On Intercept Estimation in the Sample Selection Model

Marcia M.A. Schafgans[¤]
and
Victoria Zinde-Walsh[yz]

October 1999

Abstract. We provide a proof of the consistency and asymptotic normality of the estimator suggested by Heckman (1990) for the intercept of a semiparametrically estimated sample selection model. The estimator is based on "identi...cation at in...nity" which leads to non-standard convergence rate. Andrews and Schafgans (1998) derived asymptotic results for a smoothed version of the estimator. We examine the optimal bandwidth selection for the estimators and derive asymptotic MSE rates under a wide class of distributional assumptions. We also provide some comparisons of the estimators and practical guidelines.

Keywords: Asymptotic normality, sample selection model, semiparametric estimation.
JEL Classi...cation: C14, C34, C35.

1. Introduction

Semiparametric estimation of sample selection models has attracted considerable interest in the last decade. More recently the estimation of the intercept of the semiparametric estimated sample selection model has received due attention, see Heckman (1990) and Andrews and Schafgans (1998).

The discussion around the estimation of the intercept arose, since, with the exception of Gallant and Nychka (1978), all semiparametric estimation approaches to the sample selection model precluded the estimation of the intercept; the intercept was absorbed in the nonparametric sample selection bias correction term. The semi-nonparametric estimator proposed by Gallant and Nychka (1978), however, has a drawback in that although it is consistent, its asymptotic distribution is unknown.

The importance of this intercept is evident, e.g., when using the sample selection model in the evaluation of social programs. Estimation of the intercept allows one to evaluate the net bene...t of a social program, by allowing one to compare the actual outcome of participants with the expected outcome had they chosen not to participate. Empirically, the estimation of the intercept of semiparametrically estimated sample selection models has proven desirable in the estimation of wages. Its estimation allows for a decomposition of the wage-gap between socio-economic groups (e.g., male–female) in order to assess the extent of "discrimination" (Schafgans (1998a) and allows for a discussion of its evolution over time (Buchinsky (1998)).

In Andrews and Schafgans (1998), the ...rst consistent and asymptotically normal estimator was derived for the intercept, β_0. Their estimator was based on a suggestion by Heckman (1990) to estimate β_0 using only those observations for which the probability of selection in the truncated or censored sample is close to one and in the limit as $n \to 1$ is one. The justi...cation of this approach is that the conditional mean of the errors in the outcome equation for the observations having probability of selection close to one is close to zero. Due to the di¢culty in deriving the asymptotic distribution of the Heckman (1990) estimator, arising from the non-di¤erentiability of the indicator function, Andrews and Schafgans (1998) introduced a smooth monotone [0,1]-valued function, $s(\cdot)$. Since we will make reference to the Andrews and Schafgans (1998) paper frequently, we will call it AS henceforth.

In this paper, we derive the consistency and asymptotic distribution of the Heckman estimator itself. This provides the empirical researcher with the advantage of not having to specify the smoothing function introduced by Andrews and Schafgans. We investigate a wide class of distributional assumptions for the model and derive "optimal" bandwidth parameters and corresponding asymptotic rates for mean squared error (MSE) for the two estimators. Since the solution for the optimal bandwidth may not be practical, we provide simpler bounds on the optimal bandwidth parameter; using a bound may imply preference for either AS or the Heckman estimator.

The remainder of this paper is organized as follows: Section 2 introduces the sample selection model considered and the estimators of Heckman (1990) and AS. The asymptotic normality result for the Heckman estimator is given in Section 3. Section 4 discusses the optimal selection of the bandwidth and the rate of the asymptotic mean squared error. Section 5 concludes. Various appendices follow. Appendix A contains the proof of the asymptotic normality result given in Section 3. Appendix B derives the asymptotic bias and variance for the two estimators under a class of general distributional assumptions and the optimal bandwidth choices given in Section 4.

2. Intercept Estimation

The sample selection model that we consider can be written as:

$$Y_i^* = \gamma_0 + Z_i'\mu_0 + U_i \; ;$$
$$D_i = 1(X_i'\bar{\beta}_0 > \varepsilon_i) \; ; \quad \text{and} \tag{1}$$
$$Y_i = Y_i^* D_i \quad \text{for} \quad i = 1, ..., n \; ;$$

where $(Y_i; D_i; Z_i; X_i)$ are observed random variables. The first equation is the outcome equation and the second equation is the participation equation. For convenience, we set

$$W_i = X_i'\bar{\beta}_0 : \tag{2}$$

The literature on semiparametric estimation of sample selection models gives several root-n consistent and asymptotically normal estimators for the selection parameters, $\bar{\beta}_0$ (up to some unknown scale), and the slope parameters of the outcome equation, μ_0. For instance, one could consider: Ichimura (1993), Han (1987), Newey (1988), Robinson (1988), Powell (1989), Powell, Stock, and Stoker (1989), Ichimura and Lee (1990), Andrews (1991), and Klein and Spady (1993). The existing literature and AS can deal both with censored samples (as in the model given in (1)), or truncated samples. In the latter case, that is where Y_i is observed only if $D_i = 1$, β_0 and $\bar{\beta}_0$ need to be estimated simultaneously using, e.g., Ichimura and Lee (1990). Regarding the selection parameters $\bar{\beta}_0$; furthermore, it should be noted that only the slope parameters are required in the context of estimating the intercept γ_0. The loss of identi…cation of the intercept in the selection equation, e.g., when using Ichimura (1993) or Ichimura and Lee (1990) is innocuous therefore.

A consistent and asymptotically normal estimator for the intercept, γ_0, which uses these preliminary estimators, was provided by AS. Their estimator, call it the AS estimator, is given by

$$\hat{\gamma}_s = \frac{\sum_{i=1}^n (Y_i - Z_i'\hat{\mu})D_i s(X_i'\mathbf{b} - \sigma_n)}{\sum_{i=1}^n D_i s(X_i'\mathbf{b} - \sigma_n)}; \tag{3}$$

where $s(\phi)$ is a non-decreasing $[0,1]$-valued function that has three derivatives bounded over R and for which $s(x) = 0$ for $x \cdot 0$ and $s(x) = 1$ for $x \geq b$ for some $0 < b < 1$ (AS, Assumption 3). The preliminary estimators $(\hat{\mu}; \mathbf{b})$ are root-n consistent estimators of $(\mu_0; \bar{\beta}_0)$. The parameter σ_n is called the bandwidth or smoothing parameter, where the bandwidth parameter is chosen such that $\sigma_n \to 1$ as $n \to 1$.

The Heckman (1990) estimator, on which the AS estimator was based, is given by

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n (Y_i - Z_i'\hat{\mu})D_i 1(X_i'\mathbf{b} > \sigma_n)}{\sum_{i=1}^n D_i 1(X_i'\mathbf{b} > \sigma_n)} \tag{4}$$

Comparing the two formulae (3) and (4), it is clear that the AS estimator $\hat{b}_s$ di¤ers from Heckman's (1990) $\hat{\gamma}_I$ only in that it replaces the indicator function $1(\cent)$ with a smooth function $s(\cent)$.

Heckman's estimator $\hat{\gamma}_I$ is essentially a sample average of the random variables $U_i + {}^1{}_0$ over a fraction of all observations, since $Y_i \; Z_i^0\hat{\vartheta} \; !_p \; U_i + {}^1{}_0$ as $n! \; 1$ for all $i \; \, 1$. The e¤ective sample size is equal to the number of observations used for the estimation of ${}^1{}_0$: Since AS introduced a weighting scheme for these observations, viz., the smooth function $s(\cent)$; the estimator $\hat{b}_s$ is a weighted sample average of the random variables $U_i + {}^1{}_0$; where observations with $X_i^0\mathbf{b}$ greater than $°_n$ and with $X_i^0\mathbf{b}$ close to the threshold $°_n$ are weighted less than those further away.

Estimation using the AS or Heckman estimator involves two choices, that of the bandwidth parameter $°_n$ and that of a function $s(\cent)$ (or $1(\cent)$). It is clear that the choice of $°_n$ has the most important consequences for the properties of the estimator while the impact of the function $s(\cent)$ is small in comparison. This is con...rmed in the analysis of Section 4; nevertheless there are cases when the choice of $s(\cent)$ (or $1(\cent)$) a¤ects the asymptotic rate of the MSE; results are presented in Section 4.

First, we turn to our asymptotic normality result for the Heckman estimator.

3. Asymptotic normality of the Heckman estimator

Here we prove the conjecture made by Andrews and Schafgans that the Heckman estimator also is asymptotically normal. In the unrealistic case where the true μ_0 and $\bar{}_0$ are known, Andrews and Schafgans already showed that the Heckman estimator, $\hat{\gamma}_{I;0}$, is asymptotically normal (i.e., in Andrews and Schafgans (1998) the indicator replaces the $s(\cent)$ function when the true μ_0 and $\bar{}_0$ are known).[1]

For our purposes, all we need to show now is that

$$\frac{\mathbf{p}\overline{n E D_i 1(W_i > °_n)}}{\tfrac{3}{4}} \; i \; \left(\hat{\gamma}_I \; i \; \hat{\gamma}_{I;0} \right) \; !^p \; 0; \tag{5}$$

where $\tfrac{3}{4}^2 = \text{Var}(U_i)$. Essentially, the proof requires us to deal directly with the non-di¤erentiability of the indicator function.

There are di¤erent ways of dealing with asymptotics for non-di¤erentiable functions. Typically assumptions regarding the probability density function are required. This is due to the fact that the expectation of the Dirac $\pm$-function, which is the generalized derivative of the indicator function, equals the value of the p.d.f. at zero. In our case we need to consider the non-di¤erentiable function $®_{in}$ given by

$$®_n({}^\triangle; W_i; X_i) \text{ with } ®_n(\bar{}; W; X) = 1(X^{0\bar{}} > °_n) \; i \; 1(W > °_n); \tag{6}$$

---

[1] The estimator $\hat{\gamma}_{I;0}$ is identical to $\hat{\gamma}_I$ with the preliminary estimators replaced by their true values.

where $W = X^{0-}_0$. Transform X_i via a linear transformation into the random vector partitioned as $(W_i; \xi_i)$: For our purposes, it will be convenient to let $\xi_i = X_{i(-1)}$; where $X_{i(-1)}$ is X_i with the exclusion of its ...rst component.

We add the following Assumption A to the Assumptions 1–7 of AS.[2]

Assumption A: (a) For some $A > 0$ the conditional probability density function $p:d:f:_{W;\xi_i}$ exists for all $W > A$ and declines monotonically. The marginal probability density function $p:d:f:_W$ is such that for some $d \,!\, 0$; $dn^{1-6} \,!\, 1$

$$\frac{p:d:f:_W(^\circ_n \,i\, d)}{\Pr(W_i > ^\circ_n)^{3-4}} = O(1):$$

(b) For any $W_i > A$ the conditional moment $E_{jW=W_i}(^\circ X_{i(-1)}^{\circ 3})$ exists.

