given by

$$(3.2) \qquad p_{kK}(\tau) = \tau^{k-1}.$$

Depending on the transformation $\tau(v,\eta)$, this power series can lead to several different types of sample selection corrections. Three examples are a power series in the index $\hat{v}$, in the inverse Mills ratio $\phi(\cdot)/\Phi(\cdot)$, or in the normal CDF $\Phi(\cdot)$. When a nonlinear transformation of v is used (e.g. for a power series in Φ), it may be appropriate to undo a location and scale normalization imposed on most semiparametric estimators of $v(w,\alpha)$. To this end let $\hat{\eta} = (\hat{\eta}_1,\hat{\eta}_2)'$ be the coefficients from probit estimation with regressors $(1,\hat{v}_i)$, where we do not impose normality (but will require that $\hat{\eta}$ be a $\sqrt{n}$-consistent of some population parameter). Then the transformed observations for the three examples will be

(3.3a) $\hat{\tau}_i = \hat{v}_i,$

(3.3b) $\hat{\tau}_i = \phi(\hat{\eta}_1+\hat{\eta}_2\hat{v}_i)/\Phi(\hat{\eta}_1+\hat{\eta}_2\hat{v}_i),$

(3.3c) $\hat{\tau}_i = \Phi(\hat{\eta}_1+\hat{\eta}_2\hat{v}_i).$

The power series in equation (3.3a) will have as a leading term the index $\hat{v}_i$ itself. The one from equation (3.3b) will have leading term given by the inverse Mills, so that the first term is the Heckman (1976) correction. This one also has approximating functions that preserve a shape property of $h_0(v)$ that holds when $d = 1(v+\tilde{\xi}\geq 0)$ and $(\xi,\tilde{\xi})$ are independent of v, that $h_0(v)$ goes to zero as v gets large. The last example will correspond to a power series in the selection probability for Gaussian $\tilde{\xi}$.

Replacing power series by corresponding polynomials that are orthogonal with respect to some weight function may help avoid multicollinearity. For example, for $\hat{\tau}_u \equiv \max_{i\leq n, d_i=1}\{\tau(\hat{v}_i,\hat{\eta})\}$ and $\hat{\tau}_\ell = \min_{i\leq n, d_i=1}\{\tau(\hat{v}_i,\hat{\eta})\}$ one could replace τ^{k-1} by a polynomial of order k that is orthogonal for the uniform weight on $[-1,1]$, evaluated at $\hat{\tau}_i = [2\tau(\hat{v}_i,\hat{\eta})-\hat{\tau}_u-\hat{\tau}_\ell]/(\hat{\tau}_u-\hat{\tau}_\ell)$. Of course, $\hat{\beta}$ is not affected by such a replacement, since it is just a nonsingular linear transformation of the power series.

An alternative approximation that is better in several respects than power

series is splines, that are piecewise polynomials. Splines are less sensitive to outliers and to singularities in the function being approximated. Also, as discussed below, asymptotic normality holds under weaker conditions for splines than power series. For theoretical convenience attention is limited to splines with evenly spaced knots on [-1,1]. For $b_+ \equiv 1(b > 0) \cdot b$, a spline of degree m in τ with L evenly spaced knots on $[-1,1]$ can be based on

$$(3.4) \qquad p_{kK}(\tau) = \tau^{k-1}, \quad 1 \le k \le m+1,$$

$$= \{[\tau + 1 - 2(k-m-1)/(L+1)]_+\}^m, \quad m+2 \le k \le m+1+L \equiv K.$$

An alternative, equivalent series that is less subject to multicollinearity problems is B-splines; e.g. see Powell (1981).

Fixed, evenly spaced knots is restrictive, and is motivated by theoretical convenience. Allowing the knots to be estimated may improve the approximation, but would make computation more difficult and require substantial modification to the theory of Section 4, which relies on linear in parameter approximations.

For inference it is important to have a consistent estimator of the asymptotic variance of $\hat{\beta}$. This can be formed by treating the approximation as if were exact and using formulae for parametric two-step estimators such as those of Newey (1984). The estimator will depend on a consistent estimator $\hat{V}(\hat{\alpha})$ of the asymptotic variance of $\sqrt{n}(\hat{\alpha}-\alpha_0)$. Let $\hat{\beta}$ and $\hat{\gamma}$ be the estimates from the regression of $d_i y_i$ on $d_i x_i$ and $d_i \hat{p}_i$, $\hat{\varepsilon}_i = d_i(y_i - x_i'\hat{\beta} - \hat{p}_i'\hat{\gamma})$ the corresponding residual, and $\hat{h}(v) = p^K(\tau(v,\hat{\eta}))'\hat{\gamma}$ the estimate of $h(v)$ obtained from this regression. Define $\hat{u} = (I-\hat{Q})x$ to be the matrix of residuals from the regression of $d_i x_i$ on $d_i \hat{p}_i$, so that $x'(I-\hat{Q})x = \hat{u}'\hat{u}$ and let

