Semiparametric Estimation of the Intercept of a Sample Selection Model

DONALD W. K. ANDREWS  
Yale University  
and  
MARCIA M. A. SCHAFGANS  
London School of Economics

First version received September 1995; final version accepted October 1997 (Eds.).

This paper provides a consistent and asymptotically normal estimator for the intercept of a semiparametrically estimated sample selection model. The estimator uses a decreasingly small fraction of all observations as the sample size goes to infinity, as in Heckman (1990). In the semiparametrics literature, estimation of the intercept has typically been subsumed in the nonparametric sample selection bias correction term. The estimation of the intercept, however, is important from an economic perspective. For instance, it permits one to determine the “wage gap” between unionized and nonunionized workers, decompose the wage differential between different socioeconomic groups (e.g. male–female and black–white), and evaluate the net benefits of a social programme.

1. INTRODUCTION

Semiparametric estimation of sample selection models has received considerable attention in the last decade. The reason is that parameter estimators of the sample selection model are inconsistent when incorrect distributional assumptions are made about the errors (e.g. see Goldberger (1983) and Arabmazar and Schmidt (1981, 1982)).

This paper considers semiparametric estimation of the intercept parameter, $\mu_0$, in a sample selection model. In the semiparametrics literature to date, the intercept has been absorbed in the nonparametric sample selection bias correction term. The only exceptions are the estimators of Gallant and Nychka (1987) and Heckman (1990). Gallant and Nychka's (1987) estimator of $\mu_0$ has been shown to be consistent, but its asymptotic distribution is unknown. The asymptotic distribution of Heckman's (1990) estimator also is unknown. The estimator we consider here is a slight variant of Heckman's estimator. We show that it is consistent and asymptotically normal.

The economic interpretation of an estimated sample selection model makes estimation of the intercept important. It is required for the evaluation of the “wage gap”, an issue that has received considerable attention in the literature (e.g. see Oaxaca (1973), Lewis (1986), Smith and Welch (1986), Wellington (1993), and Baker et al. (1993)), and for the evaluation of social programmes. Estimation of the intercept permits evaluation of the net benefit of a social programme by permitting comparisons of the actual outcome of participants with the expected outcome had they chosen not to participate.
The estimator, $\hat{\mu}_n$, that we consider uses a decreasingly small fraction of all observations as the sample size, $n$, goes to infinity. This approach is advocated by Heckman (1990). He suggests estimating $\mu_0$ 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. This approach works because the conditional mean of the outcome equation errors is close to zero for observations whose probability of selection is close to one. This is an example of Chamberlain’s (1986) “identification at infinity”.

Heckman suggests using all observations for which the estimated index of the participation equation, $X_i\hat{\beta}$, exceeds a certain threshold $\gamma_n$. We introduce a weighting scheme for these observations, where observations exceeding this threshold are weighted by a smooth monotone [0, 1]-valued function, $s(\cdot)$, such as a distribution function. The smoothness we impose on $s(\cdot)$, viz., differentiability of order three, is used to show that the asymptotic distribution of the estimator $\hat{\mu}_n$ is not affected by preliminary estimators such as $\hat{\beta}$. We conjecture that Heckman’s estimator also is asymptotically normal (and we show that it is in the unrealistic case where there are no preliminary estimators). But, our proof relies on smoothness when preliminary estimators are present, so we utilize the smoothing function $s(\cdot)$.

Our distribution theory for the estimator $\hat{\mu}_n$ assumes the existence of root-$n$ consistent semiparametric estimators of $(\theta_0, \beta_0)$, where $\theta_0$ is the vector of parameters of the outcome equation with the exclusion of the intercept and $\beta_0$ is the vector of parameters of the selection equation. Several such estimators are available in the literature.

The estimator $\hat{\mu}_n$ depends on the bandwidth parameter $\gamma_n$. For consistency, $\gamma_n$ is required to go to infinity as $n \to \infty$. The choice of $\gamma_n$ is constrained by the requirements that the variability and bias of $\hat{\mu}_n$ go to zero as $n \to \infty$. The first imposes an upper bound on how fast $\gamma_n$ can increase; the second a lower bound. The thinner the upper tail of the errors in the selection equation compared to the upper tail of the index $X_i\beta_0$, the greater is the latitude in the choice of $\gamma_n$. A formal method for determining an optimal bandwidth parameter $\gamma_n$ has yet to be developed.

The results show that up to cube-root-$n$ convergence of $\hat{\mu}_n$ can be achieved when the upper tail of the distribution of the regression function has the same tail thickness as the upper tail of the error distribution; up to square-root-$n$ convergence of $\hat{\mu}_n$ can be achieved when the upper tail of the distribution of the regression function is thicker than the upper tail of the error distribution; and asymptotically normality with asymptotic mean zero does not hold when the upper tail of the distribution of the regression function is less thick than the upper tail of the error distribution.

We note that the intercept is identified in our sample selection model by the assumption that the mean of the error in the main equation is zero. The same basic estimation strategy as considered here could be applied if one assumed instead that some quantile of the error, such as the median, is zero. Also, the same basic estimation strategy could be utilized in the case where the selection equation is estimated nonparametrically, as in Ahn and Powell (1993), rather than semiparametrically. Of course, the consistency and asymptotic normality results of this paper would not be directly applicable in either of these alternative scenarios.

The remainder of this paper is organized as follows: In Section 2, the sample selection model is discussed. The discussion of an application in this section motivates our interest in estimating $\mu_0$. In Section 3, our proposed estimator is defined. Consistency and asymptotic normality of this estimator are established in Section 4 under the assumption that the regressors and errors are independent. In Section 5, analogous results are established with the latter assumption relaxed. Section 6 concludes. An Appendix contains proofs of the results stated in Section 4.
2. THE SAMPLE SELECTION MODEL

The discussion of the econometric implications of sample selection started in the early seventies with the papers by Gronau (1974), Heckman (1974), and Lewis (1974). In their studies, the problem of sample selection bias is discussed in the context of the decision by women to participate in the labour force or not. The distribution of the wage offers sampled is truncated by the "self-selection" of women in the labour force, where women choose to be "in the sample" of workers if the offered wage exceeds their reservation wage. Sample selection models have been used in a wide variety of other applications, e.g., see Maddala (1983) and Amemiya (1984).

The sample selection model that we consider can be written as

\[ Y_i^* = \mu_0 + Z_i \theta_0 + U_i, \]
\[ D_i = 1(X_i \beta_0 > \epsilon_i), \]

and

\[ Y_i = Y_i^* D_i \quad \text{for } i = 1, \ldots, n, \]

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 selection equation. For convenience, we let

\[ W_i = X_i \beta_0. \]

To express the model given in (2.1) in terms of the Gronau–Heckman–Lewis model, we note that in their model \(Y_i^*\) is the latent offered wage and \(D_i\) is a dummy variable indicating whether an individual is employed, i.e. whether \(Y_i^* - Y_i^*\) exceeds zero, where \(Y_i^*\) denotes the individual's latent reservation wage. The observed wage is given by \(Y_i\). The variables influencing the decision to participate in the labour market are given by \(X_i\) and the determinants of the wage offer are given by \(Z_i\).

The standard approach to estimation of this model assumes that \((U_i, \epsilon_i)\) are bivariate normally distributed, independently of \((Z_i, X_i)\), with zero mean and unknown covariance matrix. With this assumption, the parameters can be estimated by maximum likelihood or the two-step estimator of Heckman (1976). This approach is known to yield inconsistent estimators if the normality assumption fails, e.g. see Arabmazar and Schmidt (1981, 1982) and Goldberger (1983).

Semiparametric estimation methods provide a way to overcome this deficiency. These methods consider the estimation of the parameters of interest without restricting the distribution function of the error terms or restricting the functional form of heteroskedasticity to lie in a finite-dimensional parametric family. Important progress on semiparametric estimation of selection models has been made by Gallant and Nychka (1987), Newey (1988), Robinson (1988), Powell (1989), Cosslett (1991), and Ichimura and Lee (1991). Also see Andrews (1991).

The point of departure in these papers is the conditional mean index function representation of the sample selection problem

\[ E(Y_i|D_i = 1, X_i, Z_i) = Z_i \theta_0 + E(U_i|D_i = 1, X_i) \]

and

\[ E(U_i|D_i = 1, X_i) = \kappa(X_i \beta_0), \]
where $X_i \beta_0$ is an index function and $\kappa(\cdot)$ is an unknown (smooth) function. The function $\kappa(\cdot)$ is sometimes called the sample selection correction function. It is proportional to the inverse Mill’s ratio when $(U_i, \varepsilon_i)$ are bivariate normal. In these models, conditional heteroskedasticity of $U_i$ is allowed, although only through the single index $X_i \beta_0$. The objective of these papers is to eliminate the contaminating effect of $E(U_i|D_i=1, X_i)$ in forming regression estimates of $\theta_0$.

As pointed out by Heckman (1990), all of the above papers except Gallant and Nychka (1987) absorb the intercept, $\mu_0$, into the definition of the conditional mean $E(U_i|D_i=1, X_i)$. None of these papers produces a consistent estimator of $\mu_0$. Gallant and Nychka, using a series expansion method, do obtain a consistent estimator of $\mu_0$. The distribution theory for their estimator, however, has not been developed. Gallant and Nychka assume that the errors and regressors are independent. They also impose a continuity condition on the distribution of the errors and regressors that is somewhat complicated and potentially difficult to verify.

