ON INTERCEPT ESTIMATION IN THE SAMPLE SELECTION MODEL*

Marcia M A Schafgans
Department of Economics, London School of Economics

and

Victoria Zinde-Walsh
Department of Economics, McGill University

Contents:

The Suntory Centre
Suntory and Toyota International Centres
for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 0171- 405 7686

Discussion Paper
No. EM/00/380
January 2000

* This work was supported by the Social Sciences and Humanities Research Council of Canada and by the Fonds FCAR (Québec).
The authors would like to thank an anonymous referee and Joel Horowitz for their comments and suggestions.

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 “identification at infinity” 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 Nos.: 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 benefit 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 first 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 $n \to \infty$ is one. The justification 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 difficulty in deriving the asymptotic distribution of the Heckman (1990) estimator, arising from the non-differentiability 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:

$$\begin{aligned} Y_i^{¤} &= {}^1{}_0 + Z_i^0 µ_0 + U_i \ ; \\ D_i &= 1(X_i^0 {}^{-}{}_0 > "_i) \ ; \quad \text{and} \\ Y_i &= Y_i^{¤} D_i \quad \text{for} \quad i = 1; :::; n \ ; \end{aligned} \tag{1}$$

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

$$W_i = X_i^0 {}^{-}{}_0 : \tag{2}$$

The literature on semiparametric estimation of sample selection models gives several root-n consistent and asymptotically normal estimators for the selection parameters, ${}^{-}{}_0$ (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 $D_i = 1$, $®_0$ and $¯_0$ need to be estimated simultaneously using, e.g., Ichimura and Lee (1990). Regarding the selection parameters $¯_0$; furthermore, it should be noted that only the slope parameters are required in the context of estimating the intercept $¹_0$. The loss of identification 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, $¹_0$, which uses these preliminary estimators, was provided by AS. Their estimator, call it the AS estimator, is given by

$$\hat{¹}_s = \frac{\mathbf{P}^n_{i=1}(Y_i ¡ Z_i^0\hat{µ})D_i s(X_i^0\hat{¯} ¡ °_n)}{\mathbf{P}^n_{i=1} D_i s(X_i^0\hat{¯} ¡ °_n)}; \tag{3}$$

where $s(¢)$ is a non-decreasing [0,1]-valued function that has three derivatives bounded over R and for which $s(x) = 0$ for $x · 0$ and $s(x) = 1$ for $x ¸ b$ for some $0 < b < 1$ (AS, Assumption 3). The preliminary estimators $(\hat{µ}; \hat{¯})$ are root-n consistent estimators of $(µ_0; ¯_0)$. The parameter $°_n$ is called the bandwidth or smoothing parameter, where the bandwidth parameter is chosen such that $°_n ! 1$ as $n ! 1$.

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

$$\hat{¹}_I = \frac{\mathbf{P}^n_{i=1}(Y_i ¡ Z_i^0\hat{µ})D_i 1(X_i^0\hat{¯} > °_n)}{\mathbf{P}^n_{i=1} D_i 1(X_i^0\hat{¯} > °_n)} \tag{4}$$

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

Heckman's estimator $\hat{¹}_I$ is essentially a sample average of the random variables $U_i + ¹_0$ over a fraction of all observations, since $Y_i ¡ Z_i^0\hat{µ} !_p U_i + ¹_0$ as $n ! 1$ for all $i ¸ 1$. The effective 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 $s(¢)$; the estimator $\hat{¹}_s$ is a weighted sample average of the random variables $U_i + ¹_0$; where observations with $X_i^0\hat{¯}$ greater than $°_n$ and with $X_i^0\hat{¯}$ 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(¢) (or 1(¢)). 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(¢) is small in comparison. This is confirmed in the analysis of Section 4; nevertheless there are cases when the choice of s(¢) (or 1(¢)) affects 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 $¯_0$ are known, Andrews and Schafgans already showed that the Heckman estimator, $\hat{\ }_{I;0}$, is asymptotically normal (i.e., in Andrews and Schafgans (1998) the indicator replaces the s(¢) function when the true $µ_0$ and $¯_0$ are known).[1]

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

$$\frac{\sqrt{nE\,D_i 1(W_i > °_n)}}{¾}\ ¡\ \hat{\ }_I\ ¡\ \hat{\ }_{I;0}\ ¢\ ¡\overset{p}{!}\ 0; \qquad (5)$$

where $¾^2$ = Var(U_i). Essentially, the proof requires us to deal directly with the non-differentiability of the indicator function.

There are different ways of dealing with asymptotics for non-differentiable functions. Typically assumptions regarding the probability density function are required. This is due to the fact that the expectation of the Dirac ±-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-differentiable function $®_{in}$ given by

$$®_n(\hat{¯}; W_i; X_i) \text{ with } ®_n(¯; W; X) = 1(X^0¯ > °_n)\ ¡\ 1(W > °_n); \qquad (6)$$

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

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

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:$_{W;\,i\,j\,i}$ 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

$$\frac{p{:}d{:}f{:}_W(°_n\ \mathrm{i}\ d)}{Pr(W_i > °_n)^{3=4}} = O(1):$$

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

Similar to Assumptions 4 and 7 of AS, Assumption A(a) relates to the upper tail behaviour of the selection index $W_i = X_i^{0}¯_0$. It is satisfied if W_i has a Weibull, Pareto or "combined" upper tail. If Assumption 4 of AS is satisfied with » = 0, then this condition can be replaced by:

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

In both cases, the condition is less strong than requiring a bounded hazard function on W_i; since p:d:f:$_W(°_n\ \mathrm{i}\ d)^{1=4}$ = 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 satisfied, 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 satisfies Assumption 3^0 of AS.

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

$$\text{(a)}\ \frac{p\,\overline{nED_i1(W_i > °_n)}}{¾}\ µ\ \hat{}_{I}\ \mathrm{i}\ {}^{1}{}_{0}\ \mathrm{i}\ \frac{EU_iD_i1(W_i > °_n)}{ED_i1(W_i > °_n)}\ ¶\ \mathrm{i}^{d}!\ N(0;1)$$

$$\text{(b)}\ \frac{p\,\overline{nED_i1(W_i > °_n)}}{¾}(\hat{}_{I}\ \mathrm{i}\ {}^{1}{}_{0})\ \mathrm{i}^{d}!\ N(0;1)\ \text{i¤ Assumption 8 of AS holds.}$$

[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(¢); Assumption 4 characterizes the upper tail of W_i in terms of a parameter 0 · » < 1=3 with "fatter" tails if » = 0; Assumption 5 is root-n consistency and asymptotic normality of (þ; b); Assumption 6 is $°_n$! 1 ; where its speed is restricted by Assumption 7 in terms of the tail of $W_i > °_n$:

Proof: See Appendix A. 2

To test hypotheses and construct confidence intervals for functions of $(\mu_0, \theta_0, \beta_0)$, we need a joint asymptotic normality result for $(\hat{\mu}_I, \hat{\theta}, \hat{\beta})$. This result, similar to that in AS, is given by

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

$$\begin{pmatrix} \frac{\sqrt{nED_i 1(W_i > \gamma_n)}}{\sigma}(\hat{\mu}_I - \mu_0) \\ \sqrt{n}\,\Omega^{-1/2}\begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} \end{pmatrix} \xrightarrow{d} N(0, I).$$

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 $s(\cdot)$ (or $1(\cdot)$).

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 $\omega(W)$ defined below that determines the asymptotic bias of the estimator. Specifically, let

$$\omega(W) = E_{|W=W} U_i 1(\varepsilon_i > W). \tag{7}$$

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

$$\text{abias}(\hat{\mu}_s) = \frac{-E[\omega(W_i)s(W_i - \gamma_n)]}{[ED_i s(W_i - \gamma_n)]}. \tag{8}$$

Under the bivariate normality assumption of (U_i, ε_i), $\omega(W)$ is equal to $\frac{\sigma_{\varepsilon U}}{\sigma_\varepsilon}\phi(\frac{W}{\sigma_\varepsilon})$; where $\phi(\cdot)$ denotes the standard normal density function.