$$(3.5) \qquad \hat{V}(\hat{\beta}) = \hat{M}^{-1}[\textstyle\sum_{i=1}^n \hat{u}_i \hat{u}_i' (\hat{\varepsilon}_i)^2/n + \hat{H}\hat{V}(\hat{\alpha})\hat{H}']\hat{M}^{-1},$$

$$\hat{H} = \textstyle\sum_{i=1}^n \hat{u}_i [\partial \hat{h}(\hat{v}_i)/\partial v] \partial v(w_i,\hat{\alpha})/\partial \alpha'/n.$$

This estimator is the sum of two terms, the first of which is the White (1980)

specification robust variance estimator for the second step regression and the second a term that accounts for the first-stage estimation of the parameters of the selection equation. It can also be interpreted as the block of a joint variance estimator for $\hat{\beta}$ and $\hat{\gamma}$ corresponding to $\hat{\beta}$, where the joint estimator is formed as in Newey (1984). This estimator will be consistent for the asymptotic variance of $\sqrt{n}(\hat{\beta}-\beta_0)$ under the conditions of Section 4. Note here the normalization by the total sample size n rather than the number of observations in the selected sample. For example, a 95 percent asymptotic confidence interval for β_j is $[\hat{\beta}_j-\hat{V}(\hat{\beta})_{jj}^{1/2}1.96/\sqrt{n},\ \hat{\beta}_j+\hat{V}(\hat{\beta})_{jj}^{1/2}1.96/\sqrt{n}]$.

4. Asymptotic Normality

Some regularity conditions will be used to show consistency and asymptotic normality. The first condition is about the first stage estimator.

Assumption 2: There exists $\psi(w,d)$ such that for $\psi_i = \psi(w_i,d_i)$, $\sqrt{n}(\hat{\alpha}-\alpha_0) = \sum_{i=1}^{n}\psi_i/\sqrt{n} + o_p(1)$, $E[\psi_i] = 0$, and $E[\psi_i\psi_i']$ exists and is nonsingular. Also, for $\hat{V}(\hat{\alpha}) \xrightarrow{p} V(\hat{\alpha}) = E[\psi_i\psi_i']$.

This condition requires that $\hat{\alpha}$ be asymptotically equivalent to a sample average that depends only on w and d. It is satisfied by many semiparametric estimators of binary choice models, such as that of Klein and Spady (1993).

The next condition imposes some moment conditions on the second stage.

Assumption 3: For some $\delta > 0$, $E[d\|x\|^{2+\delta}] < \infty$, $Var(x|v,d=1)$ is bounded, and for $\varepsilon \equiv d(y-x'\beta_0-h_0(v))$, $E[\varepsilon^2|v,d=1]$ is bounded.

The bounded conditional variance assumptions are standard in the literature, and will not be very restrictive here because v will also be assumed to be bounded.

To control the bias of the estimator is essential to impose some smoothness conditions on functions of v.

Assumption 4: $h_0(v)$ and $E[x|v,d=1]$ are continuously differentiable in v, of orders s and t respectively.

We also require that the transformation τ satisfy some properties.

Assumption 5: There is η_0 with $\sqrt{n}(\hat{\eta}-\eta_0) = O_p(1)$, the distribution of $\tau(v(w,\alpha_0),\eta_0)$ has an absolutely continuous component with p.d.f. bounded away from zero on its support, which is compact. Also, the first and second partial derivatives of $v(w_i,\alpha)$ and $\tau(v,\eta)$ with respect to α, v, and η are bounded for α and η in a neighborhood of α_0 and η_0 respectively.

The first condition of this assumption means that the density of τ_i is bounded away from zero, which is useful for series estimation, but is restrictive. For example, if $v = x_1 + x_2$, where x_1 and x_2 are continuously distributed and independent, then the density of v, which is a convolution of the densities of x_1 and x_2, will be everywhere continuous, and hence cannot have density bounded away from zero. It would be useful to weaken this condition, but this would be difficult and is beyond the scope of this paper.

The next assumption imposes growth rate conditions for the number of approximating terms.