The estimation of the intercept, $\mu_0$, has important economic relevance. For instance, it permits one to determine the "wage gap" between unionized and nonunionized workers (e.g., see Lewis (1986) and Heckman (1990)), decompose the wage differential between different socio-economic groups (e.g., male–female and black–white) following Oaxaca (1973), and evaluate the net benefits of a social programme (e.g., see Heckman and Robb (1985)). This is what prompted our interest in obtaining a consistent asymptotically normal estimator of this parameter.

To be more specific, consider the restrictive, but illustrative, model commonly employed to evaluate the benefit from a training programme

$$Y_i = \mu_0 + Z_i \theta + (\mu_1 - \mu_0) T_i + U_i,$$  

(2.4)

where $Y_i$ is the wage of individual $i$, $Z_i$ are wage determining variables, and $T_i$ is a dummy variable indicating whether this individual has received training. In terms of model (2.1) this can be written as

$$Y_{ik} = Z_i \theta_0 + \mu_k + U_{ik}, \quad k = 0 \text{ (non-trainee), } 1 \text{ (trainees)}$$

$$T_i = I(X_i > \varepsilon_i), \quad \text{trainee selection}$$

$$Y_i = T_i Y^*_{i0} + (1-T_i) Y^*_{i1}, \quad \text{for } i = 1, \ldots, n,$$  

(2.5)

where $Y^*_{i0}$ and $Y^*_{i1}$ denote the wages of individual $i$ with and without training, $Y_i$ is the observed wage of this individual, and $X_i$ are observed variables that affect the training status of individual $i$.

The effect of training on randomly assigned trainees in this case is given by $\mu_1 - \mu_0$. To estimate this experimental treatment effect, consistent estimates of the intercepts of the trainee wage equation $Y_{i1} = T_i Y^*_{i1}$ and non-trainee wage equation $Y_{i0} = (1-T_i) Y^*_{i0}$ are essential. Up to now, semiparametric estimation of the sample selection model only allowed the estimation of

$$E(Y_{i1}|T_i=1, Z_i, X_i) - E(Y_{i0}|T_i=0, Z_i, X_i) = \mu_1 - \mu_0 + E(U_{i1}|T_i=1, Z_i, X_i) - E(U_{i0}|T_i=0, Z_i, X_i),$$

(2.6)

which, in general, is not equal to $\mu_1 - \mu_0$. 
3. THE ESTIMATOR

The estimator we consider is

$$
\hat{\mu}_n = \frac{\sum_{i=1}^{n} (Y_i - Z_i \hat{\theta}) D_i s(X_i \hat{\beta} - \gamma_n)}{\sum_{i=1}^{n} D_i s(X_i \hat{\beta} - \gamma_n)},
$$

(3.1)

where \( s(\cdot) \) is a non-decreasing \([0, 1]\)-valued function that has three derivatives bounded over \( R \) and for which \( s(x) = 0 \) for \( x \leq 0 \) and \( s(x) = 1 \) for \( x \geq b \) for some \( 0 < b < \infty \). The preliminary estimators \((\hat{\theta}, \hat{\beta})\) are root-\( n \) consistent estimators of \((\theta_0, \beta_0)\). The parameter \( \gamma_n \) is called the bandwidth or smoothing parameter. This bandwidth parameter is chosen such that \( \gamma_n \to \infty \) as \( n \to \infty \).

The estimator suggested by Heckman (1990) is

$$
\hat{\mu}_n = \frac{\sum_{i=1}^{n} (Y_i - Z_i \hat{\theta}) D_i 1(X_i \hat{\beta} > \gamma_n)}{\sum_{i=1}^{n} D_i 1(X_i \hat{\beta} > \gamma_n)}.
$$

(3.2)

Comparing the two formulae (3.1) and (3.2), it is clear that the estimator \( \hat{\mu}_n \) differs from Heckman's (1990) \( \hat{\mu}_n \) only in that it replaces the indicator function \( 1(\cdot) \) with a smooth function \( s(\cdot) \). The introduction of this function facilitates the proof of the distribution theory.

Heckman's estimator \( \hat{\mu}_n \) is essentially a sample average of the random variables \( U_i + \mu_0 \) over a fraction of all observations, since \( Y_i - Z_i \hat{\theta} \to U_i + \mu_0 \) as \( n \to \infty \) for all \( i \geq 1 \). The effective sample size is equal to the number of observations used for the estimation of \( \mu_0 \). Since we introduce a weighting scheme for these observations, viz., the smooth function \( s(\cdot) \), our estimator \( \hat{\mu}_n \) is a weighted sample average of the random variables \( U_i + \mu_0 \), where observations with \( X_i \hat{\beta} \) greater than \( \gamma_n \) and with \( X_i \hat{\beta} \) close to the threshold \( \gamma_n \) are weighted less than those further away.

4. CONSISTENCY AND ASYMPTOTIC NORMALITY

The consistency and asymptotic normality of \( \hat{\mu}_n \) are established in Sections 4.1 and 4.2 respectively below. Section 4.3 addresses the estimation of the asymptotic covariance matrix of \((\hat{\mu}_n, \hat{\theta}, \hat{\beta})\).

4.1. Consistency

We now state the Assumptions 1-7 that are used to establish consistency of \( \hat{\mu}_n \). Each assumption is discussed below.

**Assumption 1.** \([(Z_i, X_i, U_i, e_i): i \geq 1]\) are i.i.d. random variables with \( E\|Z_i\|^{p} < \infty \), \( E\|X_i\|^{p} < \infty \), and \( E U_i^{p} < \infty \), for some \( p > 3 \).

**Assumption 2.** (a) \( E U_i = 0 \). (b) \( (U_i, e_i) \) is independent of \((Z_i, X_i)\).

**Assumption 3.** \( s(\cdot): R \to [0, 1] \) is a nondecreasing three times differentiable function with \( s(x) = 0 \) \( \forall x \leq 0 \), \( s(x) = 1 \) \( \forall x \geq b \) for some \( 0 < b < \infty \), and \( \sup_{x \in R} |s''(x)| < \infty \).

**Assumption 4.** \( P(W_i > w)^{1+\xi}/P(W_i > w + b) = O(1) \) as \( w \to \infty \) for some \( \xi \in [0, \frac{1}{3}] \) for \( b \) as in Assumption 3.
Assumption 5.

\[
\sqrt{n} \left( \hat{\theta} - \theta_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i + o_p(1)
\]

for some i.i.d. mean zero random variables \( \{Q_i : i \geq 1\} \) with \( \Omega = EQ_iQ_i' \) positive definite, \( E\|Q_i\|^\lambda < \infty \) for some \( \lambda > 2 \), and \( E\|U_iQ_i\|^p < \infty \) for some \( p > 3 \).

Assumption 6. \( \gamma_n \to \infty \).

Assumption 7. \( nP(W_i > \gamma_n + b) \to \infty \).

The first assumption imposes quite mild moment conditions on \((Z_i, X_i, U_i, e_i)\). It rules out unconditional heteroskedasticity and time series applications.

Assumption 2(a) identifies the parameter \( \mu_0 \). Assumption 2(b), independence of \((Z_i, X_i)\) and \((U_i, e_i)\), can be restrictive. Some semiparametric estimation techniques for the sample selection model have considered the less restrictive case in which the errors are allowed to depend on \( X_i \) through the index \( W_i = X_i/\sigma \). In Section 5, we relax Assumption 2 to incorporate this less restrictive case.

Assumption 3 is an assumption of smoothness of the function \( s(\cdot) \). An example of a function satisfying this condition is given by

\[
s(x) = \begin{cases} 
1 - \exp \left( -\frac{x}{b-x} \right), & \text{for } x \in (0, b); \\
0, & \text{for } x \leq 0; \\
1, & \text{for } x \geq b.
\end{cases}
\]

(4.1)

Assumption 4 requires that \( W_i \) has unbounded support from above. It rules out distributions of \( W_i \) that are too thin upper tailed. For example, suppose the upper tail of \( W_i \) decays as \( 1 - F(x) \sim \exp (-\exp (\lambda x)) \). Then, for \( \xi > 0 \), Assumption 4 is satisfied as long as \( \lambda \leq \ln (1 + \xi) / b \), where \( b \) and \( \xi \) are defined in Assumptions 3 and 4 respectively. The upper tail of \( W_i \) is too thin when \( \lambda \) exceeds this cutoff point. In this case, Assumption 4 is never satisfied for \( \xi = 0 \). If \( W_i \) has a Weibull \((\lambda, c)\) upper tail, i.e., \( 1 - F(x) \sim \exp (-\lambda x^c) \) for some \( c > 0 \) and \( \lambda > 0 \), then Assumption 4 is satisfied with \( \xi = 0 \) if \( c \leq 1 \). If \( c > 1 \), the assumption holds for all \( \xi > 0 \). (A special case is \( W_i \sim N(0, \sigma^2) \), in which case \( c = 2 \).) If the upper tail of the distribution of \( W_i \) is declining geometrically, i.e., \( W_i \) has a Pareto \((\lambda, c)\) upper tail \( 1 - F(x) \sim x^{-\lambda} \) for \( \lambda > 0 \), then Assumption 4 is satisfied for any \( \xi \geq 0 \).