[3] Since $E(U_i s(W_i > \gamma_n)) = 0$ by independence of U_i and W_i, this follows as $E(U_i D_i s(W_i > \gamma_n)) = -E(U_i(1 - D_i)s(W_i > \gamma_n)) = -E(U_i 1(\varepsilon_i > W_i)s(W_i > \gamma_n))$: By law of iterated expectations, this equals $-E(\omega(W_i)s(W_i > \gamma_n))$.

There are circumstances when $\omega(W)$ may be zero for large enough W, e.g. if the distribution of ε_i has finite support, in which case a comparison of the asymptotic variance will determine the optimal estimator. When $\omega(W)$ differs 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, the asymptotic variance (avar) of the estimator $\hat{\beta}_s$ (or $\hat{\beta}_I$ for $s(\cdot) = 1(\cdot)$) is given by

$$\mathrm{avar}(\hat{\beta}_s) = \frac{\sigma^2 E\left[D_i s^2(W_i - \gamma_n)\right]}{n\left[E D_i s(W_i - \gamma_n)\right]^2}. \tag{9}$$

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

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

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

(b) $|\mathrm{abias}(\hat{\beta}_s)| \leq |\mathrm{abias}(\hat{\beta}_I)|$; if $E1(W_i > \gamma_n)/E1(W_i > \gamma_n + b) \sim 1$ ($\xi = 0$ in Assumption 4 of AS) and $\omega(W_i) \geq 0\ \forall W_i > \gamma_n$ (or $\omega(W_i) \leq 0\ \forall W_i > \gamma_n$).

Proof: See Appendix B. 2

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 $\xi = 0$; if Assumption 4 of AS is satisfied only with $\xi > 0$, we say that W_i is "thin-tailed". Examples of fat tails of W_i are Pareto upper tails (i.e., $1 - F(W) \sim c_W W^{-\lambda}; \lambda > 0$) or Weibull $(\lambda; c)$ upper tails (i.e., $1 - F(W) \sim c_W \exp(-\lambda W^c); \lambda > 0$) with $c \leq 1$. 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 $s(\cdot)$ or $1(\cdot)$ does not affect the asymptotic variance; the asymptotic bias if unaffected if additionally $\omega(W_i)$ does not go to zero too fast.

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

(a) If $E1(W_i > °_n) = E1(W_i > °_n + b)$ ¼ 1 (» = 0 in Assumption 4 of AS) avar($\mathbf{b}_s$) ¼ avar($\mathbf{b}_l$):

(b) If additionally to (a), $jE\,!(W)1(W > °_n + b)j = jE\,!(W)1(W > °_n)j$ ¼ 1 and $!(W_i)$ ¸ 0 8$W_i > °_n$ (or $!(W_i)$ · 0 8$W_i > °_n$); then abias($\mathbf{b}_s$) ¼ abias($\mathbf{b}_l$):

Proof: See Appendix B. 2

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) ¼ $c_W W^{®} e^{¡ ¸ W^c}$ and !(W) ¼ $c_! W^{\&} e^{¡ ¹ W^d}$ where the parameters ®; ¸; c; &; ¹; d are such that the functions 1 ¡ F(W) and !(W) ! 0 as W ! 1 : This class of models includes the Pareto, Weibull (with c · or c > 1) as well as "combined" tails. If U; " and W are jointly normally distributed d = c = 2; ® = ¡1; ¸ = $\frac{1}{2¾^2_W}$; ¹ = $\frac{1}{2¾^2_W}$; 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(¢) functions satisfying the following assumption

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

$$s^{(i)}(0) = \begin{cases} 0 & i < q \\ a_q \; 6\!\!= 0 & i = q \\ \text{exists} & i = q + 1: \end{cases}$$

Note that any function that satisfies AS for which the lowest order of non-zero derivative is q · 2 satisfies 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

[4] In AS it was mistakenly claimed that the normal distribution has a Weibull tail with c = 2; in fact its tail is $W^{¡1} \exp(¡ \frac{1}{2¾^2_W} W^2)(1 + o(W^{¡1}))$:

factors, if they are present the expressions below for avar acquire $c_W^{¡1}$ and for abias $c_!$ as a factor. Furthermore, we omit the subscript n on °.

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

(a) If $1 ¡ F(W) ¼ W^{®} \exp(¡ ¸W^{c})$

$$\text{avar}(^_s) = \begin{cases} \text{avar}(^_l) = ¾^2 n^{¡1} °^{¡®} \exp(¸°^{c}) & \text{if } c · 1 \\ ¾^2 n^{¡1} ¡^{2q}_{q}¢ °^{¡®} \exp(¸°^{c}) & \text{if } c > 1; \end{cases}$$

where avar(^_l) when $c > 1$ obtains for $q = 0$ ($¡^0_0¢ = 1$).

(b) If additionally to (a) $!(W) ¼ W^{&} \exp(¡ ¹W^{d})$

$$\text{abias}(^_s) = \begin{cases} \text{abias}(^_l) = ¡ °^{&} \exp(¡ ¹°^{d}) & \text{if } c · 1; d · 1; d < c \\ \text{abias}(^_l) = ¡ \frac{¸}{¸+¹} °^{&} \exp(¡ ¹°^{c}) & \text{if } c · 1; d · 1; c = d[5] \\ \text{abias}(^_l) = ¡ \frac{¸c}{¹d} °^{&+c¡d} \exp(¡ ¹°^{d}) & \text{if } c · 1; d > c[6] \\ ¡ \frac{a_q ¸c}{(¹d)^{q+1}} °^{&+(c¡d)(1+q)} \exp(¡ ¹°^{d}) & \text{if } c · 1; d > 1[6] \\ ¡ °^{&} \exp(¡ ¹°^{d}) & \text{if } c > 1; d < c; \\ ¡ \left(\frac{¸}{¸+¹}\right)^{q+1} °^{&} \exp(¡ ¹°^{c}) & \text{if } c > 1; d = c \\ ¡ \left(\frac{¸c}{¹d}\right)^{q+1} °^{&+(c¡d)(1+q)} \exp(¡ ¹°^{d}) & \text{if } c > 1; d > c; \end{cases}$$

where abias(^_l) obtains for $q = 0$ ($0! = 1$) where it is not defined explicitly and a_0 ´ 1.

Proof: See Appendix B. 2

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

As an example of Proposition 3, if U; " and W are jointly normally distributed ($c = d = 2$), the asymptotic bias and variance of ^_s (including all relevant constant

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

[6] If $c = 0$, then the expression ¸c in the constant becomes ¡®:

factors) equals

$$\text{abias}(\text{^}_s) \quad = \quad \text{i p}\frac{\text{¾"U}}{\overline{\text{2¼}}\text{¾"}}(\frac{\text{¾}^2_{\text{"}}}{\text{¾}^2_{\text{"}} + \text{¾}^2_W})^{q+1}\exp(\text{ i }\frac{1}{2\text{¾}^2_{\text{"}}}\text{°}^2)$$

$$\text{avar}(\text{^}_s) \quad = \quad \frac{\text{¾}^2}{n}\frac{\text{p}\overline{\text{2¼}}}{\text{¾}_W}\text{°}\,\text{µ}\frac{2q}{q}\text{¶}\exp(\frac{1}{2\text{¾}^2_W}\text{°}^2):$$