Similar to Assumptions 4 and 7 of AS, Assumption A(a) relates to the upper tail behaviour of the selection index $W_i = X^{0-}_{i\ 0}$. It is satis...ed if W_i has a Weibull, Pareto or "combined" upper tail. If Assumption 4 of AS is satis...ed with $\gg = 0$, then this condition can be replaced by:

$$\frac{p:d:f:_W(^\circ_n)}{\Pr(W_i > ^\circ_n)^{3-4}} = O(1):$$

In both cases, the condition is less strong than requiring a bounded hazard function on W_i; since $p:d:f:_W(^\circ_n \,i\, d)^{1-4} = o(1)$.

The second part of Assumption A requires the existence of the conditional moment, but does not place any restrictions on its behaviour as a function of W_i. This assumption is satis...ed, for example, if the unbounded components of X_i have a joint normal or spherical distribution with W_i:

The following theorem summarizes our result for the Heckman estimator which satis...es Assumption 3^0 of AS.

Theorem 1: Under Assumptions 1, 2, 4–7 of AS and Assumption A

(a) $\dfrac{\beta\dfrac{nE\,D_i 1(W_i > ^\circ_n)}{\frac{3}{4}}\mu}{\beta\dfrac{nE\,D_i 1(W_i > ^\circ_n)}{\frac{3}{4}}}\left(\hat{\,}_1 \,i\, {}_0 \,i\, \dfrac{E\,U_i D_i 1(W_i > ^\circ_n)}{E\,D_i 1(W_i > ^\circ_n)}\right) \,i\, \overset{d}{!} \, N(0;1)$

(b) $\dfrac{\beta\dfrac{nE\,D_i 1(W_i > ^\circ_n)}{\frac{3}{4}}}{}(\hat{\,}_1 \,i\, {}_0) \,i\, \overset{d}{!} \, N(0;1)$ i¤ Assumption 8 of AS holds.

---

[2] Essentially, Assumptions 1 and 2 of AS require existence of moments and independence between $(U_i; "_i)$ and $(Z_i; X_i)$; Assumption 3 and 3^0 deal with the shape of $s(\phi)$; Assumption 4 characterizes the upper tail of W_i in terms of a parameter $0 \cdot \gg < 1=3$ with "fatter" tails if $\gg = 0$; Assumption 5 is root-n consistency and asymptotic normality of $(\hat{\beta}; \hat{b})$; Assumption 6 is $^\circ_n \,!\, 1$; where its speed is restricted by Assumption 7 in terms of the tail of $W_i > ^\circ_n$:

Proof: See Appendix A.   2

To test hypotheses and construct con...dence intervals for functions of $(\gamma_0; \mu_0; \bar{\ }_0)$, we need a joint asymptotic normality result for $(\hat{\gamma}_I; \hat{\beta}; \mathbf{b})$. This result, similar to that in AS, is given by

Theorem 2: Under Assumptions 1, 2, 4–8 of AS and Assumption A

$$\left[\frac{\sqrt{nE D_i 1(W_i > \circ_n)}}{p\frac{3}{4} \sum_{i=1}^{n} \begin{bmatrix} \hat{\beta}_i \mu_0 \\ \mathbf{b}_i \bar{\ }_0 \end{bmatrix}} (\hat{\gamma}_I ¡ \gamma_0) \right] C_i \xrightarrow{d} N(0; I):$$

Proof: See Appendix A.   2

In the following section, we compare the performance of the estimators and provide guidelines for selection of the bandwidth parameter and function $s(\text{\textcent})$ (or $1(\text{\textcent})$).

4.   Bandwidth selection and comparison of the estimators

Here we use the asymptotic MSE as a criterion for bandwidth selection and choice of the estimator. Two characteristics of the model are of importance for these choices: the tail behaviour of the selection index, W_i, and the tail behaviour of the function $!(W)$ de...ned below that determines the asymptotic bias of the estimator. Speci...cally, let

$$!(W) = E_{jW=W} U_i 1(\text{"}_i > W): \tag{7}$$

The asymptotic bias (abias) of the estimator $\hat{\mathbf{b}}_s$ (or $\hat{\mathbf{b}}_I$ for $s(\text{\textcent}) = 1(\text{\textcent})$) is given by[3]

$$abias(\hat{\mathbf{b}}_s) = \frac{¡ E[!(W_i)s(W_i ¡ \circ_n)]}{[ED_i s(W_i ¡ \circ_n)]}: \tag{8}$$

Under the bivariate normality assumption of $(U_i; \text{"}_i)$, $!(W)$ is equal to $\frac{\frac{3}{4}\text{"}_u}{\frac{3}{4}\text{"}}\acute{A}(\frac{W}{\frac{3}{4}\text{"}})$; where $\acute{A}(\text{\textcent})$ denotes the standard normal density function.

There are circumstances when $!(W)$ may be zero for large enough W, e.g. if the distribution of "_i has ...nite support, in which case a comparison of the asymptotic variance will determine the optimal estimator. When $!(W)$ di¤ers from zero for large W we need to establish the importance of the asymptotic bias relative to the asymptotic variance to select our estimator, which of course will depend on the behaviour of this conditional expectation. From Theorem 2 in AS and Theorem 1 in this paper,

---

[3] Since $E(U_i s(W_i > \circ_n)) = 0$ by independence of U_i and W_i, this follows as $E(U_i D_i s(W_i > \circ_n)) = ¡ E(U_i (1 ¡ D_i) s(W_i > \circ_n)) = ¡ E(U_i 1(\text{"}_i > W_i) s(W_i > \circ_n))$: By law of iterated expectations, this equals $¡ E(!(W_i) s(W_i > \circ_n))$.

the asymptotic variance (avar) of the estimator $\hat{\beta}_s$ (or $\hat{\beta}_I$ for $s(\cent) = 1(\cent)$) is given by

$$\text{avar}(\hat{\beta}_s) = \frac{\frac{3}{4}^2 E\,[D_i s^2(W_i \text{ } \text{ } °_n)]}{n\,[E\,D_i s(W_i \text{ } °_n)]^2} \qquad (9)$$

Proposition 1 shows that there may be a trade-off between asymptotic variance and asymptotic bias depending on the choice of function $s(\cent)$ (or $1(\cent)$). Here and below, "$a(x) \frac{1}{4} b(x)$" is defined to mean that $a(x) = b(x)(1 + o(1))$ as $x \to 1$.

Proposition 1: Under Assumptions 1–7 of AS and Assumption A, for a given sequence $°_n$

(a) $\text{avar}(\hat{\beta}_s) \geq \text{avar}(\hat{\beta}_I)$

(b) $|\text{abias}(\hat{\beta}_s)| \cdot |\text{abias}(\hat{\beta}_I)|$; if $E1(W_i > °_n) = E1(W_i > °_n + b) \frac{1}{4} 1$ ($\varkappa = 0$ in Assumption 4 of AS) and $\lambda(W_i) \geq 0\,\forall W_i > °_n$ (or $\lambda(W_i) \cdot 0\,\forall W_i > °_n$):

Proof: See Appendix B. ☐

If there is no asymptotic bias, naturally the Heckman estimator is preferred based on asymptotic MSE (and variance).

We can characterize the tail behaviour of the selection index W_i as "fat-tailed" if Assumption 4 of AS is satisfied with $\varkappa = 0$; if Assumption 4 of AS is satisfied only with $\varkappa > 0$, we say that W_i is "thin-tailed". Examples of fat tails of W_i are Pareto upper tails (i.e., $1 - F(W) \frac{1}{4} c_W W^{-\varsigma}; \varsigma > 0$) or Weibull ($\varsigma; c$) upper tails (i.e., $1 - F(W) \frac{1}{4} c_W \exp(-\varsigma W^c); \varsigma > 0$) with $c \cdot 1$: For Pareto and fat Weibull tails of W the condition in (b) is satisfied.

Proposition 2 shows that for fat-tailed distributions of W_i the choice of the function $s(\cent)$ or $1(\cent)$ does not affect the asymptotic variance; the asymptotic bias if unaffected if additionally $\lambda(W_i)$ does not go to zero too fast.

Proposition 2: Under Assumptions 1–7 of AS and Assumption A,

(a) If $E1(W_i > °_n) = E1(W_i > °_n + b) \frac{1}{4} 1$ ($\varkappa = 0$ in Assumption 4 of AS) $\text{avar}(\hat{\beta}_s) \frac{1}{4}$ $\text{avar}(\hat{\beta}_I)$:

(b) If additionally to (a), $|E\lambda(W)1(W > °_n + b)| = |E\lambda(W)1(W > °_n)| \frac{1}{4} 1$ and $\lambda(W_i) \geq 0\,\forall W_i > °_n$ (or $\lambda(W_i) \cdot 0\,\forall W_i > °_n$); then $\text{abias}(\hat{\beta}_s) \frac{1}{4} \text{abias}(\hat{\beta}_I)$:

Proof: See Appendix B. ☐

Unlike the assumptions encompassed in our Proposition 2, it is frequently assumed that W_i has thinner upper tails, e.g., the normal (Lee (1982)). We next examine

the AS and Heckman estimators for a class of models with tails of the selection index $1 - F(W) \approx c_W W^{\circledR} e^{-\gamma W^c}$ and $\lambda(W) \approx c_\lambda W^{\&} e^{-\eta W^d}$ where the parameters $\circledR; \gamma; c; \&; \eta; d$ are such that the functions $1 - F(W)$ and $\lambda(W) \to 0$ as $W \to \infty$: This class of models includes the Pareto, Weibull (with $c \le$ or $c > 1$) as well as "combined" tails. If $U; V$ and W are jointly normally distributed $d = c = 2; \circledR = -1; \gamma = \frac{1}{2\sigma_W^2}; \eta = \frac{1}{2\sigma^2}$; and $\& = 0.$[4] In order to facilitate the derivation of the asymptotic mean squared error for this class of distributions, we restrict our attention to $s(\cent)$ functions satisfying the following assumption

Assumption S: Let $s(\cent)$ be a function satisfying Assumption 3 of AS. For some q its derivatives at zero are such that

$$s^{(i)}(0) = \begin{cases} 0 & i < q \\ a_q \ne 0 & i = q \\ \text{exists} & i = q + 1: \end{cases}$$

Note that any function that satis…es AS for which the lowest order of non-zero derivative is $q \le 2$ satis…es Assumption S as well; it is only functions with two (or more) zero derivatives at 0 that require this additional assumption.

The following proposition provides expressions for the asymptotic variance and asymptotic bias. To simplify the expressions in Proposition 3 we omit the constant factors, if they are present the expressions below for avar acquire c_W^{i-1} and for abias c_λ as a factor. Furthermore, we omit the subscript n on $\circ$.

Proposition 3: Under Assumptions 1, 2, 5–7 of AS, Assumptions A and S

(a) If $1 - F(W) \approx W^{\circledR} \exp(-\gamma W^c)$

$$\operatorname{avar}(\hat{\tau}_s) = \begin{cases} \operatorname{avar}(\hat{\tau}_l) \approx \sigma^2 n^{-1 \circ i - \circledR} \exp(\gamma \circ^c) & \text{if } c \le 1 \\ \sigma^2 n^{-1 - \frac{2q}{q} \circ i - \circledR} \exp(\gamma \circ^c) & \text{if } c > 1; \end{cases}$$

where $\operatorname{avar}(\hat{\tau}_l)$ when $c > 1$ obtains for $q = 0$ ($\binom{0}{0} = 1$).