Assumption 5 requires that the preliminary estimators satisfy an asymptotic linear expansion. Almost all estimators that are root-\(n\) consistent and asymptotically normal satisfy this assumption. In particular, the estimators suggested by Ichimura (1985), Han (1987) (also see Sherman (1993)), Newey (1988), Robinson (1988), Powell (1989), Powell, Stock and Stoker (1989), Ichimura and Lee (1991), Andrews (1991), and Klein and Spady (1993) satisfy this assumption. The estimator suggested by Cosslett (1991) has not been shown to satisfy this condition. To date, only a consistency result is available for his estimator. Note that many of the papers above provide an estimator for either \( \theta_0 \) or \( \beta_0 \) and, hence, to get an estimator of \((\theta_0, \beta_0)\) one may need to employ estimators from more than one paper.

1. Here and below, "\( a(x) \sim b(x) \)" is defined to mean that the ratios \( a(x)/b(x) \) and \( b(x)/a(x) \) are \( O(1) \) as \( x \to \infty \).
The bandwidth parameter, \( \gamma_n \), is required to go to infinity as the number of observations, \( n \), goes to infinity, by Assumption 6. This guarantees that the estimation of \( \mu_0 \) is based on only those values of \( X_i \) for which \( P(D_i = 1|X_i) \) is close to one and in the limit is equal to one.

A bound on the speed with which \( \gamma_n \) is allowed to go to infinity is given by Assumption 7. This assumption implies that the variability of \( \mu_n \) is asymptotically zero. The bound on the speed depends on the upper tail probabilities of \( W_i \). For example, if \( W_i \) has a Weibull \((\lambda_1, c_1)\) upper tail, i.e. \( 1 - F(x) \sim \exp (-\lambda_1 x^{c_1}) \) for some \( c_1 > 0 \) and \( \lambda_1 > 0 \), then Assumption 7 is satisfied when \( \gamma_n \leq ((1 - \tau_1)/\lambda_1 \log n)^{1/c_1} \) for arbitrary \( 0 < \tau_1 < 1 \).

Our consistency result is given by Theorem 1.

**Theorem 1.** Under Assumptions 1–7, \( \mu_n \approx \mu_0 \).

The proofs of Theorem 1 and other results in this section are given in the Appendix.

### 4.2. Asymptotic normality

Under Assumptions 1–7, the following assumption is necessary and sufficient for \( \mu_n \) to be asymptotically normal.

**Assumption 8.** \( \sqrt{n} EU_i D_i s(W_i - \gamma_n)/(ED_i s^2(W_i - \gamma_n))^{1/2} \to 0 \).

Assumption 8 ensures that the bias of \( \mu_n \) goes to zero asymptotically. A sufficient condition for Assumption 8 is

**Assumption 8*.** \( \sqrt{n} E|U_i|1(\varepsilon_i > \gamma_n) \cdot P(W_i > \gamma_n)^{(1-\xi)/2} \to 0 \) for \( \xi \in [0, \frac{1}{2}] \) as in Assumption 4.

**Lemma 1.** Under Assumptions 1–4 and 6, Assumption 8* implies Assumption 8.

Both Assumptions 7 and 8* are related to the choice of \( \gamma_n \). The former implies an upper bound as described above; the latter implies a lower bound. A sufficient condition for the latter assumption is that \( \varepsilon_i \) is a bounded random variable, since then \( 1(\varepsilon_i > \gamma_n) \to 0 \) a.s. by Assumption 6. Alternatively, if \( \varepsilon = 0 \) and \( U_i \) is bounded, a sufficient condition is that \( n^{1/2}P(\varepsilon_i > \gamma_n)P(W_i > \gamma_n)^{1/2} \to 0 \).

For the case where \( U_i \) and \( \varepsilon_i \) are unbounded, a useful inequality that separates the randomness of \( U_i \) from that of \( \varepsilon_i \) in Assumption 8* is

\[
E|U_i|1(\varepsilon_i > \gamma_n) \leq E|U_i|1(|U_i| > \gamma_n) + \gamma_n P(\varepsilon_i > \gamma_n).
\]

Using this inequality we obtain the following results. Suppose \( W_i \) has Weibull \((\lambda_1, c_1)\) upper tail and \( U_i \) and \( \varepsilon_i \) both have Weibull \((\lambda_2, c_2)\) upper tail for some positive constants \( \lambda_1, c_1, \lambda_2, c_2 \). If \( c_1 < c_2 \), so that \( U_i \) and \( \varepsilon_i \) have thinner upper tails than \( W_i \), and \( \gamma_n = (K \log n)^{1/c}(1 + o(1)) \) for some \( c < 1, c_1, c_2 \) and any \( 0 < K < \infty \), then Assumptions 7 and 8* both hold. If \( c_1 = c_2 \) and \( \gamma_n = (K \log n)^{1/c}(1 + o(1)) \) for some \( K \in (1/(2\lambda_2 + \lambda_1), 1/\lambda_1) \), then Assumptions 7 and 8* hold. If \( c_1 > c_2 \), then Assumptions 7 and 8* (or 7 and 8) are not mutually compatible.

Asymptotic normality of \( \mu_n \) is established in the following theorem. Let \( \sigma^2 = \text{Var} (U_i) \).
Theorem 2. Under Assumptions 1-7,

\[ \frac{\sqrt{n}ED_1s(W_i - \gamma_n)}{\sigma(ED_1s^2(W_i - \gamma_n))^{1/2}} \left( \hat{\mu}_n - \mu_0 - \frac{EU_1s(W_i - \gamma_n)}{ED_1s(W_i - \gamma_n)} \right) \xrightarrow{d} N(0, 1) \]

and

\[ \frac{\sqrt{n}ED_1s(W_i - \gamma_n)}{\sigma(ED_1s^2(W_i - \gamma_n))^{1/2}} (\hat{\mu}_n - \mu_0) \xrightarrow{d} N(0, 1) \] if Assumption 8 holds.

The rate of convergence of \( \hat{\mu}_n \) is given by Theorem 2(b) to be

\[ \frac{\sqrt{n}ED_1s(W_i - \gamma_n)}{(ED_1s^2(W_i - \gamma_n))^{1/2}} \sim \frac{\sqrt{n}Es(W_i - \gamma_n)}{(Es^2(W_i - \gamma_n))^{1/2}}, \]

using Lemma A-2 of the Appendix. Under Assumption 4, the rate of convergence is at least \( n^{(1+2\epsilon)/(2+2\epsilon)} \). For example, if \( W_i \) has Weibull \((\lambda_1, c_1)\) upper tail, \( U_i \) and \( \varepsilon_i \) both have Weibull \((\lambda_2, c_2)\) upper tail, and Assumptions 1-3 and 5 hold, then (i) \( n^{(\lambda_1 - \lambda_0)} < \frac{\epsilon}{2} \) when \( c_1 < c_2 \) and \( \gamma_n = (K \log n)^{1/c} (1 + o(1)) \) for some \( c \in (c_1, c_2) \) and any \( 0 < K < \infty \) and (ii) \( n^{(\lambda_1 - \lambda_0)} \geq \frac{\epsilon}{2} \) when \( c_1 = c_2 \) and \( \gamma_n = (K \log n)^{1/c_1} (1 + o(1)) \) for \( K = (1 + \epsilon)/(2\lambda_2 + 2\lambda_1) \) and \( \epsilon \in (0, 2\lambda_2/\lambda_1) \). In case (ii), if \( \lambda_1 = \lambda_2 \), so that \( W_i, U_i, \) and \( \varepsilon_i \) all have Weibull \((\lambda_1, c_1)\) upper tails, the bound on \( \epsilon \) is \( \frac{\epsilon}{2} - \frac{1}{6} \). For \( \epsilon > 0 \) arbitrarily small, this yields a rate of convergence arbitrarily close to \( n^{1/3} \). Also in case (ii), the rate of convergence is arbitrarily close to \( n^{1/2} \) for \( \lambda_2 \) sufficiently large given any \( \lambda_1 > 0 \) and \( \epsilon_1 > 0 \).

To test hypotheses and construct confidence intervals for functions of \((\mu_0, \theta_0, \beta_0)\), we need a joint asymptotic normality result for \((\hat{\mu}_n, \hat{\theta}, \hat{\beta})\). For example, this result is needed to calculate a confidence region for \( \hat{Y}_i \). Our joint normality result establishes the block diagonality of the asymptotic covariance matrix between \( \hat{\mu}_n \) and \((\hat{\theta}, \hat{\beta})\). Intuitively this block diagonality makes sense, because \( \hat{\mu}_n \) is estimated using a decreasingly small fraction of all observations and estimators of \((\theta_0, \beta_0)\) that leave out these observations are asymptotically equivalent to \((\hat{\theta}, \hat{\beta})\).

The following asymptotic joint normality result holds.

Theorem 3. Under Assumptions 1-7,

\[ \begin{pmatrix} \frac{\sqrt{n}ED_1s(W_i - \gamma_n)}{\sigma(ED_1s^2(W_i - \gamma_n))^{1/2}} (\hat{\mu}_n - \mu_0) \\ \sqrt{n} \Omega^{-1/2} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} \end{pmatrix} \xrightarrow{d} N(0, I) \]

iff Assumption 8 holds.

