Finite sample properties for the semiparametric estimation of the intercept of a censored regression model

Marcia M.A. Schafgans*
London School of Economics and Political Science
Department of Economics

Abstract

A popular two-step estimator of the intercept of a censored regression model is compared with consistent asymptotically normal semiparametric alternatives. Using a root mean squared error criterion, the semiparametric estimators perform better for a range of bandwidth parameter choices for a variety of distributions of the errors and regressors. For error distributions that are close to the normal, however, the two-step parametric estimator performs better.

KEYWORDS: Sample selection model, consistency, asymptotic normality, semiparametric estimation, Monte Carlo simulations.

1 Introduction

Censored regression models arise frequently in biostatistics, econometrics, and other areas of statistics. Unlike in the standard linear regression model, standard parametric techniques for estimating censored regression models give rise to inconsistent estimators when based on incorrect distributional assumptions. As a result, semiparametric estimation of censored regression models has received considerable attention in the last decade. Various estimates have been shown to be $\sqrt{n}$-consistent and asymptotically normal under distribution-free assumptions. However, their finite sample performance is in doubt especially due to their presumable sensitivity to a bandwidth or smoothing parameter.

The present paper compares the finite sample performance of a particular parametric estimate, due to Heckman (1976, 1979), with consistent semiparametric alternatives, in the

*Correspondence: Marcia Schafgans, Department of Economics, London School of Economics, Houghton Street, London WC2A-2AE. Email: m.schafgans@lse.ac.uk. Financial support for this paper was provided by a C.A. Anderson Fellowship of the Cowles Foundation. I wish to thank Donald Andrews, Moshe Buchinsky, Oliver Linton, and Peter Robinson for helpful discussions. I am, of course, responsible for any remaining errors.

context of a particular censored regression or sample selection model. We focus in particular on estimation of the intercept of the 'outcome' equation of this model when Heckman's normality assumption does not hold. Given the potential sizeable magnitude of the inconsistency of the parametric two step Heckman estimator of the intercept (e.g., see Goldberger (1983), Arabmazar and Schmidt (1981, 1982), and Schafgans (1997)), this is a question well worth considering. The semiparametric estimators of the intercept considered are those given by Heckman (1990) and Andrews and Schafgans (1998). A proof of the consistency and asymptotic normality of the Heckman estimator (1990) is given by Schafgans and Zinde-Walsh (2000).

Both semiparametric estimators depend on a bandwidth or smoothing parameter. An automated method for determining this bandwidth parameter has not yet been developed. This study will evaluate the influence of choosing the bandwidth parameter on the estimates obtained. For finite sample sizes, the root mean squared errors of the semiparametric estimators are compared with that of the parametric Heckman two-step alternative. Both semiparametric estimators perform better for a range of bandwidth parameter choices for a variety of distributions of the errors and regressors. For error distributions that are close to the normal, however, the two-step parametric estimator performs better. Regarding the finite sample performance of the semiparametric Heckman (1990) estimator and the Andrews and Schafgans (1998) estimator the simulation results show that the Heckman estimator would appear to be more efficient, while the bias of the Andrews Schafgans estimator for each given bandwidth is lower in absolute value than for the Heckman estimator. Overall, our simulations tend to favour the finite sampling behaviour of the Andrews-Schafgans (1998) estimator over the Heckman estimator (1990). This is in line with findings in other nonparametric estimation problems, which show that the trade-off between bias and variance is typically better for smooth "kernels". Schafgans and Zinde-Walsh (2000) compare the asymptotic properties of the Andrews-Schafgans (1998) and Heckman (1990) estimators for a wide class of distributional assumptions and derive "optimal" bandwidth parameters.

The remainder of the paper is organized as follows: Section 2 discusses the censored regression or sample selection model used briefly. Section 3 discusses the inconsistency of the parametric two step estimator of the intercept. Section 4 gives the consistent, asymptotically normal, semiparametric estimators of the intercept and Section 5 discusses the simulation results. Section 6 concludes.