(b) If additionally to (a) $\lambda(W) \approx W^{\&} \exp(-\eta W^d)$

---

[4] In AS it was mistakenly claimed that the normal distribution has a Weibull tail with $c = 2$; in fact its tail is $W^{-1} \exp(-\frac{1}{2\sigma_W^2} W^2)(1 + o(W^{-1}))$:

$$
\text{abias}(\hat{\theta}_s) = \begin{cases}
\text{abias}(\hat{\theta}_I) = -\,{}^{\circ\&}\exp(-^{1\circ d}) & \text{if } c \cdot 1; d \cdot 1; d < c \\[4pt]
\text{abias}(\hat{\theta}_I) = -\,\frac{}{{}_{\cdot}+1}{}^{\circ\&}\exp(-^{1\circ c}) & \text{if } c \cdot 1; d \cdot 1; c = d^5 \\[4pt]
\text{abias}(\hat{\theta}_I) = -\,\frac{c}{1d}{}^{\circ\&+c-d}\exp(-^{1\circ d}) & \text{if } c \cdot 1; d \cdot 1; d > c^6 \\[4pt]
-\,\frac{a_{q\,,}c}{(1d)^{q+1}}{}^{\circ\&+(c-d)(1+q)}\exp(-^{1\circ d}) & \text{if } c \cdot 1; d > 1^6 \\[4pt]
-\,{}_3^{\circ\&}\exp(-^{1\circ d}) & \text{if } c > 1; d < c; \\[4pt]
-\,\frac{}{3^{\;}{}_{\cdot}+1}{}^{\circ\&}\exp(-^{1\circ c}) & \text{if } c > 1; d = c \\[4pt]
-\,\frac{c}{1d}^{q+1}{}^{\circ\&+(c-d)(1+q)}\exp(-^{1\circ d}) & \text{if } c > 1; d > c;
\end{cases}
$$

where $\text{abias}(\hat{\theta}_I)$ obtains for $q = 0$ ($0! = 1$) where it is not de...ned explicitly and $a_0 \,\acute{}\, 1$.

Proof: See Appendix B.   2

We see that under our assumptions on $s(x)$, the asymptotic MSE is a¤ected by the choice of function $s(\textcent)$ via q and the value of the derivative $s^{(q)}(0)$ only. When $\text{avar}(\hat{\theta}_s)$ depends on q it is an increasing function of q, while if $\text{abias}(\hat{\theta}_s)$ depends on q, its absolute value declines with q.

As an example of Proposition 3, if $U; \varepsilon$ and W are jointly normally distributed ($c = d = 2$), the asymptotic bias and variance of $\hat{\theta}_s$ (including all relevant constant factors) equals

$$
\text{abias}(\hat{\theta}_s) \;=\; -\,\text{\DJ}\frac{\sigma_{U}}{2\sigma\sigma_{\cdot}}\Big(\frac{\sigma_{\ddot{}}^2}{\sigma_{\ddot{}}^2 + \sigma_W^2}\Big)^{q+1}\exp(-\,\frac{1}{2\sigma_{\ddot{}}^2}{}^{\circ 2})
$$

$$
\text{avar}(\hat{\theta}_s) \;=\; \frac{\sigma^2}{n}\frac{\text{\DJ}\frac{}{2\sigma}}{\sigma_W}\,{}_{\circ}\,{}^{2q}\Big|\frac{}{q}\Big|\exp(\frac{1}{2\sigma_W^2}{}^{\circ 2}):
$$

The asymptotic bias and variance of $\hat{\theta}_I$ under the joint normality assumption obtains for $q = 0$:

If all the parameters determining the tail behaviour of W and the function $!(W)$ were known a solution that would provide an optimal $^{\circ\mathtt{¤}} = \text{argmin}(MSE)$ as a function of n, q (the $s(\textcent)$ function) and all those parameters could be obtained (at least via a numerical algorithm) from the formulae in Proposition 3. If the asymptotic bias is not present (in which case one would choose the Heckman estimator based on Proposition 1) the bandwidth parameter arising from reducing MSE (or equivalently avar) as $^{\circ} ! 1$ can be presented as $^{\circ\mathtt{¤}} = (\frac{\mu}{{}_{,}}\ln n)^{1=c}$ if $c \,\mathbf{6}\, 0$ and $^{\circ\mathtt{¤}} = n^{-\mu=\circledR}$

---

[6] If $c = d = 0$, the constant $\frac{}{{}_{,}+1}$ becomes $\frac{\circledR}{\circledR+\&}$:

[6] If $c = 0$, then the expression $_{,}c$ in the constant becomes $-\circledR$:

if $c = 0$ with μ close to zero and would result in a MSE proportional to $n^{-1+\mu}$. Proposition 4 deals with situations where an asymptotic bias is present and may be severe. It characterizes the bandwidth parameter σ^* and the best possible rate for MSE depending on the relation between the rate of decline in the tail of W and the function $!(W)$; we also provide simple bounds on σ^* which bring MSE close to achieving the best possible rate.

Proposition 4: Under Assumptions 1, 2, 5–7 of AS and Assumptions A and S, if $1 - F(W) \sim c_W W^{\circledR} e^{-\gamma W^c}$ and $!(W) \sim c_! W^{\&} e^{-\gamma_1 W^d}$ as $W \to \infty$:

(a) There exists a sequence σ^*_n unique up to $o(\sigma^{*-v})$ for some $v > 0$ that minimizes the asymptotic $MSE(\mathbf{1}_s)$ (or $\mathbf{1}_!$):

(b) The optimal asymptotic MSE^* can be represented as a product of a polynomial component $n^{-\zeta}; \zeta \geq 0$ and a logarithmic component $O((\ln n)^\circ)$; where ζ depends only on the parameters which characterize the leading term in the tail of W, i.e. $\circledR$ for a Pareto and $\gamma; c$ for a Weibull or combined tail, and parameters of the leading term of $!(W)$:

(c) There exist bounds σ_H and σ_L such that $\sigma_L < \sigma^* < \sigma_H$; where σ_H and σ_L are functions of the coefficients of the leading terms in tail of W and $!(W)$ only and $MSE(\sigma_H)$, $MSE(\sigma_L)$ decline at a rate with the polynomial component $n^{-\zeta(H)}; n^{-\zeta(L)}$ with one (or both) of $\zeta(H)$ and $\zeta(L)$ either equal to ζ; or arbitrarily close to ζ:

(d) When $0 \leq d < c$, that is $!(W)$ goes to zero exponentially slower than the tail of W distribution, $\zeta = 0$ and only a logarithmic rate of decline (at best) can be obtained for MSE. When, conversely, $0 \leq c < d$, $\zeta = 1$:

Proof: See Appendix B. ☐

Appendix B also provides the specific form $\zeta; \zeta(L);$ and $\zeta(H)$ take for all cases considered in Proposition 3.

As an example of Proposition 4, if $U; \varepsilon$ and W are jointly normally distributed ($c = d = 2$), σ^* is bounded by $\sigma_L = (\mu_L \ln n)^{1-2}; \mu_L < 2\sigma_{\varepsilon}^2 \sigma_W^2 = (\sigma_{\varepsilon}^2 + 2\sigma_W^2)$, and $\sigma_H = (\mu_H \ln n)^{1-2}; \mu_H > 2\sigma_{\varepsilon}^2 \sigma_W^2 = (\sigma_{\varepsilon}^2 + 2\sigma_W^2)$. The optimal asymptotic MSE has the polynomial component $n^{-\zeta}$ with $\zeta = 1 - \sigma_{\varepsilon}^2 = (\sigma_{\varepsilon}^2 + 2\sigma_W^2)$. We note that ζ can get arbitrarily close to 1 if $\sigma_W^2 \gg \sigma_{\varepsilon}^2$: (In general, if $c = d$, ζ can be made arbitrarily close to 1 given $\gamma \geq \gamma_1$). The $MSE(\sigma_H) \sim \text{avar}(\mathbf{b}_s)$ with $\zeta_H = 1 - \mu_H = (2\sigma_W^2)$, and $MSE(\sigma_L) \sim \text{abias}(\mathbf{b}_s)$ with $\zeta_L = \mu_L = \sigma_{\varepsilon}^2$: Both ζ_L and ζ_H are arbitrarily close to ζ when $\mu_L; \mu_H$ are close to $2\sigma_{\varepsilon}^2 \sigma_W^2 = (\sigma_{\varepsilon}^2 + 2\sigma_W^2)$:

After characterizing the optimal bandwidth parameter, it remains to determine the "optimal" choice of function $s(\cdot)$ (or $1(\cdot)$) in situations where an asymptotic bias is present. Following Proposition 4, if all the parameters of the tail distribution of W and in the function $!(W)$ are known solving the ...rst-order condition for $\circ$ and then substituting into the MSE and minimizing over q (where it appears) as well would give us the "optimal" estimator. When the bounds $\circ_H$, or $\circ_L$, are used instead of $\circ^{\,¤}$ they imply dominance of MSE by asymptotic variance or abias, correspondingly. This in turn implies preference for the function $1(\cdot)$ or function $s(\cdot)$ with large value of q (where it matters) correspondingly.

The use of the bounds $\circ_H$, or $\circ_L$, as the desired bandwidth (bringing MSE close or equal to its best possible rate) might be more practical since the bounds are functions of the leading terms of the tails of W_i and $!(W_i)$ only. For fat tailed distributions, we can estimate the upper tail index of a distribution e.g., using Hill (1975) and Danielsson and De Vries (1997) (see Huisman et al. (1997) for its estimation in small samples). Alternatively, a probability weighted moment estimator (or maximum likelihood) of the parameters from the generalized extreme value distribution can be considered (Hosking et al. (1985)).

5. Conclusions

The paper presents the asymptotic behaviour of the intercept in the sample selection model based on "identi...cation at in...nity," which was ...rst proposed by Heckman (1990). Technical problems in derivations arise from the non-di¤erentiability of the indicator function. This problem was circumvented by AS via introduction of a differentiable function to replace the indicator function. Here we deal with the problem by introducing an assumption on the p.d.f. that essentially permits to obtain the expectation of the "derivative" of the indicator function.

Next, the paper examines the selection of the bandwidth and choice of the estimator of the intercept (Heckman 1990 versus AS 1998) using as a criterion the asymptotic MSE. Two characteristics of the model are of importance for such a choice: the tail behaviour of the selection index, W_i, and the tail behaviour of a function $!(W)$ that determines the asymptotic bias of the estimator. A wide class of distributional assumptions for the model is investigated, speci...cally, $1 ¡ F(W) ¼ c_W W^{®} e^{i\,\circ W^c}$ and $!(W) ¼ c_!\, W^{\&} e^{i\,^1 W^d}$: This class of models includes Pareto, Weibull as well as "combined" tails.

We have shown that for fat-tailed distributions of W_i the choice of the function $s(\cdot)$ or $1(\cdot)$ does not a¤ect the asymptotic variance; the asymptotic bias if una¤ected if additionally $!(W_i)$ does not go to zero too fast. In general, however, the asymptotic MSE may be a¤ected by the choice of function $s(\cdot)$ but then via q and the value of the derivative $s^{(q)}(0)$ only, where q is the order of the ...rst non-zero derivative of $s(\cdot)$ at zero.