The asymptotic bias and variance of ^_l 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 $\text{°}^{\text{¤}}$ = argmin(MSE) as a function of n, q (the s(¢) 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 ° ! 1 can be presented as $\text{°}^{\text{¤}} = (\frac{\text{µ}}{\text{¸}}\ln n)^{1=c}$ if c 6= 0 and $\text{°}^{\text{¤}} = n^{\text{i µ=®}}$ if c = 0 with µ close to zero and would result in a MSE proportional to $n^{\text{i }1+\text{µ}}$. Proposition 4 deals with situations where an asymptotic bias is present and may be severe. It characterizes the bandwidth parameter $\text{°}^{\text{¤}}$ 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 $\text{°}^{\text{¤}}$ 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 $1 \text{ ¡ } F(W) \text{ ¼ } c_W W^{\text{®}} e^{\text{i ¸}W^c}$ and $\text{!}(W) \text{ ¼ } c_{\text{!}} W^{\text{&}} e^{\text{i ¹}W^d}$ as W ! 1:

(a) There exists a sequence $\text{°}^{\text{¤}}_n$ unique up to $o(\text{°}^{\text{¤ i}v})$ for some v > 0 that minimizes the asymptotic $MSE(\text{¹}_s)$ (or ¹_l):

(b) The optimal asymptotic $MSE^{\text{¤}}$ can be represented as a product of a polynomial component $n^{\text{i ¿}}$; ¿ ¸ 0 and a logarithmic component $O((\ln n)^{\text{°}})$; where ¿ depends only on the parameters which characterize the leading term in the tail of W, i.e. ® for a Pareto and ¸; c for a Weibull or combined tail, and parameters of the leading term of !(W):

(c) There exist bounds °_H and °_L such that $\text{°}_L < \text{°}^{\text{¤}} < \text{°}_H$; where °_H and °_L are functions of the coefficients of the leading terms in tail of W and $\omega(W)$ only and $MSE(\gamma_H)$, $MSE(\gamma_L)$ decline at a rate with the polynomial component $n^{-\tau(H)}; n^{-\tau(L)}$ with one (or both) of $\tau(H)$ and $\tau(L)$ either equal to τ; or arbitrarily close to τ:

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

Proof: See Appendix B. □

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

As an example of Proposition 4, if $U; \varepsilon$ and W are jointly normally distributed $(c = d = 2)$, γ^* is bounded by $\gamma_L = (\theta_L \ln n)^{1/2}; \theta_L < 2\sigma_\varepsilon^2\sigma_W^2/(\sigma_\varepsilon^2 + 2\sigma_W^2)$, and $\gamma_H = (\theta_H \ln n)^{1/2}; \theta_H > 2\sigma_\varepsilon^2\sigma_W^2/(\sigma_\varepsilon^2 + 2\sigma_W^2)$. The optimal asymptotic MSE has the polynomial component $n^{-\tau}$ with $\tau = 1 - \sigma_\varepsilon^2/(\sigma_\varepsilon^2 + 2\sigma_W^2)$. We note that τ can get arbitrarily close to 1 if $\sigma_W^2 \gg \sigma_\varepsilon^2$: (In general, if $c = d$, τ can be made arbitrarily close to 1 given $\lambda \ll \mu$).The $MSE(\gamma_H) \approx avar(\hat{b}_s)$ with $\tau_H = 1 - \theta_H/(2\sigma_W^2)$, and $MSE(\gamma_L) \approx abias(\hat{b}_s)$ with $\tau_L = \theta_L/\sigma_\varepsilon^2$: Both τ_L and τ_H are arbitrarily close to τ when $\theta_L; \theta_H$ 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 $\omega(W)$ are known solving the first-order condition for γ and then substituting into the MSE and minimizing over q (where it appears) as well would give us the "optimal" estimator. When the bounds γ_H, or γ_L, are used instead of γ^* 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 γ_H, or γ_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 $\omega(W_i)$ 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 "identification at infinity," which was first proposed by Heckman (1990). Technical problems in derivations arise from the non-differentiability 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 $\omega(W)$ that determines the asymptotic bias of the estimator. A wide class of distributional assumptions for the model is investigated, specifically, $1 - F(W) \approx c_W W^{\alpha} e^{-\lambda W^c}$ and $\omega(W) \approx c_{\omega} W^{\varsigma} e^{-\mu W^d}$: 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 affect the asymptotic variance; the asymptotic bias if unaffected if additionally $\omega(W_i)$ does not go to zero too fast. In general, however, the asymptotic MSE may be affected 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 first non-zero derivative of $s(\cdot)$ at zero.

If all the parameters determining the tail behaviour of W and the function $\omega(W)$ were known a solution that would provide an optimal $\gamma^* = \mathrm{argmin}(MSE)$ as a function of n, q (the $s(\cdot)$ 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 $s(\cdot)$ (or $1(\cdot)$) can be obtained. Since the solution for the optimal bandwidth (and choice of function $s(\cdot)$ (or $1(\cdot)$) 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 $\omega(W_i)$ 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 finite samples, the AS estimator might still have advantages over the Heckman estimator, in that the trade-off between bias and variance, like in nonparametric estimation problems, is better for smooth "kernels". Nevertheless, only observations at the margin are affected by the choice of the function $s(\cdot)$. In Schafgans (1998b) simulations are presented that reveal these findings 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 sufficient conditions for (5). As in the proof of Theorem A-1 of AS the left hand side of (5) can be written as $C(\frac{\hat{A}}{\hat{B}} - \frac{A}{B}) = C\frac{\hat{A}-A}{B}\frac{B}{\hat{B}} - C\frac{\hat{B}-B}{B}\frac{A}{B}\frac{B}{\hat{B}}$, where $C = \sqrt{nED_i 1(W_i > \gamma_n)}/\sigma$, $A = \sum_{i=1}^{n}(Y_i - Z_i'\theta_0)D_i 1(W_i > \gamma_n)$, $\hat{A} = \sum_{i=1}^{n}(Y_i - Z_i'\hat{\theta})D_i 1(X_i'\hat{\beta} > \gamma_n)$, $B = \sum_{i=1}^{n} D_i 1(W_i > \gamma_n)$, and $\hat{B} = \sum_{i=1}^{n} D_i 1(X_i'\hat{\beta} > \gamma_n)$. To show (5), therefore, it suffices to show that

$$
\begin{aligned}
&\text{(i)} && \hat{B}/B \xrightarrow{p} 1 \\
&\text{(ii)} && C(\hat{A} - A)/B \xrightarrow{p} 0 \\
&\text{(iii)} && A/B = O_p(1) \\
&\text{(iv)} && C(\hat{B} - B)/B \xrightarrow{p} 0.
\end{aligned}
\tag{A.1}
$$

From Assumption 7 and Lemma A-2 of AS it follows that $C \to \infty$ which means that (iv) implies (i) in (A.1). From Lemmas A-1 and A-2 of AS one gets that $\sigma C/B$ equals $n^{-1/2}\Pr(W_i > \gamma_n)^{-1/2}$ in probability and thus (ii) is implied by the following sufficient conditions.

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i+\mu_0)D_i\alpha_{in}}{\Pr(W_i>\gamma_n)^{1/2}}\xrightarrow{p}0, \tag{A.2}$$

$$\frac{-(\hat{\theta}-\theta_0)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_iD_i\alpha_{in}}{\Pr(W_i>\gamma_n)^{1/2}}\xrightarrow{p}0; \text{ and} \tag{A.3}$$

$$\frac{-(\hat{\theta}-\theta_0)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_iD_i1(W_i>\gamma_n)}{\Pr(W_i>\gamma_n)^{1/2}}\xrightarrow{p}0, \tag{A.4}$$

where α_{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

$$\frac{\frac{1}{n}\sum_{i=1}^{n}U_iD_i1(W_i>\gamma_n)}{\Pr(W_i>\gamma_n)}=O_p(1). \tag{A.5}$$

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

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_i\alpha_{in}}{\Pr(W_i>\gamma_n)^{1/2}}\xrightarrow{p}0. \tag{A.6}$$

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

$$\frac{O_p(1)E\|Z_i\|1(W_i>\gamma_n)}{\Pr(W_i>\gamma_n)^{1/2}}\le O_p(1)\left(E\|Z_i\|^3\right)^{1/3}(\Pr(W_i>\gamma_n))^{2/3-1/2}\to 0. \tag{A.7}$$

The proof of Theorem 1, therefore, requires us to show that (A.2), (A.3) and (A.6) which involve the discontinuous function α hold.