2 A Censored Regression Model

Because the presence of censored data are so common, econometricians and statisticians, alike, have denoted much efforts to the analysis of censored data (Manski (1993)). Special attention has been drawn to the censored regression or sample selection model. In latent

variable notation, the model can be written as

$$\begin{aligned} Y_i^* &= \mu_0 + U_i \ , \\ D_i &= 1(X_i'\beta_0 + \varepsilon_i > c) \ , \quad \text{and} \\ Y_i &= Y_i^* D_i \quad \text{for} \quad i = 1, ..., n, \end{aligned} \tag{1}$$

where $(Y_i,\ D_i,\ X_i)$ are observed random variable, and c is some known truncation point. The first equation is the outcome equation and the second equation is the participation equation. The outcome equation only contains an intercept, implying that we are primarily concerned with the estimation of the population mean μ_0 in this paper. In a more general setting, the outcome equation would contain a regression function as well, for instance $Z_i'\theta_0$. Theoretically, the need for using the sample selection model when interested solely in the parameters of the outcome equation, comes from the possible correlation between the outcome equation error, U_i, and the selection equation error, ε_i. This is sometimes referred to as the situation where there is a 'nonignorable nonresponse' (see, for example, Rubin, 1987). The use of a competing-risks model (Kalbfleisch and Prentice (1980)) as one statistical practice to deal with nonignorable nonresponse has a related latent variable setup.

The discussion of the econometric implications of sample selectivity 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 labor force or not. The distribution of the wage offers sampled is truncated by the 'self-selection' of women in the labor force, where women choose to be 'in the sample' of workers if the offered wage exceeds their reservation wage. Sample selection model have been used in a wide variety of other applications, e.g., see Maddala (1983) and Amemiya (1984).

To express the model given in (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 the individual is employed, i.e., whether $Y_i^* - Y_i^r$ exceeds zero, where Y_i^r denotes the individual's latent reservation wage. The observed wage is given by Y_i. The variables influencing the decision to participate in the labor market are given by X_i and μ_0 is the population mean of the offered wage.

Standard parametric approaches in econometrics to estimating the parameters of this model assume that (ϵ_i, U_i) have a bivariate normal distribution, independent of X_i, with zero mean and unknown covariance matrix. With this assumption, the parameters can be estimated by maximum likelihood or the more convenient but less efficient two-step estimator of Heckman (1976, 1979). Nevertheless, from the Heckman estimate (or any $\sqrt{N}$-consistent semiparametric estimate) one Newton-step can be performed to match the efficiency of maximum likelihood estimation.

Unlike in the standard linear regression model, deviations from normality in the censored

regression model lead to biased and inconsistent estimators. For the Tobit model Goldberger (1983) and Arabmazar and Schmidt (1982) have documented the magnitude of this inconsistency for a variety of symmetric distributions of the errors.

The present paper compares, in the context of the particular censored regression or sample selection model given in (1), the finite sample performance of a popular parametric two-step estimator with consistent asymptotically normal semiparametric alternatives. In particular, we compare the parametric two-step estimator of Heckman (1976, 1979) with the semiparametric estimators given by Heckman (1990) and Andrews and Schafgans (1998). The popular parametric two-step Heckman estimator is used rather than the maximum likelihood estimator, since it is consistent for a more general class of dependence models discussed by Olson (1980), i.e. where the errors of the outcome equation, U_i, are linear in the errors of the selection equation, ε_i and the ε_i are normally distributed.

3 The Inconsistent Two-Step Heckman Estimator of the Intercept

This paper is primarily concerned with the robustness of the estimator of the intercept to deviations from normality. Expressions for the inconsistency of the two-step Heckman estimator are derived for both the case where β_0 is known and the case where it is not known. Simulation results in Schafgans (1997) suggest that when applying the parametric Heckman two-step estimation, particular care should be taken with (i) potential skewness in the distribution of the outcome equation errors when there is much censoring, and (ii) high covariance between the selection and outcome equation errors.

For the censored regression or sample selection model with truncation at c, we have:

$$\begin{aligned} E(Y_i \,|\, D_i = 1, X_i) &= \mu_0 + E(U_i \,|\, D_i = 1, X_i) \\ &= \mu_0 + h(c - X_i'\beta_0), \end{aligned} \tag{2}$$

where $h(c - X_i'\beta_0) = E(U_i \,|\, X_i'\beta_0 + \varepsilon_i > c, X_i)$.