If all the parameters determining the tail behaviour of W and the function $!(W)$ were known a solution that would provide an optimal $°^¤ = \text{argmin(MSE)}$ as a function of n, q (the $s(\text{¢})$ function) and all those parameters could be obtained (at least via a numerical algorithm) from the formulae in Proposition 3. Similarly, the "optimal" choice of function $s(\text{¢})$ (or $1(\text{¢})$) can be obtained. Since the solution for the optimal bandwidth (and choice of function $s(\text{¢})$ (or $1(\text{¢})$) may not be practical, we provide simpler bounds on the optimal bandwidth parameter; using a bound may imply preference for either AS or the Heckman estimator.

Asymptotically, we give preference to the Heckman estimator in cases where there is no asymptotic bias and reveal the equivalence of the two estimators under fat-tailed distributions of W_i if additionally $!(W_i)$ does not go to zero too fast. For thinner-tailed distributions the decision is less clear, nevertheless, we argue that the optimal selection of the bandwidth is of primary importance. For ...nite samples, the AS estimator might still have advantages over the Heckman estimator, in that the trade-o¤ between bias and variance, like in nonparametric estimation problems, is better for smooth "kernels". Nevertheless, only observations at the margin are a¤ected by the choice of the function $s(\text{¢})$. In Schafgans (1998b) simulations are presented that reveal these ...ndings clearly.

Appendix A: Asymptotic normality result for the Heckman estimator

To prove asymptotic normality, all we need to show is that (5) holds. We start by deriving a few su¢cient conditions for (5). As in the proof of Theorem A-1 of AS the left hand side of (5) can be written as $C(\frac{\hat{A}}{\hat{B}} \, ¡ \, \frac{A}{B}) = C\frac{\hat{A}¡A}{\hat{B}}\frac{B}{\hat{B}} \, ¡ \, C\frac{\hat{B}¡B}{\hat{B}}\frac{A}{B}\frac{B}{\hat{B}}$, where $C = \frac{\sqrt{P}}{nED_i1(W_i > °_n)} = \frac{3}{4}$, $A = \sum_{i=1}^{n}(Y_i ¡ Z_i^0\mu_0)D_i1(W_i > °_n)$, $\hat{A} = \sum_{i=1}^{n}(Y_i ¡ Z_i^0\hat{b})D_i1(X_i^0\hat{b} > °_n)$, $B = \sum_{i=1}^{n}D_i1(W_i > °_n)$, and $\hat{B} = \sum_{i=1}^{n}D_i1(X_i^0\hat{b} > °_n)$. To show (5), therefore, it su¢ces to show that

(i) $\frac{\hat{B}}{B}=B \; !^p \; 1$

(ii) $C(\hat{A} ¡ A)=B \; !^p \; 0$

(iii) $A=B = O_p(1)$ (A.1)

(iv) $C(\hat{B} ¡ B)=B \; !^p \; 0:$

From Assumption 7 and Lemma A-2 of AS it follows that $C \, ! \, 1$ which means that (iv) implies (i) in (A.1). From Lemmas A-1 and A-2 of AS one gets that $\frac{3}{4}C=B$ equals $n^{¡1=2}\Pr(W_i > °_n)^{¡1=2}$ in probability and thus (ii) is implied by the following su¢cient conditions.

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i + ¹_0)D_i \circledR_{in}}{\Pr(W_i > °_n)^{1=2}} \; !^p \; 0 ; \qquad (A.2)$$

$$\frac{\iota\,(\hat{\beta}_i - \mu_0)^0 \frac{1}{n} \sum_{i=1}^{n} Z_i D_i \hat{\otimes}_{in}}{\Pr(W_i > \circ_n)^{1-2}} \to^p 0\,; \text{ and} \tag{A.3}$$

$$\frac{\iota\,(\hat{\beta}_i - \mu_0)^0 \frac{1}{n} \sum_{i=1}^{n} Z_i D_i 1(W_i > \circ_n)}{\Pr(W_i > \circ_n)^{1-2}} \to^p 0\,; \tag{A.4}$$

where $\hat{\otimes}_{in}$ is given by (6).

Condition (iii) has been shown to hold in the proof of Theorem A-1 in AS, it is equivalent to

$$\frac{\frac{1}{n} \sum_{i=1}^{n} U_i D_i 1(W_i > \circ_n)}{\Pr(W_i > \circ_n)} = O_p(1)\,: \tag{A.5}$$

Finally, Condition (iv) would follow if we show that

$$\frac{\frac{1}{n} \sum_{i=1}^{n} D_i \hat{\otimes}_{in}}{\Pr(W_i > \circ_n)^{1-2}} \to^p 0\,: \tag{A.6}$$

Note that under Assumptions 1, 5 and 6 of AS and using Hölder's inequality the expression in (A.4) is bounded by

$$\frac{O_p(1)E \|Z_i\| 1(W_i > \circ_n)}{\Pr(W_i > \circ_n)^{1-2}} \cdot O_p(1) \left[E \|Z_i\|^3\right]^{1-3} (\Pr(W_i > \circ_n))^{2-3i\,1-2} \to 0\,: \tag{A.7}$$

The proof of Theorem 1, therefore, requires us to show that (A.2), (A.3) and (A.6) which involve the discontinuous function $\hat{\otimes}$ hold.

We do this in three steps: ...rst, a technical lemma is given (Lemma 1), then a lemma is given which examines terms in the expressions of interest (Lemma 2) and ...nally we give the proofs of the theorems which combine the intermediate results.

Lemma 1: Under Assumptions 1, 2, 6, 7 of AS and Assumption A, there exists a su¢ciently slowly increasing $\{M_n\}$ such that $n^{i\,1-2}M_n < 1$ and

$$\sup_{\|\bar{-}_i - \bar{-}_0\| < \frac{M_n}{\sqrt{n}}} \frac{E\left(\frac{1}{\sqrt{n}}\|-_i\|\hat{\otimes}_n(\bar{-};W_i;X_i)\right)}{\Pr(W_i > \circ_n)^{1-2}} \to 0 \text{ as } n \to 1 \text{ for } -_i = Z_i = \frac{-}{\sqrt{n}},\ -_i = U_i,\ \text{or}\ -_i = 1;\ 8i\,:$$

Proof of Lemma 1:

It will be su¢cient to show that for the sequence $\{n^{\frac{1}{3}}\}$

$$\sup_{\|\bar{-}_i - \bar{-}_0\| < \frac{M_n}{\sqrt{n}}} \frac{\frac{1}{\sqrt{n}}E\left[\|-_i\| ¢ j\hat{\otimes}_n(\bar{-};W_i;X_i)j ¢ 1(\|X_i\| > n^{\frac{1}{3}})\right]}{\Pr(W_i > \circ_n)^{1-2}} \to 0\,; \tag{A.8}$$

and

$$\sup_{\|\bar\tau_i - \bar\tau_0\| < \frac{M_n}{\sqrt{n}}} \frac{\frac{p}{n}E\left[\|\delta_i\| \, \|j\varphi_n(\bar\tau; W_i; X_i)j\| \, 1(\|X_i\| \cdot n^{\frac{1}{3}})\right]}{\Pr(W_i > \circ_n)^{1=2}} \;!\; 0: \qquad (A.9)$$

Noting that $\|j\varphi_n(\bar\tau; W_i; X_i)j\| \cdot 1$, we get for (A.8) in the case $\delta_i = U_i$ (using the independence condition of Assumption 2(b) of AS) and similarly for $\delta_i = 1$, under Assumptions 1 and 7 of AS

$$\begin{aligned}
&\cdot \frac{\frac{p}{n}E\left(\|\delta_i\|\,\|j\varphi_n(\bar\tau;W_i;X_i)j\|\,1(\|X_i\|>n^{1=3})\right)}{\Pr(W_i>\circ_n)^{1=2}} \cdot \frac{\frac{p}{n}E\left(\|\delta_i\|\,1(\|X_i\|>n^{1=3})\right)}{\Pr(W_i>\circ_n)^{1=2}} \\
&\cdot \frac{\frac{p}{n}E\|\delta_i\|\,\Pr(\|X_i\|>n^{1=3})}{\Pr(W_i>\circ_n)^{1=2}} \cdot \frac{\frac{p}{n}E\|\delta_i\|\,E\|X_i\|^3=n}{\Pr(W_i>\circ_n)^{1=2}} \\
&= \frac{E\|\delta_i\|\,E\|X_i\|^3}{(n\,\Pr(W_i>\circ_n))^{1=2}} \;!\; 0
\end{aligned} \qquad (A.10)$$

using Jensen's and Markov's inequalities. In the case $\delta_i = Z_i = \frac{p}{\sqrt{n}}$; (A.8) converges to zero even faster, since similarly

$$\frac{E\left(\|Z_i\|\,\|j\varphi_n(\bar\tau;W_i;X_i)j\|\,1(\|X_i\|>n^{1=3})\right)}{\Pr(W_i>\circ_n)^{1=2}} \cdot \frac{\left(E\|Z_i\|^3\right)^{1=3}\left(E\|X_i\|^3\right)^{2=3}}{n^{1=6}\,(n\,\Pr(W_i>\circ_n))^{1=2}} \;!\; 0: \qquad (A.11)$$

The left hand side of (A.9) for $\delta_i = Z_i = \frac{p}{\sqrt{n}}$ is bounded by

$$\frac{\left(E\|Z_i\|^2\right)^{\frac{1}{2}}}{(n\,\Pr(W_i>\circ_n))^{1=4}} @ \left(\sup_{\|\bar\tau_i-\bar\tau_0\|<\frac{M_n}{\sqrt{n}}} \frac{\frac{p}{n}E\,\|j\varphi_n(\bar\tau;W_i;X_i)j\|\,1(\|X_i\|<n^{1=3})}{\Pr(W_i>\circ_n)^{1=2}}\right)^{1=2} \qquad (A.12)$$

using Hölder's inequality. Given Assumptions 1 and 7 of AS, therefore, it remains only to show (A.9) for $\delta_i = 1$ and U_i.