Assumption 6: $K = K_n$ such that $\sqrt{n}K^{-s-t+1} \xrightarrow{p} 0$ and a) $p^K(\tau)$ is a power series, $s \geq 5$, and $K^7/n \longrightarrow 0$; or b) $p^K(\tau)$ is a spline with $m \geq t-1$ $s \geq 3$, and $K^4/n \longrightarrow 0$.

Here, splines require the minimum smoothness conditions and the least stringent growth rate for the number of terms, with $h_0(v)$ only required to be three times continuously differentiable. It is also of note that this assumption does not required under-smoothing. The presence of t in the rate conditions means that smoothness in

$E[x|v,d=1]$ can compensate for lack of smoothness in $h_0(v)$, so that the bias of $\hat{h}(v)$ does not have to go to zero faster than the variance. This absence of an undersmoothing requirement is a feature of series estimators of semiparametric regression models that has been previously noted in Donald and Newey (1994).

Asymptotic normality of the two-step least squares estimator and consistency of the estimator of its asymptotic covariance matrix follow from the previous conditions. Let $u_i = d_i\{x_i - E[x_i|v_i,d_i=1]\}$, $\Omega = E[\varepsilon_i^2 u_i u_i']$, and $H = E[u_i\{dh_0(v_i)/dv_i\}\partial v(w_i,\alpha_0)/\partial\alpha']$.

Theorem 1: If Assumptions 1 - 6 are satisfied and Ω *is nonsingular then for* $V(\hat{\beta}) = M^{-1}(\Omega + HV(\hat{\alpha})H')M^{-1}$, $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V(\hat{\beta}))$, *and* $\hat{V}(\hat{\beta}) \xrightarrow{p} V(\hat{\beta})$.

This result gives $\sqrt{n}$-consistency and asymptotic normality of the series estimators considered in this paper, that are useful for large sample inference. It would also be useful to have a way of choosing the number of functions in practice. A K that minimizes goodness of fit criteria for the selection correction, such as cross-validation on the equation of interest, should satisfy the rate conditions of Assumption 6. In Newey, Powell and Walker (1990) such a criteria was used and gave reasonable results. However, the results of Donald and Newey (1994) and Linton (1995) for the partially linear model suggests that it may be optimal for estimation of β to undersmooth, meaning K should be larger than the minimum of a goodness of fit criteria. Such results are beyond the scope of this paper, but remain an important topic for future research.

Appendix: Proof of Theorem 1

Throughout the Appendix C will denote a positive constant that can be different in different uses. Also, we will use repeatedly the result that if $E[Y_n|X_n] \xrightarrow{p} 0$ for a sequence of positive random variables Y_n and conditioning sets X_n, then $Y_n \xrightarrow{p} 0$. To begin the proof, note that by $\partial v(w,\alpha)/\partial\alpha$ bounded and $\sqrt{n}$-consistency of $\hat{\alpha}$, and by $\partial\tau(v,\eta)/\partial v$ bounded, $\max_i|\hat{\tau}_i-\tau_i| = O_p(1/\sqrt{n})$. Also, by the density of τ_i bounded away from zero, both $\min_i\tau_i$ and $\max_i\tau_i$ will be $\sqrt{n}$-consistent for the boundary points of the support of τ_i, and hence so will $\min_i\hat{\tau}_i$ and $\max_i\hat{\tau}_i$. Therefore, by a location and scale transformation for power series, which will not change the regression, it can be assumed that $|\hat{\tau}_i| \leq 1$ and $\max_i|\hat{\tau}_i-\tau_i| = O_p(1/\sqrt{n})$. Now, it follows from Assumption 6, as in Newey (1997) that for $\|A\| = \mathrm{tr}(A'A)^{1/2}$, there is a nonsingular linear transformation of $\tilde{p}^K(\tau)$ of $p^K(\tau)$ such that

$$\text{(A.1)}\qquad E[d_i\tilde{p}^K(\tau_i)\tilde{p}^K(\tau_i)'] = I,\quad \sup_{|\tau|\leq 1}\|d^s\tilde{p}^K(\tau)/d\tau^s\| \leq \zeta_s(K),$$

$$\zeta_1(K)K^{1/2}/\sqrt{n} \longrightarrow 0,\quad \zeta_1(K)K^{-s+1} \longrightarrow 0,$$

$$\zeta_s(K) = CK^{(1+2s)/2} \text{ for splines},\quad \zeta_s(K) = CK^{1+2s} \text{ for power series}.$$