Under the assumption that (U_i, ε_i) have a bivariate normal distribution

$$\begin{aligned} h(c - X_i'\beta_0) &= \sigma_{\varepsilon U} E\left(\varepsilon_i \,|\, \varepsilon_i > c - X_i'\beta_0, X_i\right) \\ = \sigma_{\varepsilon U} \frac{\phi(c - X_i'\beta_0)}{1 - \Phi(c - X_i'\beta_0)} &= \sigma_{\varepsilon U} g^*(c - X_i'\beta_0) \equiv h^*(c - X_i'\beta_0), \end{aligned} \tag{3}$$

where $g^*(\cdot)$ is the inverse Mill's ratio (Johnson and Kotz (1970, p. 278f.) give various expansions of the inverse Mill's ratio). The variance of the selection equation errors, σ_ε^2, is normalized to equal one. More generally, as Olsen (1980) pointed out, the above equality holds when U_i is linear in ε_i and ε_i is normally distributed.

The two-step Heckman procedure, for given β_0, reduces to a least squares estimation on the uncensored observations of

$$Y_i = \mu_0 + \sigma_{\varepsilon U} g^*(c - X_i'\beta_0) + \nu_i, \qquad \text{for } i \text{ s.t. } D_i = 1, \tag{4}$$

where

$$\nu_i = U_i - \sigma_{\varepsilon U} g^*(c - X_i'\beta_0). \tag{5}$$

When (U_i, ε_i) have a bivariate normal distribution, or if U_i is linear in ε_i and ε_i is normally distributed, $E(\nu_i \mid D_i = 1, X_i) = 0$.

Equation (4) can also be written as

$$Y_i D_i = (\mu_0 + \sigma_{\varepsilon U} g^*(c - X_i'\beta_0) + \nu_i) D_i, \qquad i = 1, ..., n. \tag{6}$$

The two-step Heckman estimator $\widehat{\mu}$ solves

$$\begin{aligned} \widehat{\mu} &= \frac{1}{\frac{1}{n}\sum_{i=1}^n D_i} \left(\frac{1}{n} \sum_{i=1}^n D_i (Y_i - \widehat{\sigma}_{\varepsilon U} g^*(c - X_i'\beta_0)) \right) \\ \widehat{\sigma}_{\varepsilon U} &= \frac{\frac{1}{n}\sum_{i=1}^n (D_i Y_i - \overline{DY}\ )(D_i g^*(c - X_i'\beta_0) - \overline{Dg^*})}{\frac{1}{n}\sum_{i=1}^n \left(D_i g^*(c - X_i'\beta_0) - \overline{Dg^*} \right)^2}, \end{aligned} \tag{7}$$

where $\overline{DY} = \sum_{i=1}^n D_i Y_i / n$ and $\overline{Dg^*} = \sum_{i=1}^n D_i g^*(c - X_i'\beta_0)/n$. In the case in which $\sigma_{\varepsilon U}$ is known, we set $\widehat{\sigma}_{\varepsilon U}$ equal to $\sigma_{\varepsilon U}$ and solve (7) only for $\widehat{\mu}$.

This estimator of μ_0 (and $\sigma_{\varepsilon U}$) is consistent if (U_i, ε_i) have a bivariate normal distribution or if U_i is linear in ε_i and ε_i is normally distributed. If (U_i, ε_i) do not satisfy these conditions, then the estimator $\widehat{\mu}$ (and $\widehat{\sigma}_{\varepsilon U}$) is inconsistent, since, in general, $E(\nu_i \mid D_i = 1, X_i) \neq 0$. Under the true underlying distribution of (U_i, ε_i), we note that

$$\nu_i D_i = (h(c - X_i'\beta_0) - \sigma_{\varepsilon U} g^*(c - X_i'\beta_0) + \nu_i') D_i, \tag{8}$$

where $E(\nu_i' \mid D_i = 1, X_i) = 0$ by construction.