Using the independence of δ_i and X_i for $\delta_i = 1$ and U_i (by Assumption 2(b) of AS) we can rewrite the denominator of (A.9) as

$$E\,(\|\delta_i\|)\, \sup_{\|\bar\tau_i-\bar\tau_0\|<\frac{M_n}{\sqrt{n}}} \frac{p}{n}E\,\|j\varphi_n(\bar\tau;W_i;X_i)j\|\,1(\|X_i\|<n^{1=3}): \qquad (A.13)$$

By examining the function $\varphi_n(\bar\tau;W_i;X_i)$; (6), we can see that $\|j\varphi_n(\bar\tau;W_i;X_i)j\|$ equals 1 if either $\circ_n < W_i < \circ_n + \|jX_i^0(\bar\tau_i - \bar\tau_0)j\|$ for negative $X_i^0(\bar\tau_i - \bar\tau_0)$ or if $\circ_n i \|jX_i^0(\bar\tau_i - \bar\tau_0)j\| < W_i < \circ_n$ for positive $X_i^0(\bar\tau_i - \bar\tau_0)$ and zero otherwise. In view of the restrictions on $\|\bar\tau_i - \bar\tau_0\|$ and $\|X_i\|$ we have $\|jX_i^0(\bar\tau_i - \bar\tau_0)j\| < \frac{M_n}{n^{1=6}}$, which implies $W_i > \circ_n i \frac{M_n}{n^{1=6}}$. Hence,

$$\begin{aligned}
& \sup_{\|\bar\tau_i-\bar\tau_0\|<\frac{M_n}{\sqrt{n}}} \frac{p}{n}E\,\|j\varphi_n(\bar\tau;W_i;X_i)j\|\,1(\|X_i\|<n^{1=3}) \\
&= \sup_{\|\bar\tau_i-\bar\tau_0\|<\frac{M_n}{\sqrt{n}}} \frac{p}{n}E\,\|j\varphi_n(\bar\tau;W_i;X_i)j\|\,1(W_i>\circ_n i \tfrac{M_n}{n^{1=6}})\,1(\|X_i\|<n^{1=3}):
\end{aligned} \qquad (A.14)$$

Consider the linear transformation $L : X \to \begin{pmatrix} X_i \\ w' \end{pmatrix}$, where $\iota_i = X_{(i-1)}$ is the vector of all components of X except the first (NB the matrix of this transformation is $L = \begin{pmatrix} \mu & -\bar\delta_{01} & \cdots & -\bar\delta_{0k} \\ 0 & & I_{k-1} \end{pmatrix}$; which is non-singular). Then $X = L^{-1}\begin{pmatrix} X_i \\ w' \end{pmatrix}$ and $X'(\bar\delta_i - \bar\delta_0) =$ $W\,(L^{-1}(\bar\delta_i - \bar\delta_0))_{(1)} + \iota'(L^{-1}(\bar\delta_i - \bar\delta_0))_{(i-1)}$, where subscript (1) denotes the first component of a vector. We denote $(L^{-1}(\bar\delta_i - \bar\delta_0))_{(1)} = b$[7] and $(L^{-1}(\bar\delta_i - \bar\delta_0))_{(i-1)} = B$. Re-examining, for negative $X_i'(\bar\delta_i - \bar\delta_0)$; the condition $\alpha_n < W_i < \alpha_n + jX_i'(\bar\delta_i - \bar\delta_0)j = \alpha_n - W\,(L^{-1}(\bar\delta_i - \bar\delta_0))_{(1)} - \iota'(L^{-1}(\bar\delta_i - \bar\delta_0))_{(i-1)}$ we realize that for negative $X_i'(\bar\delta_i - \bar\delta_0)$ $j\Phi_n(\bar\delta; W_i; X_i)j$ equals 1 for $\alpha_n < W_i < \frac{\alpha_n - \iota'B}{1+b}$. Similarly for positive $X_i'(\bar\delta_i - \bar\delta_0)$ $j\Phi_n(\bar\delta; W_i; X_i)j$ equals 1 for $\frac{\alpha_n - \iota'B}{1+b} < W_i < \alpha_n$. Thus (A.14) equals

$$\sup_{k\bar\delta_i - \bar\delta_0 k < \frac{M_n}{\sqrt{n}}} \frac{p}{\sqrt{n}} E \left[\int_{\alpha_n}^{\frac{\alpha_n - \iota'B}{1+b}} 1(X_i'(\bar\delta_i - \bar\delta_0) < 0)1(kX_ik < n^{1-3})p.d.f._{W;i,\iota_i}(W; \iota_i)dW + \right.$$
$$\left. \int_{\frac{\alpha_n - \iota'B}{1+b}}^{\alpha_n} 1(X_i'(\bar\delta_i - \bar\delta_0) > 0)1(W_i > \alpha_n - \frac{M_n}{n^{1-6}})1(kX_ik < n^{1-3})p.d.f._{W;i,\iota_i}(W; \iota_i)dW \right]:$$

(A.15)

The p.d.f.$_{W;i,\iota_i}$ exists for large enough n, since the expression under the integrals is non-zero only if $W_i > \alpha_n - \frac{M_n}{n^{1-6}}$ and thus as $\alpha_n \to 1$ for $M_n = O(n^{1-6})$ becomes greater than A; the conditional p.d.f.$_{W;i,\iota_i}$ declines monotonically in W: Denote $\alpha_n - \frac{M_n}{n^{1-6}}$ by $\mathring\alpha_n$: The first integral on the right hand side of (A.15) is bounded by $p.d.f._{W;i,\iota_i}(\mathring\alpha_n; \iota_i)\left(\frac{\alpha_n - \iota'B}{1+b} - \alpha_n \right)$, the second can be bounded by $p.d.f._{W;i,\iota_i}(\mathring\alpha_n; \iota_i)\left(\alpha_n - \frac{\alpha_n - \iota'B}{1+b} \right)$: Thus the sum is bounded by

$$\sup_{k\bar\delta_i - \bar\delta_0 k < \frac{M_n}{\sqrt{n}}} 2\frac{p}{\sqrt{n}} \int p.d.f._{W;i,\iota_i}(\mathring\alpha_n; \iota_i)\left[-\frac{\iota'B}{1+b} \right] dPr_\iota$$

(A.16)

$$\cdot \frac{2M_n}{1+b} \int k\iota_i k\, p.d.f._W(\mathring\alpha_n)dPr_\iota jW=\mathring\alpha_n = \frac{2M_n}{1+b}p.d.f._W(\mathring\alpha_n)E_{jW=\mathring\alpha_n}k\iota_i k;$$

where $\iota_i = X_{(i-1)}$ and $kBk = \mathring{(L^{-1}(\bar\delta_i - \bar\delta_0))_{(i-1)}} \leq 8\mathring{(\bar\delta_i - \bar\delta_0)_{(i-1)}} \cdot \frac{M_n}{n^{1-2}}$ using the restriction on $k\bar\delta_i - \bar\delta_0 k$: The notation dPr_ι indicates that a Stiltjes integral with respect to the cumulative probability function ι is taken, if a marginal density exists

---

[7]Note: $jbj \cdot \mathring L^{-1} \langle k\bar\delta_i - \bar\delta_0 k \cdot \mathring L^{-1} \frac{M_n}{\sqrt{n}}$. Assume that M_n is such that $\mathring L^{-1} \frac{M_n}{\sqrt{n}} \cdot G < 1$.

[8]Here we use the fact that $L^{-1} = \begin{pmatrix} \frac{1}{\bar\delta_{01}} & \frac{\bar\delta_{02}}{\bar\delta_{01}} \cdots \frac{\bar\delta_{0k}}{\bar\delta_{01}} \\ 0 & I_{k-1} \end{pmatrix}$.

$dPr_i = p:d:f:_i (_i)d_i$. By Jensen's inequality, and in view of Assumption A(b) this is bounded almost surely by

$$\frac{2M_n}{1+b} \, ¢ \, p:d:f:_W (°_n) \, ¢ \, E_{jW=°_n} \left[°X_{(i\ 1)}^{°3}\right]^{1=3} : \qquad (A.17)$$

Let us consider $E_{jW=°_n} \left[°X_{(i\ 1)}^{°3}\right]$. By Assumption 1 of AS, we know that the unconditional expectation $E\left[°X_{(i\ 1)}^{°3}\right]$ is bounded; this implies that for the ...xed A of Assumption A $E\left[°X_{(i\ 1)}^{°3}1(W > A)\right]$ is bounded as well.

$$E\left[°X_{(i\ 1)}^{°3}1(W > A)\right] = \iint °X_{(i\ 1)}^{°3} 1(W > A)p:d:f:_{W;i\ jW} (_i) p:d:f:_W (W) d_i \, dW$$

$$= \int J(W)p:d:f:_W (W)1(W > A)dW; \qquad (A.18)$$

where $J(W) = \int °X_{(i\ 1)}^{°3} p:d:f:_{W;i\ jW}(_i)d_i$ is a function of W only. Since the integral $\int J(W)p:d:f:_W (W)dW$ ´ $E\left[°X_{(i\ 1)}^{°3}\right]$ exists, it implies that as $W ! 1$ $J(W)p:d:f:_W (W) = o(W^{i\ 1})$, or

$$J(W) = o\left[W^{i\ 1} (p:d:f:_W (W))^{i\ 1}\right]^{¢} : \qquad (A.19)$$

Equation (A.17), can then be rewritten as follows

$$\frac{2M_n}{1+b} \, ¢ \, p:d:f:_W (ê_n) \, ¢ \, J(ê_n)^{1=3} = \frac{2M_n}{1+b} \, ¢ \, p:d:f:_W \left(°_n \, i \, \frac{M_n}{n^{1=6}}\right)^{2=3} o(°_n^{i\ 1=3}): \qquad (A.20)$$

This implies that for $-_i = 1$ and U_i (A.9) can be bounded as

$$\frac{E\,(k-_i k)\frac{2M_n}{1+b} ¢ p:d:f:_W \left(°_n \, i \, \frac{M_n}{n^{1=6}}\right)^{2=3} o(°_n^{i\ 1=3})}{Pr(W_i > °_n)^{1=2}}$$

$$\tilde{A} \quad E\,(k-_i k)\frac{2M_n}{1+b} \left[\frac{p:d:f:_W \left(°_n \, i \, \frac{M_n}{n^{1=6}}\right)}{Pr(W_i > °_n)^{3=4}}\right]^{2=3} o(°_n^{i\ 1=3}): \qquad (A.21)$$

Set $M_n = \min f d n^{1=6}; °^{1=3}g$ with $d > 0$ as in Assumption A. Using Assumptions 1 and 6 of AS and Assumption A(a) the right-hand side of (A.21) converges to zero for M_n.
2

The next lemma will help to show that terms, involving $®_{in}$; which appear in (A.2), (A.3), and (A.6) have a zero probability limit.