We do this in three steps: first, a technical lemma is given (Lemma 1), then a lemma is given which examines terms in the expressions of interest (Lemma 2) and finally 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 sufficiently slowly increasing $\{M_n\}$ such that $n^{-1/2}M_n<\infty$ and $\sup_{\|\beta-\beta_0\|<\frac{M_n}{\sqrt{n}}}\frac{E(\sqrt{n}\|\Omega_i\|\alpha_n(\beta;W_i;X_i))}{\Pr(W_i>\gamma_n)^{1/2}}\to 0$ as $n\to\infty$ for $\Omega_i=Z_i/\sqrt{n}$, $\Omega_i=U_i$, or $\Omega_i=1;\forall i$.

Proof of Lemma 1:

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

$$\sup_{\|\beta-\beta_0\|<\frac{M_n}{\sqrt{n}}} \frac{\sqrt{n}E\left(\|\varepsilon_i\|\cdot|\alpha_n(\beta,W_i,X_i)|\cdot 1(\|X_i\|>n^{\frac{1}{3}})\right)}{\Pr(W_i>\gamma_n)^{1/2}} \to 0, \qquad \text{(A.8)}$$

and

$$\sup_{\|\beta-\beta_0\|<\frac{M_n}{\sqrt{n}}} \frac{\sqrt{n}E\left(\|\varepsilon_i\|\cdot|\alpha_n(\beta,W_i,X_i)|\cdot 1(\|X_i\|\leq n^{\frac{1}{3}})\right)}{\Pr(W_i>\gamma_n)^{1/2}} \to 0. \qquad \text{(A.9)}$$

Noting that $|\alpha_n(\beta,W_i,X_i)|\leq 1$, we get for (A.8) in the case $\varepsilon_i=U_i$ (using the independence condition of Assumption 2(b) of AS) and similarly for $\varepsilon_i=1$, under Assumptions 1 and 7 of AS

$$\begin{aligned}
&\frac{\sqrt{n}E\left(\|\varepsilon_i\|\cdot|\alpha_n(\beta,W_i,X_i)|\cdot 1(\|X_i\|>n^{1/3})\right)}{\Pr(W_i>\gamma_n)^{1/2}} \leq \frac{\sqrt{n}E\left(\|\varepsilon_i\|\cdot 1(\|X_i\|>n^{1/3})\right)}{\Pr(W_i>\gamma_n)^{1/2}}\\
\leq\ &\frac{\sqrt{n}E\|\varepsilon_i\|\cdot\Pr(\|X_i\|>n^{1/3})}{\Pr(W_i>\gamma_n)^{1/2}} \leq \frac{\sqrt{n}E\|\varepsilon_i\|\cdot E\|X_i\|^3/n}{\Pr(W_i>\gamma_n)^{1/2}}\\
=\ &\frac{E\|\varepsilon_i\|\cdot E\|X_i\|^3}{(n\Pr(W_i>\gamma_n))^{1/2}} \to 0
\end{aligned} \qquad \text{(A.10)}$$

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

$$\frac{E\left(\|Z_i\|\cdot|\alpha_n(\beta,W_i,X_i)|\cdot 1(\|X_i\|>n^{1/3})\right)}{\Pr(W_i>\gamma_n)^{1/2}} \leq \frac{\left(E\|Z_i\|^3\right)^{1/3}(E\|X_i\|^3)^{2/3}}{n^{1/6}\cdot(n\Pr(W_i>\gamma_n))^{1/2}} \to 0. \qquad \text{(A.11)}$$

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

$$\frac{\left(E\|Z_i\|^2\right)^{\frac{1}{2}}}{(n\Pr(W_i>\gamma_n))^{1/4}}\left(\sup_{\|\beta-\beta_0\|<\frac{M_n}{\sqrt{n}}}\frac{\sqrt{n}E|\alpha_n(\beta,W_i,X_i)|1(\|X_i\|<n^{1/3})}{\Pr(W_i>\gamma_n)^{1/2}}\right)^{1/2} \qquad \text{(A.12)}$$

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

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

$$E\left(\|\varepsilon_i\|\right)\cdot\sup_{\|\beta-\beta_0\|<\frac{M_n}{\sqrt{n}}}\sqrt{n}E|\alpha_n(\beta,W_i,X_i)|1(\|X_i\|<n^{1/3}). \qquad \text{(A.13)}$$

By examining the function $\alpha_n(\beta,W_i,X_i)$, (6), we can see that $|\alpha_n(\beta,W_i,X_i)|$ equals 1 if either $\gamma_n<W_i<\gamma_n+|X_i'(\beta-\beta_0)|$ for negative $X_i'(\beta-\beta_0)$ or if $\gamma_n-|X_i'(\beta-\beta_0)|<$

$W_i < °_n$ for positive $X_i^0(¯ ¡ ¯_0)$ and zero otherwise. In view of the restrictions on $k¯ ¡ ¯_0k$ and kX_ik we have $jX_i^0(¯ ¡ ¯_0)j < \frac{M_n}{n^{1=6}}$, which implies $W_i > °_n ¡ \frac{M_n}{n^{1=6}}$. Hence,

$$\begin{aligned} &\sup_{k¯ ¡ ¯_0k<\frac{M_n}{pn}} {}^p\overline{n}E\, j®_n(¯;W_i;X_i)j\,1(kX_ik < n^{1=3}) \\ = &\sup_{k¯ ¡ ¯_0k<\frac{M_n}{pn}} {}^p\overline{n}E\, j®_n(¯;W_i;X_i)j\,1(W_i > °_n ¡ \tfrac{M_n}{n^{1=6}})1(kX_ik < n^{1=3}): \end{aligned} \qquad (A.14)$$

Consider the linear transformation $L : X ! \binom{W}{¡}$, where $¡ = X_{(¡1)}$ is the vector of all components of X except the …rst (NB the matrix of this transformation is $L = \begin{pmatrix} ¯_{01} & ::: & ¯_{0k} \\ 0 & & I_{k¡1}\end{pmatrix}$; which is non-singular). Then $X = L^{¡1}\binom{W}{¡}$ and $X^0(¯ ¡ ¯_0) = W(L^{0¡1}(¯ ¡ ¯_0))_{(1)} + ¡^0(L^{0¡1}(¯ ¡ ¯_0))_{(¡1)}$, where subscript (1) denotes the …rst component of a vector. We denote $(L^{0¡1}(¯ ¡ ¯_0))_{(1)} = b^7$ and $(L^{0¡1}(¯ ¡ ¯_0))_{(¡1)} = B$. Re-examining, for negative $X_i^0(¯ ¡ ¯_0)$; the condition $°_n < W_i < °_n + jX_i^0(¯ ¡ ¯_0)j = °_n ¡ W(L^{0¡1}(¯ ¡ ¯_0))_{(1)} ¡ ¡^0(L^{0¡1}(¯ ¡ ¯_0))_{(¡1)}$ we realize that for negative $X_i^0(¯ ¡ ¯_0)$ $j®_n(¯;W_i;X_i)j$ equals 1 for $°_n < W_i < \frac{°_n ¡ ¡^0B}{1+b}$. Similarly for positive $X_i^0(¯ ¡ ¯_0)$ $j®_n(¯;W_i;X_i)j$ equals 1 for $\frac{°_n ¡ ¡^0B}{1+b} < W_i < °_n$. Thus (A.14) equals

$$\sup_{k¯ ¡ ¯_0k<\frac{M_n}{pn}} {}^p\overline{n}E \left[\int_{°_n}^{\frac{°_n ¡ ¡^0B}{1+b}} 1(X_i^0(¯ ¡ ¯_0) < 0)1(kX_ik < n^{1=3})p{:}d{:}f{:}_{W;¡j¡}(W;¡)dW + \int_{\frac{°_n ¡ ¡^0B}{1+b}}^{°_n} 1(X_i^0(¯ ¡ ¯_0) > 0)1(W_i > °_n ¡ \tfrac{M_n}{n^{1=6}})1(kX_ik < n^{1=3})p{:}d{:}f{:}_{W;¡j¡}(W;¡)dW \right]: \qquad (A.15)$$