The probability limit of $\widehat{\mu}$ is given by:

$$\mu^* \equiv \operatorname{plim} \widehat{\mu} = \mu_0 + E(h(c - X_i'\beta_0) | D_i = 1, X_i) - \sigma^*_{\varepsilon U} E(g^*(c - X_i'\beta_0) | D_i = 1, X_i), \tag{9}$$

where

$$\sigma^*_{\varepsilon U} \equiv \operatorname{plim} \widehat{\sigma}_{\varepsilon U} = \frac{\operatorname{Cov}\left(h(c - X_i'\beta_0), g^*(c - X_i'\beta_0) | D_i = 1, X_i\right)}{\operatorname{Var}(g^*(c - X_i'\beta_0) | D_i = 1, X_i)}. \tag{10}$$

These probability limits exist, for instance, if $\{(U_i, \varepsilon_i, X_i)\}$ is an i.i.d. sequence (or stationary and ergodic), $E\left(|U_i|^2 \mid \varepsilon_i > c - X_i'\beta_0, X_i\right) < \infty$, and $E\left(g^*(c - X_i'\beta_0)^2 \mid \varepsilon_i > c - X_i'\beta_0, X_i\right) < \infty$.

Using (3), the inconsistency of the intercept can be written as follows

$$\begin{aligned} \mu^* - \mu_0 = &E\left[h(c - X_i'\beta_0) - h^*(c - X_i'\beta_0) | D_i = 1, X_i\right] - \\ &(\sigma^*_{\varepsilon U} - \sigma_{\varepsilon U}) E(g^*(c - X_i'\beta_0) | D_i = 1, X_i). \end{aligned} \tag{11}$$

The inconsistency depends on the truncation point c, $h(z) = E(U_i | \varepsilon_i > z)$, the distribution of $X_i'\beta_0, \sigma_{\varepsilon U}$, and the inconsistency of the estimator of $\sigma_{\varepsilon U}$. The inconsistency of the estimator $\widehat{\mu}$ reduces to the first term on the right hand side of (11) if $\sigma_{\varepsilon U}$ is assumed to be known.

When β_0 is unknown, the inconsistency of the intercept has an additional term. This term depends non-linearly on the inconsistency of the estimator of β_0. Specifically,

$$\begin{aligned}\mu^* - \mu_0 = & E\left[h(c - X_i'\beta_0) - h^*(c - X_i'\beta_0)|D_i = 1, X_i\right] - \\ & (\sigma^*_{\varepsilon U} - \sigma_{\varepsilon U})E\left[g^*(c - X_i'\beta_0)|D_i = 1, X_i\right] - \\ & \sigma^*_{\varepsilon U}E\left[g^*(c - X_i'\beta_0) - g^*(c - X_i'\beta^*)|D_i = 1, X_i\right],\end{aligned} \tag{12}$$

where $\beta^* \equiv \operatorname{plim}\widehat{\beta}$. The existence of this probability limit requires, additionally, that ε_i has a distribution function. (Proof similar to Amemiya (1985)). In this case, the probability limit of $\widehat{\sigma}_{\varepsilon U}$ is given by:

$$\sigma^*_{\varepsilon U} = \frac{\operatorname{Cov}\left(h(c - X_i'\beta_0), g^*(c - X_i'\beta^*)|D_i = 1, X_i\right)}{\operatorname{Var}(g^*(c - X_i'\beta^*)|D_i = 1, X_i)}. \tag{13}$$

4 The Consistent Semiparametric Intercept Estimators

A consistent and asymptotically normal estimator for the intercept, μ_0, was provided by Andrews and Schafgans (1998). Their estimator is given by:

$$\hat{\mu}_n = \frac{\sum_{i=1}^n Y_i D_i s(X_i'\widehat{\beta} - c - \gamma_n)}{\sum_{i=1}^n D_i s(X_i'\widehat{\beta} - c - \gamma_n)}, \tag{14}$$

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 parameter γ_n is called the bandwidth or smoothing parameter. This bandwidth parameter is chosen such that $\gamma_n \to \infty$ as $n \to \infty$. Finally, $\widehat{\beta}$ is a $\sqrt{n}$-consistent preliminary estimator of β_0. The literature on semiparametric estimation of censored regression or sample selection models gives several root-n consistent and asymptotically normal estimators for the parameters, β_0 (up to some unknown scale) in (1). For instance, one could consider: Ichimura (1993), Powell, Stock, and Stoker (1989), and Klein and Spady (1993).