4.3. Asymptotic normality with estimated variance

To compute standard errors and test statistics one needs an estimator of the asymptotic covariance matrix of \((\hat{\mu}_n, \hat{\theta}, \hat{\beta})\). More specifically, one needs consistent estimators of \( \sigma^2 \), the asymptotic covariance matrix of \((\hat{\theta}, \hat{\beta})\), and the normalizing factors \( ED_1s(W_i - \gamma_n) \) and \( ED_1s^2(W_i - \gamma_n) \).
Define

\[ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \mu_n - Z_i \hat{\theta})^2 D_i s(X_i \hat{\beta} - \gamma_n)}{\sum_{i=1}^{n} D_i s(X_i \hat{\beta} - \gamma_n)}. \]  

(4.2)

Consistency of \( \hat{\sigma}^2 \) is established in the following theorem.

**Theorem 4.** Under Assumptions 1-7, \( \hat{\sigma}^2 \xrightarrow{p} \sigma^2 \).

For a consistent estimator of \( \Omega \), we use the results given in the literature for the particular choice of preliminary estimators \( (\hat{\theta}, \hat{\beta}) \). Assumption 9 simply formalizes the existence of such an estimator.

**Assumption 9.** \( \hat{\Omega} \xrightarrow{p} \Omega \).

The normalizing factors \( ED_i s(W_i - \gamma_n) \) and \( ED_i s^2(W_i - \gamma_n) \) can be estimated by their sample analogues \( (1/n) \sum_{i=1}^{n} D_i s(W_i - \gamma_n) \) and \( (1/n) \sum_{i=1}^{n} D_i s^2(W_i - \gamma_n) \) respectively.

The joint asymptotic normality of \( (\hat{\mu_n}, \hat{\theta}, \hat{\beta}) \) with estimated covariance matrix is given in the following theorem.

**Theorem 5.** Under Assumptions 1-7 and 9,

\[
\begin{pmatrix}
(1/\sqrt{n}) \sum_{i=1}^{n} D_i s(X_i \hat{\beta} - \gamma_n) \\
\hat{\sigma}(1/n) \sum_{i=1}^{n} D_i s^2(X_i \hat{\beta} - \gamma_n) \\
\sqrt{n} \hat{\Omega}^{-1/2} (\hat{\theta} - \theta_0) \\
\hat{\beta} - \beta_0
\end{pmatrix} \xrightarrow{d} N(0, I)
\]

iff Assumption 8 holds.

### 5. Dependence between Errors and Regressors

In the previous section, we established consistency and asymptotic normality of \( \hat{\mu_n} \) under the assumption of independence of \( (U_i, \epsilon_i) \) and \( (Z_i, X_i) \). This assumption can be restrictive. Some semiparametric estimators of \( (\theta_0, \beta_0) \) allow for conditional heteroskedasticity of the errors \( (U_i, \epsilon_i) \), although only in a restricted form. (They require that the distribution of the errors depends on \( X_i \) only through the single index \( X_i \beta_0 \).) Here we extend the results of the previous section to incorporate conditional heteroskedasticity. The proofs of the results of this section are given in Andrews and Schafgans (1996).

First we state revisions of Assumptions 1 and 2 that are used to establish consistency. Let \( W \) be the support of \( W_i (=X_i \beta_0) \).

**Assumption 1*.** \( (Z_i, X_i, U_i, \epsilon_i) \) are i.i.d. random variables with \( E\|Z_i\|^p < \infty \), \( E\|X_i\|^p < \infty \), \( E\|U_i^2 X_i\|^p < \infty \), and \( \sup_{w} E(|U_i|^4 | W_i = w) < \infty \) for some \( p > 3 \) and \( \lambda > 4 \).
Assumption 2*. (a) \( E(U_i | W_i) = 0 \) a.s.

(b) \( \sup_{w \in \mathcal{W}} P(\varepsilon_i \geq \gamma | W_i = w) \to 0 \) as \( \gamma \to \infty \).

The consistency result is given by

**Theorem 1**. Under Assumptions 1*, 2*, and 3–7, \( \hat{\mu}_n \to \mu_0 \).

To obtain asymptotic normality of \( \hat{\beta}_n \) with dependence between the errors and regressors, we need to add a new assumption and revise the sufficient condition Assumption 8* for Assumption 8.

**Assumption 10**. \( \inf_{w \in \mathcal{W}} \mathbb{E}(U_i^2 | W_i = w) > 0 \).

**Assumption 8**. \( \sqrt{n} \sup_{w \in \mathcal{W}} \mathbb{E}(| U_i | 1(\varepsilon_i > \gamma_n) | W_i = w) P(W_i > \gamma_n)^{(1 - \xi)/2} \to 0 \) for \( \xi \in [0, \frac{1}{2}] \) as in Assumption 4.

Assumption 10 ensures that \( \text{Var}(U_i D_i s(W_i - \gamma_n)) \) is positive.

An analogue of Lemma 1 is:

**Lemma 1**. Under Assumptions 1*, 2*, 3, 4, 6, and 10, Assumption 8** implies Assumption 8.

To compute standard errors and test statistics under the assumed dependence of the errors and regressors, one needs a consistent estimator of \( \text{Var}(U_i D_i s(W_i - \gamma_n)) \). Define

\[
\hat{V}_n = \frac{\sum_{i=1}^n (Y_i - \hat{\mu}_n - Z_i \hat{\theta}) D_i s^2(X_i \hat{\beta} - \gamma_n)}{\sum_{i=1}^n D_i s^2(X_i \hat{\beta} - \gamma_n)}.
\]

(5.1)

\( \hat{V}_n \) is a consistent estimator of \( \text{Var}(U_i D_i s(W_i - \gamma_n))/E(D_i s^2(W_i - \gamma_n)) \). Consistency of \( \hat{V}_n \) is established in the following analogue of Theorem 4.

**Theorem 4**. Under Assumptions 1*, 2*, 3–7, and 10, \( \hat{V}_n E(D_i s^2(W_i - \gamma_n))/\text{Var}(U_i D_i s(W_i - \gamma_n)) \xrightarrow{p} 1 \).

The joint asymptotic normality of \( (\hat{\mu}_n, \hat{\theta}, \hat{\beta}) \) with estimated covariance matrix is given by the following analogue of Theorem 5.

**Theorem 5**. Under Assumptions 1*, 2*, 3–7, 9, and 10,

\[
\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n D_i s(X_i \hat{\beta} - \gamma_n) \right)^{1/2} (\hat{\mu}_n - \mu_0) \xrightarrow{d} N(0, I)
\]

\[
\frac{\hat{V}_n^{1/2} \left( \sum_{i=1}^n D_i s^2(X_i \hat{\beta} - \gamma_n) \right)^{1/2} (\hat{\theta} - \theta_0)}{\sqrt{n} \Omega^{-1/2} (\hat{\beta} - \beta_0)} \xrightarrow{d} N(0, I)
\]

iff Assumption 8 holds.
6. CONCLUSION

In this paper we provide a consistent asymptotically normal estimator for the intercept of a semiparametrically estimated sample selection model. This parameter is of importance in determining a variety of economically interesting quantities as discussed in the second section. The estimation of this intercept has up to now been absorbed in the nonparametric sample selectivity bias correction term, with the exception of the estimator given by Gallant and Nychka. Their estimator is consistent, but its asymptotic distribution is unknown. Therefore, this paper provides a useful contribution to the literature on semiparametric sample selection models. We also present a consistent estimator of the asymptotic covariance matrix for \((\hat{\mu}_n, \hat{\theta}, \hat{\beta})\).

In Schafgans (1997a, b), the feasibility of our estimator is illustrated using various simulations and an empirical application. The simulations in Schafgans (1997a) reveal that the semiparametric estimator performs better than the two-step Heckman estimator for a range of bandwidth parameter choices for some distributions of the errors and regressors. For error distributions that are close to the normal, however, the two-step parametric estimator performs better. The empirical application in Schafgans (1997b) concerns the estimation of wages in Peninsular Malaysia. The results reveal that the parameter estimates using the semiparametric estimator are, with one exception, fairly similar to those obtained using various parametric estimation techniques. As expected, the standard errors of the semiparametric estimates are larger when a smaller proportion of the uncensored observations are used.

APPENDIX

For notational simplicity, we let \( s_n \) and \( s_n^* \) abbreviate \( s(W_i - \gamma_n) \) and \( s(X_i; \hat{\beta} - \gamma_n) \), respectively, in the proofs below.

The boundedness of the first three derivatives of \( s(\cdot) \) is not needed in all of the proofs. To indicate where it is needed, the following weaker alternative, Assumption 3', is used wherever possible. In particular, Assumption 3' is used when \((\theta_0, \beta_0)\) are assumed to be known.

Assumption 3'. \( s(\cdot): R \rightarrow [0, 1] \) is a distribution function of a random variable with support contained in \([0, b]\) for some \(0 \leq b < \infty\).

Note that Assumption 3' allows for \( s(x) = 1(x > 0) \), which generates Heckman's estimator \( \hat{\mu}_n \).