Since a nonsingular transformation does not change $\hat{\beta}$, it will be convenient to just let $\tilde{p}^K = p^K$. Then, as in Newey (1997), $\|P'P/n - I\| = O_p(\zeta_0(K)K^{1/2}/\sqrt{n}) \xrightarrow{p} 0$. Also, by the mean value theorem, $\max_i\|\hat{P}_i-P_i\| \leq \zeta_1(K)\max_i|\hat{\tau}_i-\tau_i| = O_p(\zeta_1(K)/\sqrt{n})$, so that $\|\hat{P}'\hat{P}/n - P'P/n\| \leq \|\hat{P}-P\|^2/n + \|P\|\|\hat{P}-P\|/n = O_p(\zeta_1(K)^2/n + K^{1/2}\zeta_1(K)/\sqrt{n}) \xrightarrow{p} 0$. Hence, by the triangle inequality,

$$\text{(A.2)}\qquad \|\hat{P}'\hat{P}/n - I\| \xrightarrow{p} 0.$$

It follows, as in Newey (1997), that $\lambda(\hat{P}'\hat{P}/n) \geq C$ with probability approaching one, where $\lambda(A)$ denotes the smallest eigenvalue of a symmetric matrix A.

Next, since $\tau(v,\eta_0)$ is one-to-one, conditioning on v is equivalent to

conditioning on τ, so that, for example, $h_0(v)$ can be regarded as a function of τ. Let $\mu_i = d_i E[x_i|\tau_i,d_i=1]$, $\mu = [\mu_1,\ldots,\mu_n]$, and $\hat{\mu} = \hat{Q}x$. So that $\|\hat{\mu}-\mu\|^2/n = tr(x'\hat{Q}x-2x'\hat{Q}\mu+\mu'\mu)/n$. By $\lambda(\hat{P}'\hat{P}/n) \geq C$, $\hat{Q}$ idempotent, and existence of the second moment of x_i, for $\hat{A} = \hat{P}(\hat{P}'\hat{P})^{-1}$

$$\|x'\hat{A}\|^2 = tr(x'\hat{A}\hat{A}'x) \leq O_p(1)tr(x'\hat{Q}x/n) \leq O_p(1)tr(x'x/n) = O_p(1).$$

It follows similarly that $\|x'A\| = O_p(1)$ for $A = P(P'P)^{-1}$. Also, $\|x'(\hat{P}-P)/n\| \leq \|x\|\|\hat{P}-P\|/n = O_p(\zeta_1(K)/\sqrt{n}) \xrightarrow{p}$ so that for $Q = P(P'P)^{-1}P'$,

$$\text{(A.3)} \qquad \|x'\hat{Q}x/n - x'Qx/n\| \leq \|x'(\hat{P}-P)\hat{A}'x/n\| + \|x'A(\hat{P}'\hat{P}-P'P)\hat{A}'x/n\| + \|x'A(\hat{P}-P)'x/n\|$$

$$\leq \|x'(\hat{P}-P)/n\|(\|\hat{A}'x\|+\|A'x\|) + \|x'A\|\|(\hat{P}'\hat{P}-P'P)/n\|\|\hat{A}'x\| \xrightarrow{p} 0.$$

It follows similarly that $x'\hat{Q}\mu/n - x'Q\mu/n \xrightarrow{p} 0$. Therefore,

$$\text{(A.4)} \qquad \|\hat{\mu}-\mu\|^2/n = tr(x'Qx-2x'Q\mu+\mu'\mu)/n + o_p(1) = tr(u'Qu+\mu'(I-Q)\mu)/n + o_p(1).$$

For $T = (\tau_1,\ldots,\tau_n)'$ and $D = (d_1,\ldots,d_n)'$, by independence of the observations, $E[u_i|T,D] = E[d_i(x_i-E[x_i|\tau_i,d_i=1])|\tau_i,d_i] = 0$. Therefore, $E[u_iu_j|T,D] = E[u_iu_j|\tau_i,\tau_j,d_i,d_j] = E[u_iE[u_j|u_i,\tau_i,\tau_j,d_i,d_j]|\tau_i,\tau_j,d_i,d_j] = E[u_iE[u_j|\tau_j,d_j]|\tau_i,\tau_j,d_i,d_j] = 0$. Also, by Assumption 3, $E[u_i'u_i|T,D] = E[u_i'u_i|\tau_i,d_i] \leq C$. Therefore, with probability one,

$$\text{(A.5)} \qquad E[uu'|T,D] \leq CI.$$

It follows that $E[tr(u'Qu)/n|T,D] \leq Ctr(Q)/n = CK/n \longrightarrow 0$, so that $tr(u'Qu)/n \xrightarrow{p} 0$. Also, by Assumption 4 and standard approximation theory results for power series and splines (e.g. see Newey, 1997 for references), and by $(I-Q)P = 0$ and $I-Q$ idempotent, there exists Π_K such that $E[tr(\mu'(I-Q)\mu)]/n = E[tr((\mu-P\Pi_K')'(I-Q)(\mu-P\Pi_K'))]/n \leq E[tr((\mu-P\Pi_K')'(\mu-P\Pi_K'))]/n = E[d_i\{\mu_i-\Pi_K p^K(\tau_i)\}'\{\mu_i-\Pi_K p^K(\tau_i)\}] \longrightarrow 0$. Combining these results with equation (A.4) gives