The $p.d.f._{W;¡j¡}$ exists for large enough n, since the expression under the integrals is non-zero only if $W_i > °_n ¡ \frac{M_n}{n^{1=6}}$ and thus as $°_n ! 1$ for $M_n = O(n^{1=6})$ becomes greater than A; the conditional $p.d.f._{W;¡j¡}$ declines monotonically in W: Denote $°_n ¡ \frac{M_n}{n^{1=6}}$ by $\overset{º}{°}_n$: The …rst integral on the right hand side of (A.15) is bounded by $p{:}d{:}f{:}_{W;¡j¡}(º_n;¡)¢\left(\frac{°_n ¡ ¡^0B}{1+b} ¡ °_n\right)$, the second can be bounded by $p{:}d{:}f{:}_{W;¡j¡}(º_n;¡)¢\left(°_n ¡ \frac{°_n ¡ ¡^0B}{1+b}\right)$: Thus the sum is bounded by

$$\sup_{k¯ ¡ ¯_0k<\frac{M_n}{pn}} 2\,{}^p\overline{n} \int µ p{:}d{:}f{:}_{W;¡j¡}(º_n;¡)\left|\frac{¡^0B}{1+b}\right|¶ dPr_¡ \qquad (A.16)$$

[7] Note: $jbj \cdot °L^{¡1}°¢k¯ ¡ ¯_0k \cdot °L^{¡1}° \frac{M_n}{pn}$. Assume that M_n is such that $°L^{¡1}° \frac{M_n}{pn} \cdot G < 1$.

$$\leq \tfrac{2M_n}{1+b} \int \|\Gamma\| p.d.f._W(\tilde{\gamma}_n) dPr_{\Gamma|W=\tilde{\gamma}_n} = \tfrac{2M_n}{1+b} p.d.f._W(\tilde{\gamma}_n) E_{|W=\tilde{\gamma}_n} \|\Gamma\|;$$

where $\Gamma = X_{(-1)}$ and $\|B\| = \left\|(L^{-1}(\beta - \beta_0))_{(-1)}\right\| \equiv$ [8] $\left\|(\beta - \beta_0)_{(-1)}\right\| \leq \frac{M_n}{n^{1/2}}$ using the restriction on $\|\beta - \beta_0\|$. The notation dPr_Γ indicates that a Stiltjes integral with respect to the cumulative probability function Γ is taken, if a marginal density exists $dPr_\Gamma = p.d.f._\Gamma(\Gamma)d\Gamma$. By Jensen's inequality, and in view of Assumption A(b) this is bounded almost surely by

$$\tfrac{2M_n}{1+b} \cdot p.d.f._W(\tilde{\gamma}_n) \cdot E_{|W=\tilde{\gamma}_n} \left(\left\|X_{(-1)}\right\|^3\right)^{1/3}. \tag{A.17}$$

Let us consider $E_{|W=\tilde{\gamma}_n}\left(\left\|X_{(-1)}\right\|^3\right)$. By Assumption 1 of AS, we know that the unconditional expectation $E\left(\left\|X_{(-1)}\right\|^3\right)$ is bounded; this implies that for the fixed A of Assumption A $E\left(\left\|X_{(-1)}\right\|^3 1(W > A)\right)$ is bounded as well.

$$\begin{aligned} E\left(\left\|X_{(-1)}\right\|^3 1(W > A)\right) &= \iint \left\|X_{(-1)}\right\|^3 1(W > A) p.d.f._{W;\Gamma|W}(\Gamma) p.d.f._W(W) d\Gamma dW \\ &= \int J(W) p.d.f._W(W) 1(W > A) dW; \end{aligned} \tag{A.18}$$

where $J(W) = \int \left\|X_{(-1)}\right\|^3 p.d.f._{W;\Gamma|W}(\Gamma)d\Gamma$ is a function of W only. Since the integral $\int J(W)p.d.f._W(W)dW \equiv E\left\|X_{(-1)}\right\|^3$ exists, it implies that as $W \to \infty$ $J(W)p.d.f._W(W) = o(W^{-1})$, or

$$J(W) = o\left[W^{-1}(p.d.f._W(W))^{-1}\right]. \tag{A.19}$$

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

$$\tfrac{2M_n}{1+b} \cdot p.d.f._W(\tilde{\gamma}_n) \cdot J(\tilde{\gamma}_n)^{1/3} = \tfrac{2M_n}{1+b} \cdot p.d.f._W(\gamma_n - \tfrac{M_n}{n^{1/6}})^{2/3} o(\gamma_n^{-1/3}). \tag{A.20}$$

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

$$\begin{aligned} &\frac{E(\|\varepsilon_i\|) \tfrac{2M_n}{1+b} \cdot p.d.f._W(\gamma_n - \tfrac{M_n}{n^{1/6}})^{2/3} o(\gamma_n^{-1/3})}{Pr(W_i > \gamma_n)^{1/2}} \\ &\leq E(\|\varepsilon_i\|) \tfrac{2M_n}{1+b} \left(\frac{p.d.f._W(\gamma_n - \tfrac{M_n}{n^{1/6}})}{Pr(W_i > \gamma_n)^{3/4}}\right)^{2/3} o(\gamma_n^{-1/3}). \end{aligned} \tag{A.21}$$

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

Set $M_n = \min\{dn^{1/6}, \gamma^{1/3}\}$ 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. $\square$

The next lemma will help to show that terms, involving α_{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 $\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}} \xrightarrow{p} 0$ as $n \to \infty$ for $\Omega_i = Z_i/\sqrt{n}$, $\Omega_i = U_i$, or $\Omega_i = 1$, $\forall i$.

Proof of Lemma 2:

We would like to show that for any $\zeta > 0$

$$\Pr\left(\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_{in}|}{\Pr(W_i > \gamma_n)^{1/2}} > \zeta\right) \to 0 \text{ as } n \to \infty. \qquad \text{(A.22)}$$

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

$$\Pr\left(\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}} > \zeta, \left\|\hat\beta - \beta_0\right\| \le \frac{M_n}{\sqrt{n}}\right) + \Pr\left(\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}} > \zeta, \left\|\hat\beta - \beta_0\right\| > \frac{M_n}{\sqrt{n}}\right)$$
$$\le \Pr\left(\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}} > \zeta, \left\|\hat\beta - \beta_0\right\| \le \frac{M_n}{\sqrt{n}}\right) + \Pr\left(\left\|\hat\beta - \beta_0\right\| > \frac{M_n}{\sqrt{n}}\right). \qquad \text{(A.23)}$$

The second expression on the right hand side of (A.23) converges to zero by Assumption 5 of AS, since $\sqrt{n}(\hat\beta - \beta_0) = O_p(1)$ implies $\sqrt{n}(\hat\beta - \beta_0)/M_n = o_p(1)$. The first expression on the right hand side of (A.23) is bounded by

$$\sup_{\|\beta-\beta_0\| \le \frac{M_n}{\sqrt{n}}} \Pr\left(\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_n(\beta, W_i, X_i)|}{\Pr(W_i>\gamma_n)^{1/2}} > \zeta\right) \qquad \text{(A.24)}$$
$$\le \sup_{\|\beta-\beta_0\| \le \frac{M_n}{\sqrt{n}}} \frac{E\left(\frac{\frac{1}{\sqrt{n}}\sum\|\Omega_i\||\alpha_n(\beta, W_i, X_i)|}{\Pr(W_i>\gamma_n)^{1/2}}\right)}{\zeta} = \sup_{\|\beta-\beta_0\| \le \frac{M_n}{\sqrt{n}}} \frac{\sqrt{n}E\|\Omega_i\||\alpha_n(\beta, W_i, X_i)|}{\zeta \Pr(W_i > \gamma_n)^{1/2}},$$