The proofs below use the following lemmas:

**Lemma A-1.** Under Assumptions 1, 2(b), 3', 6 and 7,

\[
\frac{(1/n) \sum D_i s(W_i - \gamma_n)}{ED_i s(W_i - \gamma_n)} \xrightarrow{D} 1.
\]

**Lemma A-2.** Under Assumptions 2(b), 3', and 6,

\[
\frac{ED_i s(W_i - \gamma_n)}{Es(W_i - \gamma_n)} \rightarrow 1, \quad \frac{ED_i s^2(W_i - \gamma_n)}{Es^2(W_i - \gamma_n)} \rightarrow 1, \quad \text{and} \quad \frac{ED_i s^3(W_i - \gamma_n)}{Es^3(W_i - \gamma_n)} \rightarrow 1.
\]

**Lemma A-3.** Under Assumptions 1, 2(b), 3', and 6,

\[
\frac{EU_i^2 D_i s^2(W_i - \gamma_n)}{EU_i^2 s^2(W_i - \gamma_n)} \rightarrow 1.
\]
If Assumption 4 also holds,

\[ \frac{\text{Var} (U_i D_i s(W_i - \gamma_s))}{\sigma^2 ED_i s^2 (W_i - \gamma_s)} \to 1. \]

**Lemma A-4.** Under Assumptions 1, 2(b), 3', 6, and 7,

\[ \frac{(1/n) \sum D_i s(W_i - \gamma_s)}{ED_i s^2 (W_i - \gamma_s)} \to 1. \]

**Proof of Lemma A-1.** The random variable \((1/n) \sum (D_i s_m - ED_i s_m)/ED_i s_m\) has mean zero and variance

\[ \frac{\text{Var} (D_i s_m)}{n(ED_i s_m)^2} \to \frac{1}{n}. \]  
(A.1)

Assumptions 1, 2(b), 3', 6, and 7 imply that

\[ \sqrt{n ED_i s_m} \to \infty, \]  
(A.2)

because

\[ \frac{\sqrt{n ED_i s_m}}{(ED_i s_m)^{1/2}} \to (n ED_i s_m)^{1/2} \]

\[ = (1 + o(1))(nED_i s_m)^{1/2} \]

\[ \geq (1 + o(1))(nP(W_i > \gamma_s + b))^{1/2}, \]  
(A.3)

where the equality uses Lemma A-2. Equations (A.1) and (A.2) give the desired result. \(\Box\)

**Proof of Lemma A-2.** We have

\[ (1 - D_i)s_m = 1(\gamma_s \leq W_i \leq \epsilon_i)s_m \leq 1(\gamma_s \leq \epsilon_i)s_m, \]  
(A.4)

using the fact that \(s_m \geq 0\) only if \(W_i - \gamma_s \geq 0\) by Assumption 3'. Thus,

\[ \frac{ED_i s_m}{E_{s_m}} - \frac{E (1 - D_i)s_m}{E_{s_m}} \leq \frac{E (1(\gamma_s \leq \epsilon_i)s_m}{E_{s_m}} = P(\gamma_s \leq \epsilon_i) \to 0, \]  
(A.5)

using Assumptions 2(b) and 6. The proof is identical with \(s_m\) replaced by \(s_i^2\) or \(s_i^*\). \(\Box\)

**Proof of Lemma A-3.** When \(s_m\) in (A.5) is replaced by \(U_i^* z_m^*\) the right-hand side becomes \(EU_i^* 1(\gamma_s \leq \epsilon_i) \to 0\) using Assumption 1.

Next, we have

\[ \frac{\text{Var} (U_i D_i s_m)}{\sigma^2 ED_i s_m^2} = \frac{EU_i^2 D_i s_m^2}{\sigma^2 ED_i s_m^2} = \frac{(EU_i D_i s_m^2)}{\sigma^2 ED_i s_m^2}, \]

\[ \frac{EU_i D_i s_m^2}{\sigma^2 ED_i s_m^2} = \frac{EU_i^2 s_m^2}{\sigma^2 s_m^2} (1 + o(1)) = 1 + o(1), \]  
(A.6)

and

\[ \frac{(EU_i D_i s_m^2)}{\sigma^2 ED_i s_m^2} \leq \frac{(EU_i^2 s_m^2)}{\sigma^2 s_m^2} (1 + o(1)) \leq \frac{(EU_i^2 s_m^2)}{\sigma^2 P(W_i > \gamma_s + b) (1 + o(1)) = o(1), \]

where the second and third equations use the results of the first part of this lemma, Lemma A-2, and Assumption 2(b) and the third equation also uses Assumptions 1, 4, and 6. \(\Box\)

**Proof of Lemma A-4.** The random variable \((1/n) \sum (D_i s_m^2 - ED_i s_m^2)/ED_i s_m^2\) has zero mean and variance

\[ \frac{\text{Var} (D_i s_m^2)}{n(ED_i s_m^2)^2} = \frac{ED_i s_m^2}{n(ED_i s_m^2)^2} \to 0, \]  
(A.7)
using Assumptions 3', 6, and 7 and Lemma A-2 since,
\[
\frac{ED_i \delta_i}{n(ED_i \delta_i)^2} \leq \frac{1}{nED_i \delta_i} = \frac{1}{nEs_{\alpha} + o(1)}
\]
\[
\leq \frac{1}{nP(W_i > \gamma_n + b)(1 + o(1))} = o(1).
\]

(A.8)

Proof of Lemma 1. By Assumption 2, \(EU_i s_{\alpha} = 0\). Hence, the absolute value of the left-hand side of Assumption 8 equals
\[
|\sqrt{nEU_i S_i/(ED_i \delta_i)^{1/2}}| \\
\leq (1 + o(1)) \sqrt{nE[|S_i|/(W_i < \varepsilon_i)](W_i > \gamma_n)/Es_{\alpha}^{1/2}} \\
\leq (1 + o(1)) \sqrt{nE[|S_i|/(W_i > \gamma_n) \cdot P(W_i > \gamma_n + b)/P(W_i > \gamma_n + b)_{1/2}} \\
= (1 + o(1)) \sqrt{nE[|S_i|/(W_i > \gamma_n) \cdot P(W_i > \gamma_n)^{1 - \varepsilon/2}(P(W_i > \gamma_n + b)_{1/2}} \\
\to 0.
\]

(A.9)

where the first inequality uses Lemma A-2, the second inequality uses Assumptions 2(b) and 3', and the convergence to zero uses Assumptions 4, 6, and 8.

Theorems 1, 2 and 3 follow from Theorems A-1 and A-2, A-1 and A-3, and A-1 and A-4, respectively, below. Define
\[
\hat{\mu}_{a0} = \sum_i U_i D_i s_{\alpha} / \sum_i D_i s_{\alpha}.
\]

(A.10)

Theorem A-1. Under Assumptions 1, 2(b), and 3, -7,
\[
\frac{\sqrt{nED_i s_i (W_i - \gamma_n)}}{(ED_i \delta_i^2 (W_i - \gamma_n)^{1/2}} (\hat{\mu}_a - \mu_a) \sim 0.
\]

Theorem A-2. Under Assumptions 1, 2, 3', 6, and 7, \(\hat{\mu}_{a0} \sim \mu_0\).

Theorem A-3. Under Assumptions 1, 2, 3', 4, 6, and 7, the results of parts (a) and (b) of Theorem 2 hold with \(\hat{\mu}_a\) replaced by \(\hat{\mu}_{a0}\).

Theorem A-4. Under Assumptions 1, 2, 3', and 4-7, the result of Theorem 3 holds with \(\hat{\mu}_a\) replaced by \(\hat{\mu}_{a0}\).

Proof of Theorem A-1. The left-hand side in the Theorem can be written as
\[
C\left(\frac{\hat{A}}{\hat{B}}\right) = \frac{\hat{A} - A}{B} - \frac{\hat{B} - B}{B} \frac{\hat{B}}{B}
\]

(A.11)

where
\[
C = \frac{\sqrt{nED_i s_{\alpha}}}{(ED_i \delta_i^2)^{1/2}} , \quad A = \sum_i (Y_i - Z_i' \theta_0) D_i s_{\alpha}, \quad \hat{A} = \sum_i (Y_i - Z_i' \hat{\theta}) D_i s_{\alpha}, \quad B = \sum_i D_i s_{\alpha}
\]

To show that the left-hand side converges in probability to zero, it suffices to show (i) \(\hat{B}/B \sim 1\), (ii) \(C(\hat{A} - A)/B \sim 0\), (iii) \(A/B = O_p(1)\), and (iv) \(C(B - B)/B \sim 0\). Since \(C \to \infty\) by Assumption 7 and (A.2), condition (i) follows from (iv). Using Lemma A-1, we find that the following two conditions (established below) are sufficient for (ii):
\[
(1/\sqrt{n}) \sum_i (U_i + \mu_0) D_i (s_{\alpha} - s_{\alpha})/(ED_i \delta_i^2)^{1/2} \sim 0,
\]

(A.12)

and
\[
(\hat{\theta} - \theta_0)' (1/\sqrt{n}) \sum_i Z_i D_i s_{\alpha} / (ED_i \delta_i^2)^{1/2} \sim 0.
\]

(A.13)

By Lemmas A-1 and A-2, the following are sufficient for (iii) and (iv) respectively:
\[
(1/n) \sum_i U_i D_i s_{\alpha} / Es_{\alpha} = O_p(1),
\]

(A.14)
and
\[ \left( \frac{1}{\sqrt{n}} \right) \sum_{i} D_i (\hat{s}_{ni} - s_{ni}) / (E_{ni}^2)^{1/2} \to 0. \] \( \text{(A.15)} \)

We establish (A.14) first. We have
\[ \left( \frac{1}{n} \right) \sum_{i} U_i D_i s_{ni} \leq \left( \frac{1}{n} \right) \sum_{i} \left( \frac{1}{n} \right) \sum_{i} (U_i D_i s_{ni} - EU_{i} D_{i} s_{ni}) \right| + \frac{E[|U_{i}|]}{E_{ni}} s_{ni}. \] \( \text{(A.16)} \)

The first term on the right-hand side is \( O_p(1) \), because it has mean zero and variance \( \sigma^2 ED_{i} s_{ni}^2 / (n ED_{i} s_{ni})^2 = o(1) \) using Lemmas A-2 and A-3, Assumption 7, and (A.2). The second term on the right-hand side equals \( E[|U_{i}|] \) using Assumption 2(b). Using Assumption 1, therefore, (A.14) holds.