Andrews and Schafgans' (1998) estimator is an adaptation of the estimator suggested by Heckman (1990)

$$\tilde{\mu}_n = \frac{\sum_{i=1}^n Y_i D_i 1(X_i'\widehat{\beta} - c > \gamma_n)}{\sum_{i=1}^n D_i 1(X_i'\widehat{\beta} - c > \gamma_n)}. \tag{15}$$

Both estimators make use of the idea of 'identification at infinity' mentioned by Chamberlain (1986). Only those observations for which the probability of selection in the censored or truncated sample is close to one and in the limit as $n \to \infty$ is one, are used for estimation of the intercept. 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.

The estimator suggested by Andrews and Schafgans (1998) differs from Heckman's (1990) $\tilde{\mu}_n$ only in that it replaces the indicator function $1(\cdot)$ with a smooth function $s(\cdot)$. The

introduction of this function facilitated them to provide the estimator with a distribution theory. The smoothness imposed on this function, viz., differentiability of order three, is used to show that the preliminary estimator $\widehat{\beta}$ does not affect the asymptotic results. The conjecture of Andrews and Schafgans that also Heckman's estimator is asymptotically normal was proven by Schafgans and Zinde-Walsh (2000).

Essentially, Heckman's estimator $\widetilde{\mu}_n$ is a sample average of the random variables $U_i + \mu_0$ over a fraction of all observations, since $Y_i \rightarrow_p U_i + \mu_0$ as $n \rightarrow \infty$ for all $i \geq 1$. The effective sample size is equal to the number of observations used for the estimation of μ_0. Since AS introduced a weighting scheme for these observations, viz., the smooth function $s(\cdot)$, the estimator $\hat{\mu}_n$ is a weighted sample average of the random variables $U_i + \mu_0$, where observations with $X_i'\widehat{\beta}$ greater than γ_n and with $X_i'\widehat{\beta}$ close to the threshold γ_n are weighted less than those further away.

The aim of the simulations presented in the next section is (a) to show that it is feasible to improve (in terms of root mean squared error) on the parametric Heckman two-step estimator using the semiparametric alternatives and (b) to reveal the sensitivity of the semiparametric estimators to the choice of the bandwidth. In addition, it allows us to compare the Heckman (1990) estimator with the Andrews and Schafgans (1998) estimator.

5 Simulation of the Semiparametric Estimator

In this section, simulation results are presented for the semiparametric estimator of μ_0 under non-normality. The simulation results in this section are based on 1,000 random draws of the censored regression or sample selection model given in (1) with 1,000 and 500 observations. The amount of censoring considered is equal to 20, 50, and 80 percent. The true parameter vector $(\mu_0, \beta_0)'$ is given by $(0, 1, 1)'$. For this purpose, c is chosen so that $P(X_i'\beta_0 < c) = 0.2$, 0.5, and 0.8 as in Section 4.

The distributions of the selection equation errors considered are the Student t distribution and the chi-squared distribution with degrees of freedom set low (specifically three and five, each standardized to have zero mean and unit variance). In these instances, the parametric two-step Heckman estimator gives rise to the largest deviations from the true population mean (see, Schafgans (1997)). The standardization ensures that comparisons among these distributions are not confused with differences in scale. The distribution of the outcome equation errors are determined by the class of dependence models given by Olson (1980), i.e., $U_i = \sigma_{\varepsilon U}\, \varepsilon_i + V_i$, where ε_i and V_i are independent. In the results presented, V_i has a normal distribution with variance equal to $\sigma_V^2 = \sigma_U^2(1 - \rho_{\varepsilon U}^2)$, where σ_U^2 is set equal to two. Explorations with other choices of the distribution of V_i seem to indicate that the results are not very sensitive to the specific choice made for this distribution.