Lemma 2: Under Assumptions 1, 2, 5–7 of AS, and Assumption A

$$\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}} \to^p 0 \text{ as } n \to \infty \text{ for } -_i = Z_i = \sqrt{n}, \ -_i = U_i, \text{ or } -_i = 1; \ 8i:$$

Proof of Lemma 2:

We would like to show that for any $3 > 0$

$$\Pr\left(\left|\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}}\right| > 3\right) \to 0 \text{ as } n \to \infty: \tag{A.22}$$

From Lemma 1, let M_n satisfy $M_n o(\circ_n^{1=3}) \to 0$ and $\frac{M_n}{n^{1=6}} \cdot d$. The left hand side of (A.22) is equal to

$$\Pr\left(\left|\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}}\right| > 3; \left\|b_i - {}_0\right\| \cdot \frac{M_n}{\sqrt{n}}\right) + \Pr\left(\left|\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}}\right| > 3; \left\|b_i - {}_0\right\| > \frac{M_n}{\sqrt{n}}\right)$$

$$\cdot \Pr\left(\left|\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}}\right| > 3; \left\|b_i - {}_0\right\| \cdot \frac{M_n}{\sqrt{n}}\right) + \Pr\left(\left\|b_i - {}_0\right\| > \frac{M_n}{\sqrt{n}}\right): \tag{A.23}$$

The second expression on the right hand side of (A.23) converges to zero by Assumption 5 of AS, since $\sqrt{n}(b_i - {}_0) = O_p(1)$ implies $\sqrt{n}(b_i - {}_0)=M_n = o_p(1)$. The ...rst expression on the right hand side of (A.23) is bounded by

$$\sup_{\|k^-_i - {}_0\| \cdot \frac{M_n}{\sqrt{n}}} \Pr\left(\left|\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_n(^-;W_i;X_i) j}{\Pr(W_i > \circ_n)^{1=2}}\right| > 3\right) \tag{A.24}$$

$$\cdot \sup_{\|k^-_i - {}_0\| \cdot \frac{M_n}{\sqrt{n}}} \frac{E\left[\left|\frac{\frac{1}{n}\sum k_{-i} k_j \otimes_n(^-;W_i;X_i) j}{\Pr(W_i > \circ_n)^{1=2}}\right|\right]}{3} = \sup_{\|k^-_i - {}_0\| \cdot \frac{M_n}{\sqrt{n}}} \frac{\sqrt{n} E\left[k_{-i} k_j \otimes_n(^-;W_i;X_i) j\right]}{3 \Pr(W_i > \circ_n)^{1=2}};$$

where the inequality is based on Markov's inequality. This term converges to zero for all $3 > 0$ by Lemma 1. $\quad\blacksquare$

Proof of Theorem 1: For our proof it is su¢cient to show (A.2), (A.3), and (A.6). By Assumptions 1 and 5 for b_i μ_0 of AS, the left-hand sides of (A.2), (A.3) can be bounded by:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} jU_i j \xi j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}} + j^1 {}_0 j \frac{\frac{1}{n}\sum_{i=1}^{n} j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}}; \tag{A.25}$$

$$\frac{O_p(1)\frac{1}{n}\sum_{i=1}^{n} kZ_i k \xi j \otimes_{in} j}{\Pr(W_i > \circ_n)^{1=2}}; \tag{A.26}$$

respectively. From Lemma 2, (A.25) and (A.26) have zero probability limits.

Finally, to prove (A.6), we note that its left-hand side can be bounded by

$$\frac{\frac{1}{n}\mathbf{P}_{i=1}^{n}\mathbf{j}\circledast_{i n}\mathbf{j}}{\Pr(W_i > \,^\circ_n)^{1=2}}; \tag{A.27}$$

which by Lemma 2 again converges in probability to zero. This completes the proof of our theorem. 2

Proof of Theorem 2: By Cramer-Wold device, this result follows directly from (5) and Theorem A-4 in AS. In the latter, the result in Theorem 2 is shown for the case where $\mathbf{b}_l$ is replaced by $\mathbf{b}_{l;0}$. 2

Appendix B: Asymptotic variance and bias: Selection of Bandwidth and Estimator

In this Appendix we provide the asymptotic bias, variance, and mean squared error of the AS and Heckman estimator. Using Lemma A–2 of AS and Theorem 1 of this paper, we write the asymptotic bias and variance as

$$\text{abias}(\mathbf{b}_s) \; \text{\textonequarter} \; _i \; \frac{E \, ! \, (W_i) s(W_i \, _i \, ^\circ_n)}{E s(W_i \, _i \, ^\circ_n)}; \quad \text{abias}(\mathbf{b}_l) \; \text{\textonequarter} \; _i \; \frac{E \, ! \, (W_i) 1(W_i > \,^\circ_n)}{E1(W_i > \,^\circ_n)}; \tag{B.1}$$

$$\text{avar}(\mathbf{b}_s) \; \text{\textonequarter} \; \text{\textthreequarters}^2 n^{i \, 1} \frac{E s^2(W_i \, _i \, ^\circ_n)}{[E s(W_i \, _i \, ^\circ_n)]^2}; \quad \text{avar}(\mathbf{b}_l) \; \text{\textonequarter} \; \text{\textthreequarters}^2 n^{i \, 1} \, (E1(W_i > \,^\circ_n))^{i \, 1} \tag{B.2}$$

In the following we let $! \, (W_i) \, _{\text{\textcurrency}} \, 0 \, 8 W_i > \,^\circ_n$ (similar proofs can be given when $! \, (W_i) \cdot \, 0 \, 8 W_i > \,^\circ_n$).

Proof of Proposition 1: For (a), we note that by Cauchy-Schwartz inequality

$$E s(W_i \, _i \, ^\circ_n) = E s(W_i \, _i \, ^\circ_n) 1(W_i > \,^\circ_n) \cdot \, [E s^2(W_i \, _i \, ^\circ_n)]^{1=2} [E1(W_i > \,^\circ_n)]^{1=2}: \tag{B.3}$$

This inequality combined with (B.2) gives the result that (a) holds for any sequence $\,^\circ_n$.

Next we turn to (b). Using (B.1), we get

$$\text{jabias}(\mathbf{b}_s)\mathbf{j} \; \text{\textonequarter} \; \frac{E \, ! \, (W_i) s(W_i \, _i \, ^\circ_n)}{E s(W_i \, _i \, ^\circ_n)} \cdot \frac{E \, ! \, (W_i) 1(W_i > \,^\circ_n)}{E1(W_i > \,^\circ_n + b)}$$
$$\text{\textonequarter} \; \text{jabias}(\mathbf{b}_l)\mathbf{j} \frac{E1(W_i > \,^\circ_n)}{E1(W_i > \,^\circ_n + b)}: \tag{B.4}$$

Since $E1(W_i > \,^\circ_n) = E1(W_i > \,^\circ_n + b) \; \text{\textonequarter} \; 1$ (b) follows. 2

Proof of Proposition 2: From Proposition 1(a), we know

$$\frac{E s^2(W_i \, _i \, °_n)}{[E s(W_i \, _i \, °_n)]^2} \, _, \, [E1(W_i > °_n)]^{i\,1}: \qquad (B.5)$$

Furthermore as $E s^2(W_i \, _i \, °_n) \cdot E s(W_i \, _i \, °_n)$ and $E s(W_i \, _i \, °_n) \, _, \, E1(W_i > °_n + b)$;

$$\frac{E s^2(W_i \, _i \, °_n)}{[E s(W_i \, _i \, °_n)]^2} \cdot [E1(W_i > °_n + b)]^{i\,1} \, ¼ \, [E1(W_i > °_n)]^{i\,1}; \qquad (B.6)$$

and (a) follows.

Under the same assumptions, we know from Proposition 1(b) that

$$\frac{E! \, (W_i) s(W_i \, _i \, °_n)}{E s(W_i \, _i \, °_n)} \cdot \frac{E! \, (W_i)1(W_i > °_n)}{E1(W_i > °_n + b)} \, ¼ \, \frac{E! \, (W_i)1(W_i > °_n)}{E1(W_i > °_n)}: \qquad (B.7)$$

In addition, we have

$$\frac{E! \, (W_i) s(W_i \, _i \, °_n)}{E s(W_i \, _i \, °_n)} \, _, \, \frac{E! \, (W_i)1(W_i > °_n + b)}{E1(W_i > °_n)} \, ¼ \, \frac{E! \, (W_i)1(W_i > °_n)}{E1(W_i > °_n)}: \qquad (B.8)$$

Combining these inequalities we obtain (b).   2

For the remainder of this appendix we have omitted the subscript n on ° to simplify notation. For the proof of Proposition 3, we make use of the following technical lemma:

Lemma 3: As ° ! 1,

(a) $\int_°^1 \, _,c(W \, _i \, °)^i W^v \exp(_i \, _, W^c)dW = °^{i+v_i\,c+1_i\,ci}\exp(_i \, _, °^c)(_,c)^{i\,i}i!(1 + o(°^{i\,c}))$

(b) For a>f, $\int_°^1 \, ^1a(W \, _i \, °)^i W^v \exp(_i \, ^1W^a \, _i \, _, W^f)dW =$

$$°^{i+v_i\,a+1_i\,ai}\exp(_i \, ^1°^a \, _i \, _, °^f)(^1a)^{i\,i}i!(1 + o(°^{i\,a})):$$

Proof: The integral in (a), I, can be rewritten as

$$I \quad = \quad \int_°^Z \, _,c \sum_{j=0}^{x} \binom{\mu_i}{j} (_i \, 1)^{i+j\,°i_i\,j} W^{j+v} \exp(_i \, _, W^c)dW$$

$$= \quad \sum_{j=0}^{x} \binom{\mu_i}{j} (_i \, 1)^{i+j\,°i_i\,j} \int_°^Z \, _,c W^{j+v} \exp(_i \, _, W^c)dW: \qquad (B.9)$$

By setting $W^c = z$, we get

$$I = \sum_{j=0}^{\mu_i \P} \binom{\mu_i}{j} (\i 1)^{i+j\circ i_i j} \Cent \int_{\circ c}^{Z_1} z^{\frac{j+v_i c+1}{c}} \exp(\i z) dz: \qquad (B.10)$$

Combining 3.381#3 and 8.357 in Gradshteyn and Ryzhik (1994)[9] we obtain

$$
\begin{aligned}
I &= \sum_{j=0}^{\mu_i \P} \binom{\mu_i}{j} (\i 1)^{i+j\circ i_i j} \i \frac{j+v+1}{c} \i \left(\frac{j+v+1}{c}; \circ c\right) \\
&= \sum_{j=0}^{\mu_i \P} \binom{\mu_i}{j} (\i 1)^{i+j\circ i_i j} \Cent^{\circ j+v_i c+1} \exp(\i \circ c) \Cent \\
&\qquad \left(\sum_{m=0}^{\tilde{A}_1} \frac{(\i 1)^m \i (1 \i \frac{j+v+1}{c} + m)}{\circ c m_m \i (1 \i \frac{j+v+1}{c})} + o(\circ i cL)\right) \qquad (B.11) \\
&= (\i 1)^{i\circ i+v_i c+1} \exp(\i \circ c) \sum_{j=0}^{\mu_i \P} \binom{\mu_i}{j} (\i 1)^j \left(\sum_{m=0}^{\tilde{A}_1} \frac{\sum_{s=0}^{P_m} \cdot_s (\frac{j}{c})^s}{\circ c m_m} + o(\circ i cL)\right);
\end{aligned}
$$

where we have substituted $(\i 1)^m \i (1 \i \frac{j+v+1}{c} + m) = \i (1 \i \frac{j+v+1}{c}) = \sum_{s=0}^{P_m} \cdot_s (\frac{j}{c})^s$, with $\cdot_s = 1$ for $s = m$ and some known constant for $s = 0; ::::; m \i 1$. Using the fact that $\sum_{j=0}^{P \i \Cent} \binom{\i}{j} (\i 1)^j j^m = 0$ for $\i \geq m+1 \geq 1; 0^0 \approx 1$ (see 0.154#3 in Gradshteyn and Ryzhik (1994)), we get

$$
\begin{aligned}
I &= (\i 1)^{i\circ i+v_i c+1} \exp(\i \circ c) \sum_{j=0}^{\mu_i \P} \binom{\mu_i}{j} (\i 1)^j \left(\frac{(\frac{j}{c})^i}{\circ c i_i} (1 + o(\circ i c))\right) \\
&= \circ i+v_i c+1_i ci \exp(\i \circ c)(\cdot c)^{i \i} i!(1 + o(\circ i c)); \qquad (B.12)
\end{aligned}
$$

where the second equality uses 0.154#4 in Gradshteyn and Ryzhik (1994).