Next, we establish (A.15). For notational simplicity, suppose \( \hat{\beta} = 0 \) and \( \beta_0 \) are scalars. The argument carries over to the vector case. Let \( G_n \) denote the left-hand side of (A.15). A three-term Taylor expansion of \( s(X_{i} \hat{\beta} - \gamma_{n}) \) about \( \beta_0 \) gives
\[ |G_n| \leq \left( \frac{1}{n} \right) \sum_{i} X_{i} D_{i} s'(X_{i} \beta_{0} - \gamma_{n}) \sqrt{n}(\hat{\beta} - \beta_0) \big/ (E_{ni}^2)^{1/2} \right| + \left( \frac{1}{2n} \right) \sum_{i} X_{i}^2 D_{i} s''(X_{i} \beta_{0} - \gamma_{n}) \sqrt{n}(\hat{\beta} - \beta_0)^2 \big/ (n E_{ni}^2)^{1/2} \big| \right| + \left( \frac{1}{6n} \right) \sum_{i} X_{i}^3 D_{i} s'''(X_{i} \beta_{0} - \gamma_{n}) \sqrt{n}(\hat{\beta} - \beta_0)^3 \big/ (n^2 E_{ni}^2)^{1/2} \big| \right| \leq O_p(1) \left( \frac{1}{n} \right) \sum_{i} |X_{i}| \big/ (E_{ni}^2)^{1/2} \big| + O_p(1) \left( \frac{1}{n} \right) \sum_{i} X_{i}^2 \big/ (n E_{ni}^2)^{1/2} + O_p(1) \left( \frac{1}{n} \right) \sum_{i} |X_{i}| \big/ (n^2 E_{ni}^2)^{1/2}, \] \( \text{(A.17)} \)

where \( s'(\cdot) \) denotes the first derivative of \( s(\cdot) \), etc., \( \beta_0 \) is on the line segment joining \( \hat{\beta} \) and \( \beta_0 \), and the second inequality uses Assumption 3, Assumption 5 (which implies that \( n(\hat{\beta} - \beta_0) = O_p(1) \)), and the inequalities
\[ |s'(x)| \leq 1(x > 0) \sup_{x \in \mathbb{R}} |s'(x)|, \] \[ |s''(x)| \leq 1(x > 0) \sup_{x \in \mathbb{R}} |s''(x)|, \] and \( \sup_{x \in \mathbb{R}} |s'''(x)| < \infty. \)

For any identically distributed random variables \( \{ H_i : i \geq 1 \} \) with \( 0 < E[|H_i|] < \infty, \) \( \left( \frac{1}{n} \right) \sum_{i} H_i / E[|H_i|] = O_p(1) \) by Markov's inequality. Thus, the first term on the right-hand side of (A.17) is
\[ O_p(1) \left( \frac{E[|X_{i}|]}{E_{ni}^2} \right)^{1/2} \leq O_p(1) \left( \frac{E[|X_{i}|]}{E_{ni}^2} \right)^{1/2} P(W_{i} > \gamma_{n})^{1/2} = o_p(1), \] \( \text{(A.18)} \)

where \( 1/p + 1/q = 1, p \) is as in Assumption 1, and the equality holds by Assumptions 4 and 6 since \( 2/q > 1 + \xi. \)

The second term on the right-hand side of (A.17) is
\[ O_p(1) \left( \frac{EX_{i}^2 1(W_{i} > \gamma_{n})}{(n E_{ni}^2)^{1/2}} \right) \leq O_p(1) \left( \frac{EX_{i}^3 1(W_{i} > \gamma_{n})}{(n E_{ni}^2)^{1/2}} \right) P(W_{i} > \gamma_{n} + b)^{1/2} = o_p(1), \] \( \text{(A.19)} \)

where the equality holds by Lemma A-2, Assumptions 4, 6 and 7, and (A.2).

The third term on the right-hand side of (A.17) is
\[ O_p(1) \left( \frac{EX_{i}^2}{(n E_{ni}^2)^{1/2}} \right) = O_p(1) \left( \frac{n E_{ni}}{E_{ni}^2} \right)^{1/2} = o_p(1). \] \( \text{(A.20)} \)

The first equality uses Assumption 1 and the last equality uses Lemma A-2, Assumptions 4, 6, and 7, and (A.2) and requires that \( \xi \leq 1 \) in Assumption 4. This completes the proof of (A.15).

The proof of (A.12) is the same as that of (A.15) except that the factor \( |U_{i} + \mu_0| \) appears in various sums and expectations. In particular, using Assumption 2(b), each expression in (A.18)-(A.19) is multiplied by \( E[|U_{i}| + \mu_0|]. \)
We now establish (A.13). By Assumption 5 (which implies that \( \sqrt{n}(\hat{\beta} - \beta_0) = O_p(1) \)) and a two-term Taylor expansion about \( \beta_0 \), the absolute value of the left-hand side of (A.13) is bounded by

\[
O_p(1) \left( \frac{1}{n} \sum \|Z_i\|s_{ni} \right)^{1/2} \leq O_p(1) \left( \frac{1}{n} \sum \|Z_i\|s_{ni} \right)^{1/2} + O_p(1) \left( \frac{1}{n} \sum \|Z_i\| \cdot |X_i| \cdot |s'(X_i, \beta_0 - \gamma_0)| \cdot \sqrt{n}(\hat{\beta} - \beta_0) \right) \\
+ O_p(1) \left( \frac{1}{n} \sum \|Z_i\|X_i^2 \cdot |s'(X_i, \beta_0 - \gamma_0)| \cdot \sqrt{n}(\hat{\beta} - \beta_0) \right)^{1/2} \\
\leq O_p(1) \left( \frac{1}{n} \sum \|Z_i\|X_i^2 \right)^{1/2} + O_p(1) \left( \frac{1}{n} \sum \|Z_i\| \cdot |X_i| \cdot (W_i > \gamma_0) \right)^{1/2} \\
+ O_p(1) \left( \frac{1}{n} \sum \|Z_i\|X_i^2 \right)^{1/2},
\tag{A.21}
\]

where \( \beta_0 \) lies on the line segment joining \( \hat{\beta} \) and \( \beta_0 \) and the inequalities hold for the same reasons as in (A.17). The right-hand side of (A.21) is the same as the right-hand side of (A.17) except that \( |X_i|, X_i^2 \), and \( |X_i| \) are replaced by \( |Z_i|, Z_i^2 \cdot |X_i|, \) and \( |Z_i| \cdot X_i^2 \) respectively. In consequence, the right-hand side of (A.21) is \( o_p(1) \) by the same proof as for (A.17), provided the following moment conditions hold: \( E|Z_i|^{p} < \infty \) for \( p > 3 \), \( E|Z_i|^3 \cdot |X_i|^3 < \infty \), and \( E|Z_i| \cdot X_i^2 < \infty \). The latter hold by Assumption 1 and the proof of (A.13) is complete. \( \square \)

Proof of Theorem A-2. We have

\[
\bar{\beta}_{ni} - \mu_0 = \frac{(1/n) \sum U_i D_i s_{ni}}{ED_i s_{ni}^2} \times \frac{ED_i s_{ni}}{(1/n) \sum D_i s_{ni}}, \tag{A.22}
\]

The second multiplicand on the right-hand side is \( 1 + o_p(1) \) by Lemma A-1. The first multiplicand on the right-hand side has mean \( EU_i D_i s_{ni}/ED_i s_{ni} \), whose absolute value by Assumption 2 is

\[
\frac{|EU_i (1 - D_i) s_{ni}|}{ED_i s_{ni}} \leq \frac{EU_i |1(\gamma_n < \epsilon_i) s_{ni}|}{Es_{ni}(1 + o(1))} = E|U_i|1(\gamma_n < \epsilon_i)(1 + o(1)) = o(1), \tag{A.23}
\]

where the inequality uses (A.4) and Lemma A-2, the first equality uses Assumption 2, and the second equality uses Assumptions 1 and 6. The first multiplicand on the right-hand side of (A.22) has variance equal to

\[
\frac{Var(U_i D_i s_{ni})}{n(ED_i s_{ni})^2} = \frac{\sigma^2 ED_i s_{ni}^2 (1 + o(1))}{n(ED_i s_{ni})^2} \rightarrow 0, \tag{A.24}
\]

using Lemma A-2, Assumption 7, and (A.2). \( \square \)

The proof of Theorem A-3 uses the following condition and lemma.