Lastly, two distributions of the selection index $X_i'\beta_0$ are considered, the normal and the chi-squared distribution. There are two regressors in X_i, X_{1i} and X_{2i}. In the first case

($X_i'\beta_0$ has a normal distribution) X_{1i} and X_{2i} are two independently drawn normal random variables, and in the second case ($X_i'\beta_0$ has a chi-squared distribution), X_{1i} and X_{2i} are two independently drawn chi-squared random variables with two degrees of freedom. The regressors X_{1i} and X_{2i} are standardized in the following way: (i) both regressors have a variance equal to a half (ii) the first regressor X_{1i} has a mean equal to zero, and (iii) the second regressor X_{2i} has a mean which varies in such a way that the amount of censoring is equal to 20, 50, and 80 percent respectively. In each case, the variance of $X_i'\beta_0 \equiv X_{1i} + X_{2i}$ equals one.

Both in the parametric and the semiparametric case the estimation of the true parameters (μ_0, β_0) follow a two step approach. In the parametric case, a probit regression precedes the ordinary least square regression with the inverse Mill's ratio as one of the explanatory variables. In the semiparametric case, the average derivative estimator (Powell et al.(1989)) precedes the semiparametric estimation of the intercept described in detail in the previous section. Following the suggestion in Andrews and Schafgans (1998), the function $s(\cdot)$ in (14) is defined by:

$$s(x) = \begin{cases} 1 - \exp(-\frac{x}{b-x}) & \text{for } x \in (0, b) \\ 0 & \text{for } x \leq 0 \\ 1 & \text{for } x \geq b, \end{cases} \tag{16}$$

where we vary b from one, a half, and zero. When b is set equal to zero, we get the semiparametric estimator for the intercept given by Heckman (1990).

The primary criterion used for comparison of the semiparametric and parametric estimators is the root mean squared error. Below, the abbreviation RMSE is used for the root mean squared error. The feasibility of finding a bandwidth parameter for which the semiparametric estimator is better in terms of RMSE than the parametric alternative is evaluated by computing the RMSE ratio (defined by RMSE of the semiparametric estimator over the RMSE of the parametric Heckman two-step estimator) for a wide range of bandwidth choices. A secondary criterion for comparison is the simulated probability of rejecting the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at a five percent level of significance using t-tests based on the parametric Heckman two-step and semiparametric estimators. For the computation of the rejection rates of parametric Heckman two-step estimator, the bivariate normality assumption is maintained. For the semiparametric estimators, the asymptotic normality results given in Andrews and Schafgans (1998) and Schafgans and Zinde-Walsh (2000) are used. The actual bandwidth chosen for the estimation of μ_0 should coincide with a simulated probability of rejection of the null hypothesis equal to five percent.

A wide range of bandwidth choices γ_n is considered. The choices considered are based on the percentage of the uncensored observations used in the estimation of the semiparametric

estimator $\hat{\mu}_n$. In the simulations, this percentage is computed as

$$\Big(P(X_i'\beta_0 - c > \gamma_n)/P(X_i'\beta_0 - c > 0)\Big) * 100 = \% \text{ Uncensored Observations.} \qquad (17)$$

The actual bandwidth declines with the proportion of uncensored observations. Note that γ_n indicating the use of 100% of the uncensored observations for the estimation of $\hat{\mu}_n$, i.e., $\gamma_n = 0$, does not coincide with the naive truncated regression model. For the naive truncated regression model, we need to set γ_n equal to $-\infty$. For consistency, the bandwidth parameter is required to approach infinity as the number of observations, n, goes to infinity. This means that the probability of $X_i'\beta_0 - c$ exceeding γ_n needs to approach zero as n goes to infinity. This will guarantee that the estimation of μ_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.

The simulation results are presented in Graphs 1 through 6. The horizontal lines in all graphs correspond to the parametric Heckman two-step estimator (since this estimator does not depend on the bandwidth), the remaining lines correspond to the semiparametric estimators.

With the exception of Graph 6, all graphs are based on simulations with 1,000 observations and 1,000 replications.

Distribution of the Selection Equation Errors.
Graph 1 presents the results for the parametric Heckman two-step estimator and the semiparametric estimators of the intercept for a range of bandwidth choices. In part A of the graph, the average bias (in absolute terms) and the standard deviation of the estimators are graphed against the bandwidth. The bandwidth is reported in terms of the percentage of uncensored observations used for the semiparametric estimation of μ_0. In each graph, the results for different choices of the parameter b in the $s(\cdot)$ function are presented for different distributions of the selection equation errors. In part B of the graph, the results from part A are combined to give the RMSE ratio against the bandwidth. Together with this criterion to compare the parametric with the semiparametric estimators, part B of the graph also reports the simulated rejection rate of the t test for the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at a five percent level of significance against the bandwidth.