(A.6) $\|\hat{\mu}-\mu\|^2/n \xrightarrow{p} 0.$

This implies that $\hat{M} - u'u/n \xrightarrow{p} 0$, while $u'u/n \xrightarrow{p} M$ follows by the law of large numbers. The triangle inequality then gives $\hat{M} \xrightarrow{p} M$.

Next, let $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)'$, and $W = [w_1',\dots,w_n']'$. It follows similarly to eq. (A.5) that $E[\varepsilon\varepsilon'|W,D] \le CI$. Then, since $\hat{Q}$ and Q are functions of W and D,

$$E[\|x'(\hat{Q}-Q)\varepsilon/\sqrt{n}\|^2|W,D] = tr\{x'(\hat{Q}-Q)E[\varepsilon\varepsilon'|W,D](\hat{Q}-Q)x\}/n \le Ctr(x'(\hat{Q}-Q)(\hat{Q}-Q)x)/n.$$

It follows similarly to equation (A.3) that $x'(\hat{Q}-Q)\hat{Q}x/n \xrightarrow{p} 0$ and $x'(\hat{Q}-Q)Qx/n \xrightarrow{p} 0$, so that $\|x'(\hat{Q}-Q)\varepsilon/\sqrt{n}\| \xrightarrow{p} 0$, and hence $x'(I-\hat{Q})\varepsilon/\sqrt{n} = x'(I-Q)\varepsilon/\sqrt{n} + o_p(1)$. It follows as in Donald and Newey (1994) that

(A.7) $x'(I-Q)\varepsilon/\sqrt{n} = u'\varepsilon/\sqrt{n} + o_p(1).$

For both power series and splines it follows as in Newey (1997) that there are γ_K and π_K such that for $h_K(\tau) = p^K(\tau)'\gamma_K$ and $\mu_K(\tau) = p^K(\tau)'\pi_K$,

(A.8) $\sup_{|\tau|\le 1}|h_0(\tau)-h_K(\tau)| \le CK^{-s+1}$, $\sup_{|\tau|\le 1}|dh_0(\tau)/d\tau-dh_K(\tau)/d\tau| \le CK^{-s+1}$,

$\sup_{|\tau|\le 1}|\mu(\tau)-\mu_K(\tau)| \le CK^{-t}.$

Let $\tilde{h}_i = h(\hat{\tau}_i)$, $h_i = h(\tau_i)$, $\tilde{h}_{Ki} = h_K(\hat{\tau}_i)$, $h_{Ki} = h_K(\tau_i)$, $\tilde{\mu}_i = \mu(\hat{\tau}_i)$, $\tilde{\mu}_{Ki} = \mu_K(\hat{\tau}_i)$, $\mu_{Ki} = \mu_K(\tau_i)$, and let expressions without the i subscript denote corresponding matrices over all observations multiplied by selection indicators, e.g. $\tilde{\mu}_K = [d_1\tilde{\mu}_{K1},\dots,d_n\tilde{\mu}_{Kn}]'$. Then $x'(I-\hat{Q})h/\sqrt{n} = x'(I-\hat{Q})(h-\tilde{h})/\sqrt{n} + (x-\tilde{\mu}_K)'(I-\hat{Q})(\tilde{h}-\tilde{h}_K)/\sqrt{n}$. Let $\theta = (\alpha',\eta')'$, $\tau(w,\theta) = \tau(v(w,\alpha),\eta)$, and $h_{\theta i} = \partial h(\tau(w_i,\theta_0))/\partial\theta'$. Since $\partial\tau(w,\theta_0)/\partial\eta$ depends only on v and $E[u_i a(v_i)] = 0$ for any function $a(v_i)$ with finite mean-square, $E[u_i h_{\theta i}] = E[u_i\{dh_0(v_i)/dv\}\partial v(w_i,\alpha_0)/\partial\alpha',0] = [H,0]$. It follows similarly to $\hat{M} \xrightarrow{p} M$ that $x'(I-\hat{Q})h_\theta/n \xrightarrow{p} E[u_i h_{\theta i}']$. Then by a second-order expansion and $\sqrt{n}$-consistency of $\hat{\theta}$,