Condition L.

\[
\forall \epsilon > 0, \lim_{n \rightarrow \infty} E \left( \frac{(B_n - EB_n)^2}{\text{Var}(B_n)} \right) \left( \frac{(B_n - EB_n)^2}{\text{Var}(B_n)} > n \epsilon \right) = 0,
\]

where \( B_n = U_i D_i s_{ni} \).

Condition L is the Lindeberg condition for the triangular array \( \{U_i D_i s_{ni} : i \leq n, n \geq 1\} \).

Lemma A-5. Under Assumptions 1, 2, 3', 4, 6, and 7, Condition L holds.

Proof of Lemma A-5. We have: \( \forall \delta > 0, \)

\[
E|B_n - EB_n|^2 + 2\delta \leq 4^{1 + \delta} E|B_n|^2 + 2\delta = 4^{1 + \delta} E|U_i D_i s_{ni}|^2 + 2\delta \\
\quad \leq 4^{1 + \delta} E|U_i|^2 + 2\delta E s_{ni}^2 + 2\delta \leq 4^{1 + \delta} E|U_i|^2 + 2\delta E s_{ni}^2, \tag{A.25}
\]
where the first inequality holds by Minkowski's inequality, the second inequality uses Assumption 2(b), and the last inequality uses Assumption 3'. Now, using Lemmas A-2 and A-3, for $0 < \delta \leq 1$, the left-hand side of Condition L equals

$$\lim_{n \to \infty} \frac{E[(B_n - \mu_0)^2]}{\sigma^2(\sum D_i s_{ni})^2 \epsilon(1 + o(1))} \leq \lim_{n \to \infty} \frac{E[B_n - \mu_0]^2}{\sigma^2(\sum D_i s_{ni})^2} \epsilon^{1 + \delta} \leq \lim_{n \to \infty} \frac{4^{1 + \delta} E[U_i^2]}{\sigma^2(\sum D_i s_{ni})^2} \epsilon^\delta (nP(W_i > \gamma_n + b)) \to 0,$$

(A.26)

where the second inequality uses (A.25) and the convergence to zero holds by Assumptions 1 and 7.

**Proof of Theorem A-3.** For part (a), we write

$$\hat{\mu}_n - \mu_0 = \frac{\sum U_i D_i s_{ni}}{\sum D_i s_{ni}},$$

(A.27)

and so,

$$\frac{\sqrt{n}ED_i s_{ni}}{\sigma(ED_i s_{ni})^{1/2}} \left( \frac{\sum U_i D_i s_{ni}}{\sum D_i s_{ni}} - \frac{EU_i D_i s_{ni}}{1/n \sum D_i s_{ni}} \right) = \frac{ED_i s_{ni} Var^{1/2}(U_i D_i s_{ni})}{\sigma(ED_i s_{ni})^{1/2}} \left( \frac{1/n \sum U_i D_i s_{ni} - EU_i D_i s_{ni}}{Var^{1/2}(U_i D_i s_{ni})} \right).$$

(A.28)

The second multiplicand of the right-hand side is asymptotically $N(0, 1)$ by the CLT for "infinitesimal" independent non-identically distributed random variables because the Lindeberg condition (Condition L) holds by Lemma A-5, see Chow and Teicher (1978, Cor. 12.2.2, p. 434). (We note that the summands of the triangular array are infinitesimal, because $\sup P(1/n) (U_i D_i s_{ni} - EU_i D_i s_{ni})/Var^{1/2}(U_i D_i s_{ni}) > 0 \leq 1/(n^2) \to 0$ using Markov's inequality.) The first multiplicand on the right-hand side of (A.28) equals $1 + o_p(1)$ by Lemmas A-1 and A-3.

For part (a), it remains to show that

$$\frac{\sqrt{n}ED_i s_{ni}}{\sigma(ED_i s_{ni})^{1/2}} \left( \frac{EU_i D_i s_{ni} - EU_i D_i s_{ni}}{1/n \sum D_i s_{ni}} \right) = o_p(1).$$

(A.29)

The left-hand side of (A.29) equals

$$- \frac{EU_i D_i s_{ni}}{\sigma(ED_i s_{ni})^{1/2}} \left( \frac{1/n \sum D_i s_{ni} - ED_i s_{ni}}{1/n \sum D_i s_{ni}} \right) = - \frac{EU_i D_i s_{ni}}{\sigma ED_i s_{ni}} \frac{1/n \sum D_i s_{ni} - ED_i s_{ni}}{(ED_i s_{ni})^{1/2}} (1 + o_p(1))$$

$$= - \frac{EU_i D_i s_{ni}}{\sigma ED_i s_{ni}} o_p(1) = o_p(1),$$

(A.30)

where the first equality uses Lemma A-1, the second equality holds because the second multiplicand has mean zero and variance less than or equal to one for all $n \geq 1$, and the third equality holds by (A.23). This completes the proof of part (a).

Part (b) follows from part (a), because the left-hand side of part (b) differs from that of part (a) by a non-stochastic quantity that goes to zero as $n \to \infty$ if and only if Assumption 8 holds.

**Proof of Theorem A-4.** The converse holds by the converse result of Theorem A-3(b).
Hence, we suppose Assumption 8 holds. We use the Cramer–Wold device and the proof of Theorem A-3. Let \( \mathbf{c} = (c_1, c_2) \) be an arbitrary unit vector. We need to show that

\[
\frac{\sqrt{n} E D_i, s_{n}}{\sigma(E D_i, s_{n})^{1/2}} c_1 (\mu_{n0} - \mu_0) + \sqrt{n} c_2 \Omega^{-1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} N(0, 1). \tag{A.31}
\]

By the proof of Theorem A-3 and Assumptions 5 and 8, the left-hand side above equals

\[
(1 + o_p(1)) (1/\sqrt{n}) \sum_i \left( \frac{c_1 U_i D_i s_{n} - E U_i D_i s_{n}}{\text{Var}^{1/2} (U_i D_i s_{n})} + c_2 \Omega^{-1/2} Q_i \right) + o_n(1). \tag{A.32}
\]

The summands are easily seen to be infinitesimal and we show below that the sum of their variances equals \( 1 + o(1) \), so the same CLT as used in the proof of Theorem A-3 applies here provided the Lindeberg condition hold.

We now compute the variance of the summands of (A.32)

\[
\text{Var} \left( \frac{c_1 U_i D_i s_{n} - E U_i D_i s_{n}}{\text{Var}^{1/2} (U_i D_i s_{n})} + c_2 \Omega^{-1/2} Q_i \right) = 1 + 2 \text{Cov} \left( \frac{c_1 U_i D_i s_{n}}{\text{Var}^{1/2} (U_i D_i s_{n})}, c_2 \Omega^{-1/2} Q_i \right). \tag{A.33}
\]

and

\[
\left| \text{Cov} \left( \frac{c_1 U_i D_i s_{n}}{\text{Var}^{1/2} (U_i D_i s_{n})}, c_2 \Omega^{-1/2} Q_i \right) \right| \leq \sigma(E s_{n})^{1/2} (1 + o(1))
\]

\[
\leq \sigma^{-1} (E |U_i c_2 \Omega^{-1/2} Q_i|^p)^{1/p} \left( \frac{P(W_i > \gamma_n \alpha + b)}{P(W_i > \gamma_n)} \right)^{1/2} (1 + o(1)) = o(1). \tag{A.34}
\]

where the first inequality uses Lemmas A-2 and A-3, \( p \) is as in Assumption 5, \( 1/p + 1/q = 1 \), and the last equality holds by Assumptions 4–6 since \( 2/q > 1 + \xi \).

It remains to verify the Lindeberg condition. Let

\[
A_n = c_1 (U_i D_i s_{n} - E U_i D_i s_{n}) / \text{Var}^{1/2} (U_i D_i s_{n}) \quad \text{and} \quad Q = c_2 \Omega^{-1/2} Q_i. \tag{A.35}
\]

Lemma A-5 implies that the Lindeberg condition holds for \( \{ A_n : n \geq 1 \} : \)

\[
\lim_{n \to \infty} E A_n^2 1(A_n^2 \geq n \varepsilon) = 0 \quad \forall \varepsilon > 0. \tag{A.36}
\]

We need to show that the Lindeberg condition holds for \( \{ A_n + Q : n \geq 1 \} : \)

\[
\lim_{n \to \infty} E (A_n + Q)^2 1((A_n + Q)^2 > n \varepsilon) = 0 \quad \forall \varepsilon > 0. \tag{A.37}
\]

The left-hand side of (A.37) is less than or equal to

\[
\lim_{n \to \infty} E (2A_n^2 + 2Q^2) 1(2A_n^2 + 2Q^2 > n \varepsilon). \tag{A.38}
\]

We consider the two summands of (A.38) separately:

\[
\lim_{n \to \infty} E Q^2 1(2A_n^2 + 2Q^2 > n \varepsilon) \leq \lim_{n \to \infty} (E |Q|^2)^{1/2} P(2A_n^2 + 2Q^2 > n \varepsilon)^{1/2} = 0,
\]

\[
\leq (E |Q|^2)^{1/2} \lim_{n \to \infty} \left( \frac{2E A_n^2 + 2EQ^2}{n \varepsilon} \right)^{1/2} = 0, \tag{A.39}
\]

\[
\leq (E |Q|^2)^{1/2} \lim_{n \to \infty} \left( \frac{2E A_n^2 + 2EQ^2}{n \varepsilon} \right)^{1/2} = 0, \tag{A.39}
\]

Downloaded from https://academic.oup.com/restud/article-abstract/65/3/497/1565376 by guest on 17 October 2019
where the equality holds because \( E A_n^2 = 1 \) and \( E|Q|^4 < \infty \) by Assumption 5. Next, we have

\[
\lim_{n \to \infty} E A_n^2 (A_n > n \epsilon / 2) \\
\leq \lim_{n \to \infty} E A_n^2 (Q > M_n) + \lim_{n \to \infty} E A_n^2 (A_n \leq n, Q > M_n) + \lim_{n \to \infty} E A_n^2 (A_n > n \epsilon / 2 - M_n)
\]

where the last inequality holds by the Lindeberg condition for \( \{A_n: n \geq 1\} \) and Markov’s inequality, \( \lambda > 2 \) is as in Assumption 5, and the equality holds by taking \( M_n = n \epsilon / 4 \) using Assumption 5 and the Lindeberg condition for \( \{A_n: n \geq 1\} \). This completes the proof of the Lindeberg condition for \( \{A_n + Q: n \geq 1\} \).