   Next consider the integral (b), I^0, for $a > f$. We notice that I^0 can be rewritten as follows,

$$
\begin{aligned}
I^0 &= \exp(\i \circ f) \int_{\circ}^{Z_1} 1a(W \i \circ)^i W^v \exp(\i 1 W^a) \left[\exp(\i (W^f \i \circ f))\right]^{\Cent} dW \quad (B.13) \\
&= \exp(\i \circ f) \int_{\circ}^{Z_1} 1a(W \i \circ)^i W^v \exp(\i 1 W^a) \left(\sum_{r=0}^{\tilde{A}_x} \frac{(\i 1)^r \i r (W^f \i \circ f)^r}{r!}\right) dW;
\end{aligned}
$$

---

[9]There is a typographical error in 3.381#3 in G&R; the correct formula reads $\int_u^{R_1} x^{v_i 1} e^{\i 1x} dx = \i^v \i (v; \i 1u):$

where we have substituted the series representation for the last exponential function in the integral. By dominating convergence theorem, we can interchange the integral and summation, giving

$$I^0 = \exp(_{\!\!\scriptscriptstyle\ldots} {}^\circ f) \sum_{r=0}^{\infty} \frac{(_{\!\!\scriptscriptstyle i} 1)^r {}_{\!\!\scriptscriptstyle\ldots}{}^r}{r!} \int_{\circ}^{1} {}^1 a(W^f {}_{\!\!\scriptscriptstyle i} {}^\circ f)^r (W {}_{\!\!\scriptscriptstyle i} {}^\circ)^i W^v \exp(_{\!\!\scriptscriptstyle i} {}^1 W^a) dW: \quad (B.14)$$

To complete the proof, we need to reapply the steps taken in (B.9)–(B.12). The ...nal result is obtained by setting $r = 0$ (all remaining terms converge to zero faster). A detailed proof of this can be obtained from the authors. $\quad 2$

Proof of Proposition 3: In this proof we will not attempt to formally show all cases considered, but indicate the method used to derive the results, pointing primarily to the more complex derivations.

Using the results from Proposition 2, $\mathrm{avar}(^1{}_{\!\!\scriptscriptstyle i}) \frac{1}{4} \mathrm{avar}(^1{}_{\!\!\scriptscriptstyle s})$ if $1 {}_{\!\!\scriptscriptstyle i} F(W) \frac{1}{4} W^{\circledR}e^{i {}_{\!\!\scriptscriptstyle\ldots}W^c}$ with $c \cdot 1$, since Assumption 4 of AS holds with $\gg = 0$ and $E1(W > {}^\circ)=E1(W > {}^\circ + b) \frac{1}{4} 1$. The variance in these cases can be easily derived using $\frac{3}{4}^2=(nE1(W_i > {}^\circ))$; see (B.2).[10]

When on the other hand $1 {}_{\!\!\scriptscriptstyle i} F(W) \frac{1}{4} W^{\circledR}e^{i {}_{\!\!\scriptscriptstyle\ldots}W^c}$ with $c > 1$ we need to derive $\mathrm{avar}(^1{}_{\!\!\scriptscriptstyle s})$ using the de...nition in (B.2) (Assumption 4 of AS does not hold with $\gg = 0$). For this we need to apply Lemma 3(a). Notice that for any $^2 > 0$ we can write $Es(W_i {}_{\!\!\scriptscriptstyle i} {}^\circ)$ as

$$
\begin{aligned}
Es(W_i {}_{\!\!\scriptscriptstyle i} {}^\circ) = {} & \frac{a_q}{q!} \int_{\circ}^{1} (W {}_{\!\!\scriptscriptstyle i} {}^\circ)^q {}_{\!\!\scriptscriptstyle\ldots} cW^{\circledR+c_i 1} \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}W^c) dW \\
& {}_{\!\!\scriptscriptstyle i} \frac{a_q}{q!} \int_{\circ+2}^{1} (W {}_{\!\!\scriptscriptstyle i} {}^\circ)^q {}_{\!\!\scriptscriptstyle\ldots} cW^{\circledR+c_i 1} \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}W^c) dW \quad (B.15) \\
& + \int_{\circ}^{\circ+2} \Big[s(W {}_{\!\!\scriptscriptstyle i} {}^\circ) {}_{\!\!\scriptscriptstyle i} \frac{a_q}{q!}(W {}_{\!\!\scriptscriptstyle i} {}^\circ)^q \Big]_{\!\!\scriptscriptstyle\ldots} cW^{\circledR+c_i 1} \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}W^c) dW \\
& + \int_{\circ+2}^{1} s(W {}_{\!\!\scriptscriptstyle i} {}^\circ) {}_{\!\!\scriptscriptstyle\ldots} cW^{\circledR+c_i 1} \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}W^c) dW:
\end{aligned}
$$

By Lemma 3(a), the ...rst integral is $\frac{1}{4} a_q {}^\circ q+{}^{\circledR}i^{cq}(_{\!\!\scriptscriptstyle\ldots}c)^i q \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}{}^\circ c)$; the second is similarly $\frac{1}{4} {}_{\!\!\scriptscriptstyle i} a_q({}^\circ + {}^2)^{q+{}^{\circledR}i^{cq}}(_{\!\!\scriptscriptstyle\ldots}c)^i q \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}({}^\circ + {}^2)^c)$ which goes to zero at an expo- nentially faster rate than the ...rst as long as $^2{}^\circ c_i 1 ! 1$ since $\exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}({}^\circ + {}^2)^c) = \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}{}^\circ c) \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}c^2{}^\circ c_i 1(1 + o(\frac{2}{\circ})))$. Using Assumption S and Lemma 3(a), the ab- solute value of the third integral can be bounded by $O(1)^{2q+1}{}^\circ{}^{\circledR} \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}{}^\circ c)(1 + o({}^\circ i {}^c))$

---

[10] For $W > {}^\circ$ the p:d:f:(W) equals $_{\!\!\scriptscriptstyle\ldots} cW^{\circledR+c_i 1} \exp(_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}W^c)(1 + o({}^\circ i {}^c))$, where $^{\circledR} = {}_{\!\!\scriptscriptstyle i} {}_{\!\!\scriptscriptstyle\ldots}$ if $c = 0$ (Pareto tail case).

(we apply a $(q + 1)^{th}$ order Taylor expansion to $s(\mathcal{C})$ around zero). If $2q+1 \circ (c_i 1)q \; ! \; 0$ this implies that the third integral goes to zero faster than the ...rst. Finally, the fourth integral can be bounded by one where the function $1(\mathcal{C})$ is substituted for $s(\mathcal{C})$, thus by Lemma 3(a) it is bounded by $(° + 2)^® \exp(_i \, (° + 2)^c)(1 + o(° i \, c))$ and (similarly to the second integral) goes to zero exponentially faster than the ...rst one if $2 \circ c_i 1 \; ! \; 1$. If the conditions are met, the ...rst integral dominates $Es(W_i \; _i \; °)$, i.e.,

$$Es(W_i \; _i \; °) \; ¼ \; a_q \circ q + ® i \, cq (_,c)^{i \, q} \exp(_i \, _,°^c): \qquad (B.16)$$

For $Es^2(W \; _i \; °)$, we get similarly four terms, the ...rst of which $¼ \; a_q^{2 i 2q} \overset{\mathcal{C}}{}_q \circ 2q(1_i c) + ® \mathcal{C} (_,c)^{i \, 2q} \exp(_i \, _,°^c)$; the second is the corresponding integral from $° + 2$ to in...nity and requires $"° c_i 1 \; ! \; 1$ to go to zero exponentially faster. The third one can analogously be bounded by

$$Z_{° + "} \; \left[\; ^-s^2(W \; _i \; °) \; _i \; \frac{a_q}{q!} ^3 \right.2 (W \; _i \; °)^{2q} ^- p:d:f:_W(W)dW \right.$$

$$\cdot \quad O(1)^{22q+1 \circ ®} \exp(_i \, _,°^c)(1 + o(° i \, c))$$

and thus needs $22q+1 \circ (c_i 1)2q \; ! \; 0$: And ...nally, the fourth integral can be bounded again by replacing $s^2(\mathcal{C})$ by $1(\mathcal{C})$, which using Lemma 3(a) goes to zero exponentially faster than the ...rst one if $2 \circ c_i 1 \; ! \; 1$. For $2 = ° i \frac{(c_i 1)(2q+\{)}{2q+1}$ with $0 < \{ < 1$ the ...rst integral dominates $Es^2(W_i \; _i \; °)$, i.e.,

$$Es^2(W_i \; _i \; °) \; ¼ \; a_q^{2 i 2q} \overset{\mathcal{C}}{}_q \circ 2q + ® i \, 2cq (_,c)^{i \, 2q} \exp(_i \, _,°^c): \qquad (B.17)$$

This 2 also satis...es the requirement for (B.16) to hold. Combining (B.16) and (B.17) give $avar(^1_s)$:

Concerning the results presented in (b), if $1 _i F(W) \; ¼ \; W^® e^{i \, _,W^c}$ with $c \cdot 1$ and $!(W) \; ¼ \; W^\& \exp(_i \, ^1 W^d)$ with $d \cdot 1$ all assumptions in Proposition 2 are satis...ed and $abias(^1_l) \; ¼ \; abias(^1_s)$. The derivations of the asymptotic bias in that case requires us to compute $_i E(!(W_i)1(W_i > °))$ (see (B.1)). Substituting $1 _i F(W)$ and $!(W)$

$$E(!(W_i)1(W_i > °)) = _,c \int_°^{Z_1} W^{\& + ® + c_i 1} \exp(_i \, ^1 W^d \; _i \; _,W^c)(1 + o(° i \, c)dW: \qquad (B.18)$$

When $d = c$, this expectation can be obtained straightforwardly using the analysis of the asymptotic variance given above. When $d \; ¢ \; c$, Lemma 3(b) gives

$$E(!(W_i)1(W_i > °)) \; ¼ \; \begin{cases} \frac{_,c}{^1 d} ° \& + ® + c_i d \exp(_i \, ^1 \circ d \; _i \; _,°^c)(1 + o(° i \, c)) & d > c \\ °^{\& + ®} \exp(_i \, ^1 \circ d \; _i \; _,°^c)(1 + o(° i \, c)) & d < c: \end{cases} \qquad (B.19)$$