(A.9) $x'(I-\hat{Q})(h-\tilde{h})/\sqrt{n} = -[x'(I-\hat{Q})h_\theta/n]\sqrt{n}(\hat{\theta}-\theta_0) + o_p(1) = -E[u_i h_{\theta i}]\sqrt{n}(\hat{\theta}-\theta_0) + o_p(1)$

$= H\sqrt{n}(\hat{\alpha}-\alpha_0) + o_p(1).$

Also, by eq. (A.8) and $I-\hat{Q}$ idempotent, $(\tilde{\mu}-\tilde{\mu}_K)'(I-\hat{Q})(\tilde{h}-\tilde{h}_K)/\sqrt{n} = O_p(\sqrt{n}K^{-s-t+1}) \xrightarrow{p} 0$. Also, $(\tilde{\mu}-\mu)'(I-\hat{Q})(\tilde{h}-\tilde{h}_K)/\sqrt{n} = O_p(K^{-s+1}) \xrightarrow{p} 0$ and $u'(I-\hat{Q})(\tilde{h}-\tilde{h}_K-h+h_K)/\sqrt{n} = O_p(K^{-s+1}) \xrightarrow{p} 0$. Also, it follows similarly to eq. (A.4) that for $\epsilon_K = h-h_K$, $u'(Q-\hat{Q})\epsilon_K/\sqrt{n} = O_p(\zeta_1(K)K^{-s+1}) \xrightarrow{p} 0$. Also, $E[\|u'Q\epsilon_K\|^2/n|T,D] = \epsilon_K'QE[uu'|T,D]Q\epsilon_K/n \le C\epsilon_K'Q\epsilon_K/n \le \epsilon_K'\epsilon_K/n \to 0$, so that $u'Q\epsilon_K/\sqrt{n} \xrightarrow{p} 0$. The triangle inequality then gives

(A.10) $x'(I-\hat{Q})\hat{h}/\sqrt{n} = (x-\tilde{\mu}_K)'(I-\hat{Q})(\tilde{h}-\tilde{h}_K)/\sqrt{n} \xrightarrow{p} 0.$.

Combining equations (A.7), (A.9), and (A.10), we obtain

(A.11) $x'(I-\hat{Q})(\varepsilon+h)/\sqrt{n} = u'\varepsilon/\sqrt{n} + H\sqrt{n}(\hat{\alpha}-\alpha_0) + o_p(1) = \sum_{i=1}^n (u_i\varepsilon_i + H\psi_i)/\sqrt{n} + o_p(1).$

The first conclusion then follows from the Lindberg-Levy central limit theorem and $E[u_i\varepsilon_i\psi_i'] = E[u_iE[\varepsilon_i|w_i,d_i]\psi_i'] = 0$.

To show the second conclusion, note that $d\hat{h}(\hat{v}_i)/dv = [d\hat{h}(\hat{\tau}_i)/d\tau]d\tau(\hat{v}_i,\hat{\eta})/dv$, and it follows from the Assumption 5 that $\sup_{i\le n}|d\tau(\hat{v}_i,\hat{\eta})/dv - d\tau(v_i,\eta)/dv| = O_p(1/\sqrt{n})$. Also, $\hat{h}(\tau) = p^K(\tau)'\hat{\gamma}$, $\hat{\gamma} = \hat{A}'(y-x\hat{\beta})$. Similarly to eq. (A.3), $\|\hat{A}'x(\hat{\beta}-\beta_0)\| \le O_p(1)\|x(\hat{\beta}-\beta_0)/\sqrt{n}\| = O_p(1/\sqrt{n})$, $\|\hat{A}'(h-\tilde{h})\| \le O_p(1)\|(h-\tilde{h})/\sqrt{n}\| = O_p(1/\sqrt{n})$, and $\|\hat{A}'(\tilde{h}-\hat{P}\gamma_K)\| \le O_p(1)\|(\tilde{h}-\hat{P}\gamma_K)/\sqrt{n}\| = O_p(K^{-s+1})$. Similarly to previous results, $E[\varepsilon'\hat{Q}\varepsilon|D,W] \le CK$, so that $\|\hat{A}'\varepsilon\|^2 = \varepsilon'\hat{A}\hat{A}'\varepsilon = O_p(1)\varepsilon'\hat{Q}\varepsilon/n = O_p(K/n)$. Then by $\hat{\gamma}-\gamma_K = \hat{A}'x(\hat{\beta}-\beta_0) + \hat{A}'\varepsilon + \hat{A}'(h-\tilde{h}) + \hat{A}'(\tilde{h}-\hat{P}\gamma_K)$ and the triangle inequality, $\|\hat{\gamma}-\gamma_K\| = O_p((K/n)^{1/2}) + O_p(K^{-s+1})$. Then for $s = 1$ or 2,