Theorem 4 follows from Theorems A-5 and A-6 below. Define

\[
\delta^2_{x0} = \frac{1}{n} \sum D_i s_n / D_i s_{nl}.
\]

**Theorem A-5.** Under Assumptions 1, 2(b), 3', 4, 6 and 7, \( \delta^2_{x0} \to \sigma^2 \).

**Theorem A-6.** Under Assumptions 1, 2(b), and 3-7, \( \delta^2_{x0} - \delta^2_{x0} \to 0 \).

**Proof of Theorem A-5.** By Lemma A-1,

\[
\delta^2_{x0} = \frac{1}{n} \sum U_i^2 D_i s_{nl} / E D_i s_{nl} (1 + o_p(1)).
\]

Let \( V_{nl} = (1/n) \sum U_i^2 D_i s_{nl} / E D_i s_{nl} \). We have

\[
EV_{nl} = \frac{E U_i^2 D_i s_{nl}}{ED_i s_{nl}} = \frac{E U_i^2 s_{nl} (1 + o(1))}{ED_i s_{nl}} = \frac{\sigma^2 E s_{nl} (1 + o(1))}{ED_i s_{nl}} = \sigma^2 (1 + o(1)),
\]

using Lemmas A-2 and A-3 and Assumption 2(b). We also have

\[
\text{Var} (V_{nl}) = \frac{\text{Var} (U_i^2 D_i s_{nl})}{n(ED_i s_{nl})^2} \leq \frac{E U_i^2 s_{nl} (1 + o(1))^2}{n(ED_i s_{nl})^2} = \frac{E U_i^2 ED_i s_{nl} (1 + o(1))}{n(ED_i s_{nl})^2} = o(1),
\]

where the second to last equality holds by Lemma A-2 and Assumption 2(b) and the last equality holds by Assumptions 1 and 7 and (A.2).

**Proof of Theorem A-6.** The left-hand side in this theorem can be written as:

\[
\left( \frac{\hat{A} - A}{\hat{A} - A} \right) = \frac{\hat{A} - A}{\hat{B} - B} \left( \frac{\hat{B} - B}{\hat{B} - B} \right).
\]

where

\[
A = \sum U_i^2 D_i s_{nl}, \quad \hat{A} = \sum Y_i - \hat{\mu}_n - Z_t \hat{\delta} D_i s_{nl},
\]

\[
B = \sum D_i s_{nl}, \quad \hat{B} = \sum D_i s_{nl}.
\]

To show that the left-hand side converges in probability to zero it suffices to show that: (i) \( \hat{B} / B \to 1 \), (ii) \( (\hat{A} - A) / B \to 0 \), and (iii) \( A / B = O_p(1) \). By the proof of Theorem A-1, (i) holds. (It is implied by condition (iv) of that proof.) By Theorem A-5, \( A / B \to \sigma^2 \), so (iii) is satisfied.
Using Lemmas A-1, A-2 and A-3, Assumptions 3, 5, 7, equation (A.2) and Theorem 1, we find that the following six conditions are sufficient for (ii)

\[
\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i D_i (\delta_{si} - s_{si}) \right) / (ED_i s_{si}^2)^{1/2} \xrightarrow{p} 0, \tag{A.46}
\]

\[
\left( \frac{1}{n} \sum_{i=1}^{n} D_i s_{si} \right) / E s_{si} = O_p(1), \tag{A.47}
\]

\[
\left( \frac{1}{n} \sum_{i=1}^{n} Z_i D_i s_{si} \right) / E s_{si} = O_p(1), \tag{A.48}
\]

\[
\left( \frac{1}{n} \sum_{i=1}^{n} U_i D_i s_{si} \right) / E s_{si} = O_p(1), \tag{A.49}
\]

\[
\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i U_i D_i s_{si} \right) / (ED_i s_{si}^2)^{1/2} \xrightarrow{p} 0, \tag{A.50}
\]

and

\[
\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i Z_i D_i s_{si} (\hat{\theta} - \theta_0) \right) / (ED_i s_{si}^2)^{1/2} \xrightarrow{p} 0. \tag{A.51}
\]

The proof of (A.46) is the same as that of (A.12) except that the factor \( U_i^2 \) appears in various sums and expectations instead of \( |U_i + \mu_0| \). By Assumption 2(b), this requires \( EU_i^2 < \infty \), which holds by Assumption 1.

By the proof of Theorem A-1, (A.47), (A.48) and (A.49) hold. In particular, equation (A.47) is implied by condition (i) of that proof together with Lemmas A-1 and A-2. Equation (A.48) follows from (A.2) and (A.12). Equation (A.49) follows from (A.2), (A.12) and (A.14).

The proof of (A.50) is the same as that of (A.13) except that the factor \( |U_i| \) appears in various sums and expectations. In analogy with the proof of (A.12), using Assumption 2(b), this requires \( E |U_i| < \infty \), which holds by Assumption 1.

We now establish (A.51). By Assumption 5 and a one-term Taylor expansion about \( \beta_0 \), the absolute value of the left-hand side of (A.51) is bounded by

\[
O_p(1) \left( \frac{1}{n^{3/2}} \sum_i Z_i E_s \right)^{1/2} \leq O_p\left( \frac{1}{n} \sum_i \|Z_i\|^3 \right) s_{si} + O_p\left( \frac{1}{n} \sum_i \|Z_i\|^3 \right) X_i (X_i \beta_0 - \gamma_0) \sqrt{m(\hat{\beta} - \beta_0)} \right) / (n E s_{si}^2)^{1/2} \leq O_p(1) \left( \frac{1}{n} \sum_i \|Z_i\|^3 (W_i \gamma_0) \right) / (n E s_{si}^2)^{1/2} + O_p(1) \left( \frac{1}{n} \sum_i \|Z_i\|^3 \|X_i\| \right) / (n E s_{si}^2)^{1/2}. \tag{A.52}
\]

By the same proof as for (A.21), the right-hand side of (A.52) is \( o_p(1) \) provided \( E \|Z_i\|^p < \infty \) for \( p > 3 \) and \( E \|Z_i\|^3 \|X_i\| < \infty \). The latter hold by Assumption 1.

**Proof of Theorem 5.** By the proof of Theorem A-1, \( (1/n) \sum_{i=1}^{n} D_i s_{si} / ((1/n) \sum_{i=1}^{n} D_i s_{si}^2) \xrightarrow{p} 1. \) (It is condition (i) of that proof.) If we show, analogously, that

\[
(1/n) \sum_{i=1}^{n} D_i s_{si}^2 / ((1/n) \sum_{i=1}^{n} D_i s_{si}^2) \xrightarrow{p} 1, \tag{A.53}
\]

then the left-hand side of the theorem equals

\[
\left( \frac{\sqrt{n} E D_i s_{si}}{\sigma(ED_i s_{si}^2)^{1/2}} \right) \left( 1 + O_p(1) (\hat{\mu}_a - \mu_0) \right), \tag{A.54}
\]

\[
\left( \frac{\sqrt{n} \Omega^{1/2}}{\sigma(ED_i s_{si}^2)^{1/2}} (\hat{\beta} - \theta_0) \right),
\]

where

\[
\Omega = \left( \frac{1}{n} \sum_{i=1}^{n} \|Z_i\|^3 \|X_i\| \right) / (n E s_{si}^2)^{1/2}.
\]
using Lemmas A-1 and A-4, Assumption 9, and Theorem 4. Theorem 5 follows from (A.54) and Theorem 3. It remains to establish (A.53). Using Lemma A-4, we find that the following condition is sufficient:

\[
\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} D_j (\tilde{\beta}_j - \hat{\beta}_j) / \sqrt{\text{Var}(\tilde{\beta}_j)} \right) \to 0.
\]

(A.55)

The proof is analogous to that of (A.15) with \(s_m\) replaced by \(s_m^v\) and \(s_m^i\) replaced by \(s_m^i_v\). In addition to the assumptions required for (A.15), the proof uses (A.8) of the proof of Lemma A-4.

Acknowledgements. This paper was partially supported by National Science Foundation Grant SES-9121914 and by a C. A. Anderson Fellowship of the Cowles Foundation. The authors wish to thank two referees and James Heckman for their comments, Glena Ames for her able assistance with the typing of this document, and Rosemarie Lewis for proofreading the manuscript.

REFERENCES