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

Proof of Theorem 1: For our proof it is sufficient to show (A.2), (A.3), and (A.6). By Assumptions 1 and 5 for $\hat\theta - \theta_0$ of AS, the left-hand sides of (A.2), (A.3) can be bounded by:

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|U_i|\cdot|\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}}+|\mu_0|\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}}; \tag{A.25}$$

$$\frac{O_p(1)\frac{1}{n}\sum_{i=1}^{n}\|Z_i\|\cdot|\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}}, \tag{A.26}$$

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

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\alpha_{in}|}{\Pr(W_i>\gamma_n)^{1/2}}, \tag{A.27}$$

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

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 $\hat{\boldsymbol{\beta}}_l$ is replaced by $\hat{\boldsymbol{\beta}}_{l,0}$. $\Box$

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

$$\text{abias}(\hat{\boldsymbol{\beta}}_s)\approx-\frac{E\omega(W_i)s(W_i-\gamma_n)}{Es(W_i-\gamma_n)};\quad \text{abias}(\hat{\boldsymbol{\beta}}_l)\approx-\frac{E\omega(W_i)1(W_i>\gamma_n)}{E1(W_i>\gamma_n)}; \tag{B.1}$$

$$\text{avar}(\hat{\boldsymbol{\beta}}_s)\approx\sigma^2 n^{-1}\frac{Es^2(W_i-\gamma_n)}{[Es(W_i-\gamma_n)]^2};\quad \text{avar}(\hat{\boldsymbol{\beta}}_l)\approx\sigma^2 n^{-1}(E1(W_i>\gamma_n))^{-1}. \tag{B.2}$$

In the following we let $\omega(W_i)\geq 0\ \forall W_i>\gamma_n$ (similar proofs can be given when $\omega(W_i)\leq 0\ \forall W_i>\gamma_n$).

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

$$Es(W_i-\gamma_n)=Es(W_i-\gamma_n)1(W_i>\gamma_n)\leq[Es^2(W_i-\gamma_n)]^{1/2}[E1(W_i>\gamma_n)]^{1/2}. \tag{B.3}$$

This inequality combined with (B.2) gives the result that (a) holds for any sequence $°_n$.

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

$$\text{jabias}(\mathbf{b}_s)\text{j} ¼ \frac{E\,!(W_i)s(W_i\ ¡\ °_n)}{Es(W_i\ ¡\ °_n)} \quad · \quad \frac{E\,!(W_i)1(W_i > °_n)}{E1(W_i > °_n + b)}$$

$$¼ \text{jabias}(\mathbf{b}_l)\text{j} \frac{E1(W_i > °_n)}{E1(W_i > °_n + b)}: \qquad \text{(B.4)}$$

Since $E1(W_i > °_n)=E1(W_i > °_n + b)$ ¼ 1 (b) follows. 2

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

$$\frac{Es^2(W_i\ ¡\ °_n)}{[Es(W_i\ ¡\ °_n)]^2}\ ¸\ [E1(W_i > °_n)]^{¡1}: \qquad \text{(B.5)}$$

Furthermore as $Es^2(W_i\ ¡\ °_n)\ ·\ Es(W_i\ ¡\ °_n)$ and $Es(W_i\ ¡\ °_n)\ ¸\ E1(W_i > °_n + b)$;

$$\frac{Es^2(W_i\ ¡\ °_n)}{[Es(W_i\ ¡\ °_n)]^2}\ ·\ [E1(W_i > °_n + b)]^{¡1}\ ¼\ [E1(W_i > °_n)]^{¡1}; \qquad \text{(B.6)}$$

and (a) follows.

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

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

In addition, we have

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

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) $\mathbf{R}_{°}^{1}\ ¸c(W\ ¡\ °)^i W^v \exp(¡\,¸W^c)dW = °^{i+v¡c+1¡ci}\exp(¡\,¸°^c)(¸c)^{¡i} i!(1 + o(°^{¡c}))$

(b) For a>f, $\int_{\gamma}^{\infty}\mu a(W-\gamma)^{i}W^{v}\exp(-\mu W^{a}-\lambda W^{f})dW =$

$$\gamma^{i+v-a+1-ai}\exp(-\mu\gamma^{a}-\lambda\gamma^{f})(\mu a)^{-i}i!(1+o(\gamma^{-a})).$$

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

$$\begin{aligned} I &= \int_{\gamma}^{\infty}\lambda c\sum_{j=0}^{i}\binom{i}{j}(-1)^{i+j}\gamma^{i-j}W^{j+v}\exp(-\lambda W^{c})dW \\ &= \sum_{j=0}^{i}\binom{i}{j}(-1)^{i+j}\gamma^{i-j}\int_{\gamma}^{\infty}\lambda cW^{j+v}\exp(-\lambda W^{c})dW. \end{aligned} \qquad (B.9)$$

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

$$I = \sum_{j=0}^{i}\binom{i}{j}(-1)^{i+j}\gamma^{i-j}\cdot\lambda\int_{\gamma^{c}}^{\infty}z^{\frac{j+v-c+1}{c}}\exp(-\lambda z)dz. \qquad (B.10)$$

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

$$\begin{aligned} I &= \sum_{j=0}^{i}\binom{i}{j}(-1)^{i+j}\gamma^{i-j}\lambda\cdot\lambda^{-\frac{j+v+1}{c}}\Gamma\left(\frac{j+v+1}{c};\lambda\gamma^{c}\right) \\ &= \sum_{j=0}^{i}\binom{i}{j}(-1)^{i+j}\gamma^{i-j}\lambda\cdot\gamma^{j+v-c+1}\exp(-\lambda\gamma^{c})\cdot \\ &\qquad \left(\sum_{m=0}^{L-1}\frac{(-1)^{m}\Gamma(1-\frac{j+v+1}{c}+m)}{\gamma^{cm}\lambda^{m}\Gamma(1-\frac{j+v+1}{c})}+o(\gamma^{-cL})\right) \qquad (B.11) \\ &= (-1)^{i}\gamma^{i+v-c+1}\exp(-\lambda\gamma^{c})\sum_{j=0}^{i}\binom{i}{j}(-1)^{j}\left(\sum_{m=0}^{L-1}\frac{\sum_{s=0}^{m}\kappa_{s}(\frac{j}{c})^{s}}{\gamma^{cm}\lambda^{m}}+o(\gamma^{-cL})\right); \end{aligned}$$

where we have substituted $(-1)^{m}\Gamma(1-\frac{j+v+1}{c}+m)/\Gamma(1-\frac{j+v+1}{c}) = \sum_{s=0}^{m}\kappa_{s}(\frac{j}{c})^{s}$, with $\kappa_{s}=1$ for $s=m$ and some known constant for $s=0,\ldots,m-1$. Using the fact that $\sum_{j=0}^{i}\binom{i}{j}(-1)^{j}j^{m}=0$ for $i\geq m+1\geq 1$; $0^{0}\equiv 1$ (see 0.154#3 in Gradshteyn and Ryzhik

[9] There is a typographical error in 3.381#3 in G&R; the correct formula reads $\int_{u}^{\infty}x^{v-1}e^{-\mu x}dx = \mu^{-v}\Gamma(v;\mu u)$.