(A.12) $\sup_{|\tau|\le 1}|d^s\hat{h}(\tau)/d\tau^s - d^sh_0(\tau)/d\tau^s| \le \sup_{|\tau|\le 1}|[d^sp^K(\tau)/d\tau^s]'(\hat{\gamma}-\gamma_K)|$

$+ \sup_{|\tau|\le 1}|d^s[p^K(\tau)'\gamma_K]/d\tau^s - d^sh_0(\tau)/d\tau^s| \le \zeta_s(K)\|\hat{\gamma}-\gamma_K\| + O(K^{-s+1})$

$= O_p(\zeta_s(K)[(K/n)^{1/2}+K^{-s+1}]) = o_p(1).$

It follows that $\max_{i\leq n}|d\hat{h}(\hat{\tau}_i)/d\tau - dh_0(\hat{\tau}_i)/d\tau| \xrightarrow{p} 0$. Also, since the conditions require that $h_0(\tau)$ be at least twice differentiable with bounded derivative, $\max_{i\leq n}|dh_0(\hat{\tau}_i)/d\tau - dh_0(\tau_i)/d\tau| \xrightarrow{p} 0$, implying $\max_{i\leq n}|d\hat{h}(\hat{v}_i)/dv - dh_0(v_i)/dv| \xrightarrow{p} 0$. Then, by boundedness of $\partial v(w_i,\hat{\alpha})/\partial\alpha$, for $\tilde{H} = n^{-1}\sum_{i=1}^n \hat{u}_i[\partial\tau(w_i,\hat{\alpha})/\partial\alpha']dh_0(v_i)/dv$,

$$\|\hat{H}-\tilde{H}\| \leq tr(\hat{u}'\hat{u}/n)^{1/2}(\sum_{i=1}^n \|\partial\tau(w_i,\hat{\alpha})/\partial\alpha\|^2/n)^{1/2}\max_{i\leq n}|dh_0(\hat{\tau}_i)/d\tau - dh_0(\tau_i)/d\tau| \xrightarrow{p} 0.$$

It also follows by eq. (A.3) that for $\bar{H} = n^{-1}\sum_{i=1}^n u_i[\partial\tau(w_i,\alpha_0)/\partial\alpha']dh_0(v_i)/dv$, $\|\tilde{H}-\bar{H}\| \xrightarrow{p} 0$. Then since $\bar{H} \xrightarrow{p} H$ by the law of large numbers, $\hat{H} \xrightarrow{p} H$ follows by the triangle inequality.

Now, let $\hat{\Delta}_i = x_i'(\hat{\beta}-\beta_0) + \hat{h}_i - h_i$. By eq. (A.12), $\max_{i\leq n}|\hat{h}_i - h_i| \leq \max_{i\leq n}|\hat{h}(\hat{\tau}_i) - h_0(\hat{\tau}_i)| + \max_{i\leq n}|h_0(\hat{\tau}_i) - h_0(\tau_i)| \xrightarrow{p} 0$. Also, $\max_{i\leq n}|x_i'(\hat{\beta}-\beta_0)| \leq \max_{i\leq n}\|x_i\|\|\hat{\beta}-\beta_0\| \leq n^{1/(2+\delta)}(\sum_{i=1}^n \|x_i\|^{2+\delta}/n)^{1/(2+\delta)}O_p(1/\sqrt{n}) = n^{1/(2+\delta)}O_p(1)O_p(1/\sqrt{n}) \xrightarrow{p} 0$. Then by the triangle inequality $\max_{i\leq n}|\hat{\Delta}_i| \xrightarrow{p} 0$. Furthermore, by Assumption 3, $E[|\varepsilon_i||W,D] \leq C$, so that $E[\sum_{i=1}^n \|\hat{u}_i\|^2|\varepsilon_i|/n|W,D] = \sum_{i=1}^n \|\hat{u}_i\|^2 E[|\varepsilon_i||W,D]/n \leq C\sum_{i=1}^n \|\hat{u}_i\|^2/n = O_p(1)$, and hence $\sum_{i=1}^n \|\hat{u}_i\|^2|\varepsilon_i|/n = O_p(1)$. Therefore,

$$\text{(A.13)} \qquad \|\sum_{i=1}^n \hat{u}_i\hat{u}_i'\hat{\varepsilon}_i^2/n - \sum_{i=1}^n \hat{u}_i\hat{u}_i'\varepsilon_i^2/n\| \leq \sum_{i=1}^n \|\hat{u}_i\|^2|\hat{\varepsilon}_i^2-\varepsilon_i^2|/n = \sum_{i=1}^n \|\hat{u}_i\|^2|(\varepsilon_i-\hat{\Delta}_i)^2-\varepsilon_i^2|/n$$

$$2(\sum_{i=1}^n \|\hat{u}_i\|^2|\varepsilon_i|/n)\max_{i\leq n}|\hat{\Delta}_i| + (\sum_{i=1}^n \|\hat{u}_i\|^2/n)\max_{i\leq n}|\hat{\Delta}_i|^2 \xrightarrow{p} 0.$$