When c or d (or both) exceed 1, the asymptotic bias of the Heckman and AS estimator are not equal any more. In that case we need to extend the analysis above

to derive $_i E(! (W_i)s(W_i _i °))$. As in (B.15), we write $E(! (W_i)s(W_i _i °))$ as a sum of four integrals, where $^2 > 0$

$$E(!_Z (W_i)s(W_i _i °))$$

$$= \frac{a_q}{q!} \int_{\circ}^{Z_1} (W _i °)^q {}_\cdot cW^{\&+®+c_i 1} \exp(_i {}^1W^d {}_i {}_\cdot W^c)dW \qquad \text{(B.20)}$$

$$_i \frac{a_q}{q!} \int_{\circ+2}^{Z_1} (W _i °)^q {}_\cdot cW^{\&+®+c_i 1} \exp(_i {}^1W^d {}_i {}_\cdot W^c)dW$$

$$+ \int_{\circ+2}^{Z_{\circ+2}} \mathbf{h} s(W _i °) _i \frac{a_q}{q!}(W _i °)^q \mathbf{i} {}_\cdot cW^{\&+®+c_i 1} \exp(_i {}^1W^d {}_i {}_\cdot W^c)dW$$

$$+ \int_{\circ+2}^{Z^\circ_1} s(W _i °) {}_\cdot cW^{\&+®+c_i 1} \exp(_i {}^1W^d {}_i {}_\cdot W^c)dW:$$

In the case where $c = d$ the result follows directly from Lemma 3(a). Using a similar discussion when $c \neq d$, Lemma 3(b) gives us

$$\frac{E}{\frac{1}{2}}(! (W_i)s(W_i _i °))$$

$$\frac{1}{4} \quad \begin{cases} a_q \frac{c}{(^1d)^{q+1}} \exp(_i {}_\cdot °^c {}_i {}^{1\circ d})° q+\&+®+c_i d_i dq & d > c \\ a_q(_\cdot c)^{i\ q} \exp(_i {}_\cdot °^c {}_i {}^{1\circ d})° q+\&+®_i cq & d < c: \end{cases} \qquad \text{(B.21)}$$

This completes the derivations required for the proof. $\quad 2$

Proof of Proposition 4: According to Proposition 3, the asymptotic MSE for all cases to be considered has the form

$$MSE = an^{i\ 1\circ i\ ®} \exp(_\cdot °^c) + b°^{2´} \exp(_i 2^{1\circ d}); \qquad \text{(B.22)}$$

where

$$´ = \begin{cases} \& + (c _i d)(1 + q) & \text{if } d > c; d > 1 \\ \& + (c _i d) & \text{if } d > c; d \cdot 1 \\ \& & \text{otherwise.} \end{cases} \qquad \text{(B.23)}$$

Thus

$$\frac{@MSE}{@°} = an^{i\ 1\circ i\ ®i\ 1} \exp(_\cdot °^c)(_i ® + _\cdot c°^c) + b°^{2´i\ 1} \exp(_i 2^{1\circ d})(2´ _i 2^1 d°^d): \qquad \text{(B.24)}$$

The optimal bandwidth $°^¤$ solves the ...rst order condition in which we ignore terms that go to zero faster than the ones we keep, i.e., $°^¤$ solves:

(i) $_i ® a \exp(_\cdot)n^{i\ 1\circ ¤i\ ®i\ 1} + 2´ b \exp(_i 2^1)°^{¤2´i\ 1} = 0$ if $c = 0; d = 0 (®; ´ < 0)$

(ii) $_\cdot c a n^{i\ 1\circ i\ ®+c_i 1} \exp(_\cdot °^c) + 2´ b \exp(_i 2^1)°^{2´i\ 1} = 0$ if $c > 0; d = 0 (´ < 0))$

(iii) $_i ® a \exp(_\cdot)n^{i\ 1\circ i\ ®i\ 1} _i 2^1 d b°^{2´+d_i 1} \exp(_i 2^{1\circ d}) = 0$ if $c = 0; d > 0 (® < 0)$

(iv) $_\cdot c a n^{i\ 1\circ i\ ®+c_i 1} \exp(_\cdot °^c) _i 2^1 d b°^{2´+d_i 1} \exp(_i 2^{1\circ d}) = 0$ if $c > 0; d > 0:$

We discuss each case separately.

Case (i). In this case we get an analytic solution $\beta^* = (\frac{2\hat{b}\exp(-2\gamma)}{\alpha a\exp(\delta)}n)^{\frac{1}{\gamma-\alpha+2\delta}}$ and

$MSE(\beta^*) = a\exp(\delta)(\frac{2\hat{b}\exp(-2\gamma)}{\alpha a\exp(\delta)})^{\frac{\alpha}{\gamma+2\delta}}n^{\frac{2\gamma}{\alpha+2\gamma}}(1 + \frac{b\exp(-2\gamma)}{a\exp(\delta)})$, so $\iota = \frac{2\gamma}{\alpha+2\gamma}$: Here for

any $\beta = \mu n^{\frac{1}{\gamma-\alpha+2\delta}}$, the corresponding rate equals ι.

Case (ii). Substituting from the first order conditions we can express $MSE(\beta^*)$ as
$\frac{1}{4}b\beta^{-2\gamma}\exp(-2\gamma)(1 + \frac{2\gamma}{c}\beta^{-\alpha}c) \approx \frac{1}{4}b\beta^{-2\gamma}\exp(-2\gamma)$. If $\beta = (\mu\ln n)^{1-c}$ the first term in
the derivative is proportional to $n^{-1+\delta\mu}(\ln n)^{\frac{\alpha+c-1}{c}}$ while the second is proportional
to $(\ln n)^{\frac{2\gamma-1}{c}}$; thus for $\beta_L = (\mu_L\ln n)^{1-c}$ with $\mu_L < 1$, the second negative term domi-
nates the growth in the derivative (abias² dominates $MSE(\beta_L)$) and $\iota(L) = 0$. For
$\beta_H = (\mu_H\ln n)^{1-c}$ $\mu_H > 1$, the first term dominates the derivative (avar dominates
$MSE(\beta_H)$) and $\iota(H) = 1 - \delta\mu_H$ (note: only if $\mu = 1$ does the variance not increase
with n). Since $\beta_L < \beta^* < \beta_H$, we get (a), (b), (c), and (d).

Case (iii). Here similarly, $MSE(\beta^*) = an^{-1+\delta\alpha}\exp(\delta)(1 - \frac{\alpha}{2\gamma d}\beta^{-\alpha}d) \approx \frac{1}{4}an^{-1+\delta\alpha}\exp(\delta)$.
If $\beta_L = (\mu_L\ln n)^{1-d}$ and $\mu_L < \frac{1}{2\gamma}$ the derivative is negative, $MSE(\beta_L)$ is dominated by
abias² and declines at a rate with polynomial component $n^{-\iota(L)}$ with $\iota(L) = 2\gamma\mu_L$. If
$\beta_H = (\mu_H\ln n)^{1-d}$ and $\mu_H > \frac{1}{2\gamma}$ the derivative is positive, avar dominates and declines
with $\iota(H) = 1$; (a), (b), (c), and (d) follow.

Case (iv). Note that substituting from the first order condition here $MSE(\beta^*) \approx$
$an^{-1+\delta\alpha}\exp(\delta\beta^{-\alpha c}) \approx - 2\frac{1}{c}d\beta^{-2\gamma+di}c\exp(-2\gamma\beta^{-\alpha d})$. If $c < d$, set $\beta_L = (\mu_L\ln n)^{1-d}; 0 <$
$\mu_L < \frac{1}{2\gamma}$ and $\beta_H = (\mu_H\ln n)^{1-c}; 0 < \mu_H < 1$, then for $MSE(\beta_L) \approx$ abias² we have
$\iota_L = 2\gamma\mu_L$; $MSE(\beta_H) \approx$ avar we have $\iota_L = 1 - \delta\mu_H$. As μ_L (μ_H) is selected
close to $\frac{1}{2\gamma}$ (0) we approach $\iota = 1$ and (a), (b), (c), and (d) follow. If $c > d$,
$\beta_L = (\mu_L\ln n)^{1-c}; 0 < \mu_L < 1$ and $\beta_H = (\mu_H\ln n)^{1-d}; 0 < \mu_H < \frac{1}{2\gamma}$. We get $\iota(L) =$
$\iota(H) = 0$: Since here the expression for $MSE(\beta^*)$ for $\beta^* > \beta_L$ grows faster than
$n^{-1}\exp(\delta\mu_L\ln n) = n^{-1+\delta\mu_L}$ for any μ_L, $\iota = 0$. Thus (a), (b), (c) and (d) follow. For
$c = d$, set $\beta_L = (\mu_L\ln n)^{1-c}; 0 < \mu_L < \frac{1}{\delta+2\gamma}$ and $\beta_H = (\mu_H\ln n)^{1-c}; \frac{1}{\delta+2\gamma} < \mu_H < 1$. In
the expression for $MSE(\beta^*)$ with $\beta^* > \beta_L$ we have $\iota < 1 - \mu_L\delta$ for all $0 < \mu_L < \frac{1}{\delta+2\gamma}$,
therefore considering μ_L arbitrarily close to $\frac{1}{\delta+2\gamma}$ we get $\iota = \frac{2\gamma}{\delta+2\gamma}$: Thus (a), (b),
(c), and (d) follow. If $\alpha = 2\gamma$ an analytic solution $\beta^* = (\frac{1}{\delta+2\gamma})^{1-c}(\ln n + \ln(\frac{2\gamma b}{\delta a}))^{1-c}$
obtains.

This completes the proof. $\blacksquare$

References

Andrews, D.W.K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," Econometrica, 59, 307–345.

Andrews, D.W.K. and M.M.A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," Review of Economic Studies, 65, 497–518.

Buchinsky, M. (1998): "The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach," Journal of Applied Econometrics, 13, 1–30.

Danielsson, J. and C.G. de Vries (1997): "Beyond the Sample: Extreme Quantile and Probability Estimation," unpublished manuscript.

Gallant, R. and D. Nychka (1978): "Semi-Nonparametric Maximum Likelihood Estimation," Econometrica, 55, 363–390.

Gradshteyn, I.S. and I.M. Ryzhik (1994): Table of Integrals, Series, and Products, ed. by A. Je¤rey. Translated from the Russian by Scripta Technica, Inc. London: Academic Press. Fifth Edition.

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model," Journal of Econometrics, 35, 303–316.

Heckman, J.J. (1990): "Varieties of Selection Bias," American Economic Review, 80, 2, 313–318.

Hosking, J., J. Wallis, and E. Wood (1985): "Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments," Technometrics, 27, 251–261.

Huisman, R., K. Koedijk, C. Kool, and F. Palm (1997): "Fat Tails in Small Samples," unpublished manuscript.

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

Ichimura, H. and L.F. Lee (1990): "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics, ed. by W.A. Barnett, J. Powell, and G.E. Tauchen. Cambridge: Cambridge University Press.

Klein, R.W. and R.H. Spady (1993): "An E¢cient Semiparametric Estimator for Binary Response Models," Econometrica, 61, 387–421.

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

Newey, W. (1988): "Two Step Series Estimation of Sample Selection Models," unpublished manuscript, Department of Economics, Princeton University.

Powell, J. (1989): "Semiparametric Estimation of Censored Selection Models," unpublished manuscript, University of Wisconsin-Madison.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coe¢cients," Econometrica, 57, 1435–1460.

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

Schafgans, M.M.A. (1998a): "Ethnic Wage Di¤erences in Malaysia: Parametric and Semiparametric Estimation of the Chinese–Malay Wage-Gap", Journal of Applied Econometrics, 13, 481–504.

_____ (1998b):"A Monte Carlo Study of the Semiparametric Estimation of the Intercept of a Sample Selection Model," unpublished manuscript.