(1994)), we get

$$\begin{aligned} I &= (-1)^{i}\gamma^{i+v-c+1}\exp(-\lambda\gamma^{c})\sum_{j=0}^{i}\binom{i}{j}(-1)^{j}\left(\frac{(\frac{j}{c})^{i}}{\gamma^{ci}\lambda^{i}}(1+o(\gamma^{-c}))\right) \\ &= \gamma^{i+v-c+1-ci}\exp(-\lambda\gamma^{c})(\lambda c)^{-i}i!(1+o(\gamma^{-c})); \qquad \text{(B.12)} \end{aligned}$$

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

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

$$\begin{aligned} I' &= \exp(-\lambda\gamma^{f})\int_{\gamma}^{\infty}\mu a(W-\gamma)^{i}W^{v}\exp(-\mu W^{a})\left[\exp(-\lambda(W^{f}-\gamma^{f}))\right]dW \qquad \text{(B.13)} \\ &= \exp(-\lambda\gamma^{f})\int_{\gamma}^{\infty}\mu a(W-\gamma)^{i}W^{v}\exp(-\mu W^{a})\left(\sum_{r=0}^{\infty}\frac{(-1)^{r}\lambda^{r}(W^{f}-\gamma^{f})^{r}}{r!}\right)dW; \end{aligned}$$

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

$$I' = \exp(-\lambda\gamma^{f})\sum_{r=0}^{\infty}\frac{(-1)^{r}\lambda^{r}}{r!}\int_{\gamma}^{\infty}\mu a(W^{f}-\gamma^{f})^{r}(W-\gamma)^{i}W^{v}\exp(-\mu W^{a})dW. \qquad \text{(B.14)}$$

To complete the proof, we need to reapply the steps taken in (B.9)–(B.12). The final 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. 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, $\text{avar}(\mu_l) \approx \text{avar}(\mu_s)$ if $1-F(W) \approx W^{\alpha}e^{-\lambda W^{c}}$ with $c \leq 1$, since Assumption 4 of AS holds with $\xi = 0$ and $E1(W > \gamma)/E1(W > \gamma + b) \approx 1$. The variance in these cases can be easily derived using $\sigma^2/(nE1(W_i > \gamma))$; see (B.2).[10]

When on the other hand $1-F(W) \approx W^{\alpha}e^{-\lambda W^{c}}$ with $c > 1$ we need to derive $\text{avar}(\mu_s)$ using the definition in (B.2) (Assumption 4 of AS does not hold with $\xi = 0$).

[10] For $W > \gamma$ the p.d.f.(W) equals $\lambda cW^{\alpha+c-1}\exp(-\lambda W^{c})(1+o(\gamma^{-c}))$, where $\alpha = -\lambda$ if $c = 0$ (Pareto tail case).

For this we need to apply Lemma 3(a). Notice that for any $\epsilon > 0$ we can write $Es(W_i - \gamma)$ as

$$\begin{aligned} Es(W_i - \gamma) &= \frac{a_q}{q!}\int_{\gamma}^{\infty} (W - \gamma)^q \lambda c W^{\alpha+c-1}\exp(-\lambda W^c)dW \\ &\quad - \frac{a_q}{q!}\int_{\gamma+\epsilon}^{\infty} (W - \gamma)^q \lambda c W^{\alpha+c-1}\exp(-\lambda W^c)dW \qquad (B.15) \\ &\quad + \int_{\gamma}^{\gamma+\epsilon}\left[s(W - \gamma) - \frac{a_q}{q!}(W - \gamma)^q\right]\lambda c W^{\alpha+c-1}\exp(-\lambda W^c)dW \\ &\quad + \int_{\gamma+\epsilon}^{\infty} s(W - \gamma)\lambda c W^{\alpha+c-1}\exp(-\lambda W^c)dW. \end{aligned}$$

By Lemma 3(a), the first integral is $\approx a_q \gamma^{q+\alpha-cq}(\lambda c)^{-q}\exp(-\lambda\gamma^c)$; the second is similarly $\approx -a_q(\gamma+\epsilon)^{q+\alpha-cq}(\lambda c)^{-q}\exp(-\lambda(\gamma+\epsilon)^c)$ which goes to zero at an exponentially faster rate than the first as long as $\epsilon\gamma^{c-1} \to \infty$ since $\exp(-\lambda(\gamma+\epsilon)^c) = \exp(-\lambda\gamma^c)\exp(-\lambda c\epsilon\gamma^{c-1}(1+o(\frac{\epsilon}{\gamma})))$. Using Assumption S and Lemma 3(a), the absolute value of the third integral can be bounded by $O(1)\epsilon^{q+1}\gamma^{\alpha}\exp(-\lambda\gamma^c)(1+o(\gamma^{-c}))$ (we apply a $(q+1)$th order Taylor expansion to $s(\cdot)$ around zero). If $\epsilon^{q+1}\gamma^{(c-1)q} \to 0$ this implies that the third integral goes to zero faster than the first. Finally, the fourth integral can be bounded by one where the function $1(\cdot)$ is substituted for $s(\cdot)$, thus by Lemma 3(a) it is bounded by $(\gamma+\epsilon)^{\alpha}\exp(-\lambda(\gamma+\epsilon)^c)(1+o(\gamma^{-c}))$ and (similarly to the second integral) goes to zero exponentially faster than the first one if $\epsilon\gamma^{c-1} \to \infty$. If the conditions are met, the first integral dominates $Es(W_i - \gamma)$, i.e.,

$$Es(W_i - \gamma) \approx a_q\gamma^{q+\alpha-cq}(\lambda c)^{-q}\exp(-\lambda\gamma^c). \qquad (B.16)$$

For $Es^2(W - \gamma)$, we get similarly four terms, the first of which $\approx a_q^2\binom{2q}{q}\gamma^{2q(1-c)+\alpha}\cdot(\lambda c)^{-2q}\exp(-\lambda\gamma^c)$; the second is the corresponding integral from $\gamma + \epsilon$ to infinity and requires $\varepsilon\gamma^{c-1} \to \infty$ to go to zero exponentially faster. The third one can analogously be bounded by

$$\begin{aligned} &\int_{\gamma}^{\gamma+\varepsilon}\left|s^2(W - \gamma) - \left(\frac{a_q}{q!}\right)^2(W - \gamma)^{2q}\right| p.d.f._W(W)dW \\ \le\ & O(1)\epsilon^{2q+1}\gamma^{\alpha}\exp(-\lambda\gamma^c)(1 + o(\gamma^{-c})) \end{aligned}$$

and thus needs $\epsilon^{2q+1}\gamma^{(c-1)2q} \to 0$. And finally, the fourth integral can be bounded again by replacing $s^2(\cdot)$ by $1(\cdot)$, which using Lemma 3(a) goes to zero exponentially faster than the first one if $²°^{c¡1}$! 1. For $² = °^{¡\frac{(c¡1)(2q+\{)}{2q+1}}$ with $0 < \{ < 1$ the first integral dominates $Es^2(W_i ¡ °)$, i.e.,

$$Es^2(W_i ¡ °) ¼ a_q^2 \mathbf{i}{}^{2q}_{q}\mathbf{¢} °^{2q+®¡2cq}(¸c)^{¡2q}\exp(¡¸°^c): \qquad (B.17)$$

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

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

$$E(!(W_i)1(W_i > °)) = ¸c \int_°^1 W^{\&+®+c¡1}\exp(¡¹W^d ¡ ¸W^c)(1 + o(°^{¡c})dW: \quad (B.18)$$

When d = c, this expectation can be obtained straightforwardly using the analysis of the asymptotic variance given above. When d 6= c, Lemma 3(b) gives

$$E(!(W_i)1(W_i > °)) ¼ \begin{cases} \frac{¸c}{¹d}°^{\&+®+c¡d}\exp(¡¹°^d ¡ ¸°^c)(1 + o(°^{¡c)}) & d > c \\ °^{\&+®}\exp(¡¹°^d ¡ ¸°^c)(1 + o(°^{¡c}) & d < c: \end{cases} \quad (B.19)$$