Also, note that $E[\sum_{i=1}^n \|\varepsilon_i\hat{u}_i - \varepsilon_i u_i\|^2/n|W,D] = E[\sum_{i=1}^n \varepsilon_i^2\|\hat{\mu}_i-\mu_i\|^2/n|W,D] \leq \sum_{i=1}^n E[\varepsilon_i^2|W,D]\|\hat{\mu}_i-\mu_i\|^2/n \leq C\sum_{i=1}^n \|\hat{\mu}_i-\mu_i\|^2/n \xrightarrow{p} 0$ by equation (A.6). Therefore, $\sum_{i=1}^n \|\varepsilon_i\hat{u}_i - \varepsilon_i u_i\|^2/n \xrightarrow{p} 0$. It follows that $\sum_{i=1}^n \hat{u}_i\hat{u}_i'\hat{\varepsilon}_i^2/n - \sum_{i=1}^n u_iu_i'\varepsilon_i^2/n \xrightarrow{p} 0$. Then by the law of large numbers, $\sum_{i=1}^n u_iu_i'\varepsilon_i^2/n \xrightarrow{p} E[u_iu_i'\varepsilon_i^2] = \Omega$, so by the triangle inequality, $\sum_{i=1}^n \hat{u}_i\hat{u}_i'\hat{\varepsilon}_i^2/n \xrightarrow{p} \Omega$. The second conclusion then follows by consistency of $\hat{V}(\hat{\alpha})$ and the Slutzky theorem.

References

Ahn, H. and J.L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics* 58, 3-29.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 32, 189-218.

Cosslett, S.R. (1983): "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* 51, 765-782.

Cosslett, S.R. (1991): "Distribution-Free Estimator of a Regression Model With Sample Selectivity," in W.A. Barnett, J.L. Powell and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Cambridge, Cambridge University Press.

Donald, S.G. and W. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis* 50, 30-40.

Gallant, A.R. and D.W. Nychka (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

Gronau, R. (1973): "The Effects of Children on the Housewife's Value of Time," *Journal of Political Economy* 81, S168-S199.

Heckman, J.J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica* 42, 679-693.

Heckman, J.J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Vairables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475-492.

Heckman, J.J. and R. Robb (1987): "Alternative Mehods for Evaluating the Impact of Interventions," Ch. 4 of *Longitudinal Analysis of Labor Market Data*, J.J. Heckman and B. Singer eds., Cambridge, UK: Cambridge University Press.

Ichimura, H. (1993). Estimation of single index models. *Journal of Econometrics* 58, 71-120.

Klein, R.W. and R.S. Spady (1993): "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica* 61, 387-421.

Lee, L.F. (1982): "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies* 49, 355-372.

Linton, O. (1995): "Second Order Approximation in a Partially Linear Regression Model," *Econometrica* 63, 1079-1112.

Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3, 205-228.

Newey, W.K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

Newey, W.K. and J.L. Powell (1993): "Efficiency Bounds for Semiparametric Selection Models," *Journal of Econometrics* 58, 169-184.

Newey, W.K., J.L. Powell, and J.R. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review Papers and Proceedings*, May.

Powell, J.L. (1994): "Estimation of Semiparametric Models," in R.F. Engle and D. McFadden, eds., *Handbook of Econometrics: Volume 4*, New York: North-Holland.

Powell, J.L., J.H. Stock, and T.M. Stoker (1989). Semiparametric Estimation of Index Coefficients *Econometrica* 57, 1403-1430.

Powell, J.L. (1987): "Semiparametric Estimation of Bivariate Limited Dependent Variable Models," manuscript, University of California, Berkeley.

Powell, M.J.D. (1981): *Approximation Theory and Methods*, Cambridge, UK, Cambridge University Press.

Robinson, P. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica* 56, 931-954.

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

Stone, C.J. (1985): "Additive Regression and Other Nonparametric Models, *Annals* of *Statistics* 13, 689-705.

White, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review* 21, 149-170.