As expected, the bias of the parametric Heckman two-step estimator decreases as the selection equation errors are closer to the normal distribution. The bias of the semiparametric estimators of μ_0 decreases when the bandwidth parameter is increased (or similarly a lower proportion of the uncensored observations are used in the estimation). On the other hand, the standard deviation of the semiparametric estimators rises when a larger γ_n is chosen. This is to be expected since the sample size on which the estimator of μ_0 is based decreases with γ_n. For bandwidth choices that are based on using a large proportion of the uncensored observations for the semiparametric estimator, the standard deviation of the semiparametric estimator actually is smaller than that of the parametric estimator. As the proportion of uncensored

observations used declines, the standard deviation of the semiparametric estimators surpasses that of the parametric alternative.

When the selection equation errors have a chi-squared distribution, the simulation results indicate that the semiparametric estimators are better than the parametric Heckman two-step estimator for a large range of bandwidth choices. This can be seen by looking at the range of bandwidth choices for which the RMSE ratio is less than one. The improvement in the RMSE of the semiparametric estimator over the parametric estimator is not general. This becomes particularly clear when one considers the simulation results where the selection equation errors have a Student t distribution.

Graph 1 also allows us to compare the different semiparametric estimators, i.e., with b set equal to zero, a half, and one. Panel A shows that for any bandwidth choice the semiparametric Heckman estimator ($b = 0$) has the lowest standard deviation (with the difference increasing with the bandwidth parameter), a finding supported by Schafgans and Zinde-Walsh (2000). In fact, the standard deviation for a given bandwidth parameter decreases with b. In addition, in order to achieve the same level of bias, the bandwidth parameter needs to be based on a smaller proportion of the uncensored observations as b decreases (with the difference decreasing with the bandwidth parameter). This clear dependence between the choice of b and the selection of the bandwidth parameter is further supported by Panel B. First, the RMSE ratio graphs reveal that a larger selection of b should be accompanied by a lower bandwidth parameter (the RMSE curves are shifted to the left). Second, the range for which in addition the simulated rejection rate equals five percent shifts also to the left when b increases.

When the selection equation errors have a chi-squared distribution with three degrees of freedom and censoring equal to 50 percent, the range of bandwidth choices for which the RMSE ratio is less than one is 18–100 percent when b is equal to one, with b equal to a half and zero the range is 16–85 and 10–70 percent respectively. The RMSE ratio does appear to be less sensitive to the choice of the bandwidth parameter for higher values of b (the RMSE curves are flatter). As a result, the range of bandwidth parameters for which the Andrews-Schafgans estimator does better in terms of RMSE than the parametric two-step Heckman estimator widens with b.

When the degrees of freedom are increased, the range of bandwidth choices for which the RMSE ratio is less than one diminishes and moves to the right, indicating the use of a smaller fraction of the uncensored observations. The bandwidth ranges for which, in addition, the simulated rejection rate equals five percent are smaller. Specifically, when the selection equation errors have a chi-squared distribution with three degrees of freedom and censoring equal to 50 percent, the range becomes 18–50 when b is equal to one, with b equal to a half and zero the range becomes 16–40 and 10–35 respectively.

When the selection equation errors are normal and censoring is equal to 50 percent, the

Heckman semiparametric estimator ($b = 0$) can at best be 70 percent worse than Heckman's parametric estimator using 20 percent of the uncensored observations. The same relative efficiency can be reached for the other semiparametric estimators, with b equal to a half and one, but is reached at a higher proportion of the uncensored observations, 27 percent and 35 percent respectively. At these bandwidth choices, however, the rejection rates of the semiparametric estimators still exceed five percent. At a bandwidth choice which is more consistent with the asymptotic normality result of the semiparametric estimator (using only 10 percent of the uncensored observations), the semiparametric estimators only do twice as bad as the parametric alternative when $b = 0$, two and a quarter as bad when $b = 1/2$ and two and a half as bad when $b = 1$.