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

$$\begin{aligned} & E(!(W_i)s(W_i ¡ °)) \\ = & \tfrac{a_q}{q!}\int_°^1 (W ¡ °)^q ¸cW^{\&+®+c¡1}\exp(¡¹W^d ¡ ¸W^c)dW \qquad (B.20) \\ & ¡\tfrac{a_q}{q!}\int_{°+²}^1 (W ¡ °)^q ¸cW^{\&+®+c¡1}\exp(¡¹W^d ¡ ¸W^c)dW \\ & + \int_°^{°+²} \mathbf{h} s(W ¡ °) ¡ \tfrac{a_q}{q!}(W ¡ °)^q \mathbf{i} ¸cW^{\&+®+c¡1}\exp(¡¹W^d ¡ ¸W^c)dW \\ & + \int_{°+²}^1 s(W ¡ °)¸cW^{\&+®+c¡1}\exp(¡¹W^d ¡ ¸W^c)dW: \end{aligned}$$

In the case where c = d the result follows directly from Lemma 3(a). Using a similar discussion when c 6= d, Lemma 3(b) gives us

$$E(!(W_i)s(W_i ¡ °))$$

$$¼ \begin{cases} a_q \frac{¸c}{(¹d)^{q+1}} \exp(¡¸°^c ¡ ¹°^d)°^{q+\&+®+c¡d¡dq} & d > c \\ a_q(¸c)^{¡q} \exp(¡¸°^c ¡ ¹°^d)°^{q+\&+®¡cq} & d < c: \end{cases} \qquad (B.21)$$

This completes the derivations required for the proof. 2

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

$$MSE = an^{¡1}°^{¡®} \exp(¸°^c) + b°^{2´} \exp(¡2¹°^d); \qquad (B.22)$$

where

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

Thus

$$\frac{@MSE}{@°} = an^{¡1}°^{¡®¡1} \exp(¸°^c)(¡® + ¸c°^c) + b°^{2´¡1} \exp(¡2¹°^d)(2´ ¡ 2¹d°^d): \quad (B.24)$$

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:

$$\begin{array}{lll} \text{(i)} & ¡®a \exp(¸)n^{¡1}°^{¤¡®¡1} + 2´b \exp(¡2¹)°^{¤2´¡1} = 0 & \text{if } c = 0; d = 0\ (®; ´ < 0) \\ \text{(ii)} & ¸can^{¡1}°^{¡®+c¡1} \exp(¸°^c) + 2´b \exp(¡2¹)°^{2´¡1} = 0 & \text{if } c > 0; d = 0\ (´ < 0)) \\ \text{(iii)} & ¡®a \exp(¸)n^{¡1}°^{¡®¡1} ¡ 2¹db°^{2´+d¡1} \exp(¡2¹°^d) = 0 & \text{if } c = 0; d > 0\ (® < 0) \\ \text{(iv)} & ¸can^{¡1}°^{¡®+c¡1} \exp(¸°^c) ¡ 2¹db°^{2´+d¡1} \exp(¡2¹°^d) = 0 & \text{if } c > 0; d > 0: \end{array}$$

We discuss each case separately.

Case (i). In this case we get an analytic solution $°^¤ = (\frac{2´b \exp(¡2¹)}{®a \exp(¸)}n)^{\frac{1}{¡®¡2\&}}$ and $MSE(°^¤) = a \exp(¸)(\frac{2´b \exp(¡2¹)}{®a \exp(¸)})^{\frac{®}{®+2\&}} n^{¡\frac{2´}{®+2´}})(1 + \frac{b \exp(¡2¹)}{a \exp(¸)})$, so $¿ = \frac{2´}{®+2´}$: Here for any $° = µn^{\frac{1}{¡®¡2\&}}$, the corresponding rate equals $¿$.

Case (ii). Substituting from the …rst order conditions we can express $MSE(°^¤)$ as $¼ b°^{¤2´} \exp(¡2¹)(1 + \frac{2´}{¸c}°^{¤¡c}) ¼ b°^{¤2´} \exp(¡2¹)$. If $° = (µ \ln n)^{1=c}$ the …rst term in the derivative is proportional to $n^{¡1+¸µ}(\ln n)^{\frac{¡®+c¡1}{c}}$ while the second is proportional to $(\ln n)^{\frac{2´¡1}{c}}$; thus for $°_L = (µ_L \ln n)^{1=c}$ with $µ_L < \frac{1}{¸}$, the second negative term dominates the growth in the derivative (abias^2 dominates MSE($°_L$)) and ¿(L) = 0. For $°_H = (µ_H \ln n)^{1=c}$ $µ_H > \frac{1}{¸}$, the …rst term dominates the derivative (avar dominates MSE($°_H$)) and ¿(H) = 1 ¡ $¸µ_H$ (note: only if $µ = \frac{1}{¸}$ does the variance not increase with n). Since $°_L < °^¤ < °_H$, we get (a), (b), (c), and (d).

Case (iii). Here similarly, MSE($°^¤$) = $an^{¡1°¤¡®}\exp(¸)(1 ¡ \frac{®a}{2¹d}°^{¤¡d}) ¼ an^{¡1°¤¡®}\exp(¸)$. If $°_L = (µ_L \ln n)^{1=d}$ and $µ_L < \frac{1}{2¹}$ the derivative is negative, MSE($°_L$) is dominated by abias^2 and declines at a rate with polynomial component $n^{¡¿(L)}$ with ¿(L) = $2¹µ_L$. If $°_H = (µ_H \ln n)^{1=d}$ and $µ_H > \frac{1}{2¹}$ the derivative is positive, avar dominates and declines with ¿(H) = 1; (a), (b), (c), and (d) follow.

Case (iv). Note that substituting from the …rst order condition here MSE($°^¤$) ¼ $an^{¡1°¤¡®}\exp(¸°^{¤c})$ ¼ $¡2\frac{1d}{¸c}°^{¤2´+d¡c}\exp(¡2¹°^{¤d})$. If c < d, set $°_L = (µ_L \ln n)^{1=d}; 0 < µ_L < \frac{1}{2¹}$ and $°_H = (µ_H \ln n)^{1=c}; 0 < µ_H < \frac{1}{¸}$, then for MSE($°_L$) ¼ abias^2 we have $¿_L = 2¹µ_L$; MSE($°_H$) ¼ avar we have $¿_L = 1 ¡ ¸µ_H$. As $µ_L$ ($µ_H$) is selected close to $\frac{1}{2¹}$ (0) we approach ¿ = 1 and (a), (b), (c), and (d) follow. If c > d, $°_L = (µ_L \ln n)^{1=c}; 0 < µ_L < \frac{1}{¸}$ and $°_H = (µ_H \ln n)^{1=d}; 0 < µ_H < \frac{1}{2¹}$. We get ¿(L) = ¿(H) = 0: Since here the expression for MSE($°^¤$) for $°^¤ > °_L$ grows faster than $n^{¡1}\exp(¸µ_L \ln n) = n^{¡1+¸µ_L}$ for any $µ_L$, ¿ = 0. Thus (a), (b), (c) and (d) follow. For c = d, set $°_L = (µ_L \ln n)^{1=c}; 0 < µ_L < \frac{1}{¸+2¹}$ and $°_H = (µ_H \ln n)^{1=c}; \frac{1}{¸+2¹} < µ_H < \frac{1}{¸}$. In the expression for MSE($°^¤$) with $°^¤ > °_L$ we have ¿ < 1 ¡ $µ_L¸$ for all $0 < µ_L < \frac{1}{¸+2¹}$, therefore considering $µ_L$ arbitrarily close to $\frac{1}{¸+2¹}$ we get ¿ = $\frac{2¹}{¸+2¹}$: Thus (a), (b), (c), and (d) follow. If ¡® = 2´ an analytic solution $°^¤ = (\frac{1}{¸+2¹})^{1=c}(\ln n + \ln(\frac{2¹b}{¸a}))^{1=c}$ obtains.

This completes the proof. 2

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 Efficient 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 Coefficients,” Econometrica, 57, 1435–1460.

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

Schafgans, M.M.A. (1998a): “Ethnic Wage Differences 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.