Amount of Censoring.
Graph 2 presents the results for the parametric Heckman two-step estimator and the semiparametric estimator of the intercept for a range of bandwidth choices and levels of censoring. In this graph as in the graphs which are to follow, only the RMSE ratio and simulated rejection rates are reported since these are the primary tools with which to compare the parametric and semiparametric estimators. The separate graphs for the average bias and standard deviation are available upon request. In part A of the graph, the semiparametric Heckman estimator ($b = 0$) is compared with the parametric Heckman estimator, in part B, the semiparametric Andrews–Schafgans estimator with $b = 1$ is considered.

Although not separately shown, the bias of the parametric Heckman two-step estimator decreases when the amount of censoring decreases, as expected. We also see a decrease in the bias of the semiparametric estimators when the amount of censoring decreases, using a given percentage of the uncensored observations for the semiparametric estimation of the intercept. In addition, the variance increases with the level of censoring, both parametrically and semiparametrically.

The simulation results indicate that the semiparametric estimator improves relative to the parametric Heckman two-step estimator when the amount of censoring increases. When we keep the bandwidth parameter choice in terms of the percentage of uncensored observations used in the estimation of the intercept constant, the RMSE ratio generally decreases as the amount of censoring increases. The range of bandwidth choices for which the RMSE ratio is less than one, therefore increases as the amount of censoring increases when the selection equation errors have a chi-squared distribution. When the selection equation errors have a student t distribution, increasing the amount of censoring increases the feasibility for the semiparametric estimators to do better in terms of RMSE than the parametric Heckman two-step estimator. It appears therefore that the semiparametric estimators are better than the parametric estimator in particular in cases where the inconsistency of the parametric Heckman two-step procedure is most severe. As discussed in Schafgans (1997) this is true for

the chi-squared distribution and for the Student t distribution at high levels of censoring.

For example, in the comparison of the semiparametric Heckman estimator ($b = 0$) with the parametric estimator the following result emerge. When the selection equation errors have a chi-squared distribution with three degrees of freedom and censoring equal to 20 percent, the range of bandwidth choices for which the RMSE ratio is less than one is 35–75 percent, with censoring equal to 50 and 80 percent the range is 10–70 and 10–100 percent respectively. When the degrees of freedom are increased, the range of bandwidth choices for which the RMSE ratio is less than one diminishes and moves to the right, indicating the use of a smaller fraction of the uncensored observations for the estimation of the intercept. The bandwidth ranges for which, in addition, the simulated rejection rate equals five percent indicate that a smaller proportion of the uncensored observations should be used for the semiparametric estimator. In this case, with censoring equal to 20 percent, the range becomes 35–60, with censoring equal to 50 and 80 percent the range becomes 10–35 and 10–25 respectively. When we repeat this comparison for the Andrews-Schafgans estimator with b set equal to 1, the same tendencies are revealed. The only difference is that the range of bandwidths for which the Andrews-Schafgans estimator ($b = 1$) outperforms the parametric two-step Heckman estimator is somewhat wider than the range of bandwidths for which the semiparametric Heckman estimator ($b = 0$) outperforms it.

When the selection equation errors are normal, the Heckman semiparametric estimator ($b = 0$) can at best be 50 percent worse than Heckman's parametric estimator in terms of RMSE when there is 20 percent censoring in the data. With higher amounts of censoring, the semiparametric estimator is relatively worse off. At best, the semiparametric estimator ($b = 0$) can be 70 percent worse than the parametric estimator when censoring is 50 percent, and 90 percent worse when censoring is 80 percent. At a bandwidth choice where the simulated rejection rate equals five percent (using only 10 percent of the uncensored observations), the semiparametric estimators at best do twice as badly as the parametric alternative in terms of RMSE when the amount of censoring is either 20, 50 or 80 percent.

The Selection Index.

In Graph 3, the impact of changing the distribution of the selection index on the semiparametric Heckman estimator can be seen. The two distributions of $X_i'\beta_0$ considered are the standard normal (part A), and the chi-squared distribution with four degrees of freedom (part B). In addition, the graph reveals what impact the availability of the selection equation parameters, β, has on the relative performance of the semiparametric Heckman estimator versus the parametric estimator.

In discussing the impact of changing the distribution of the selection index on the semiparametric Heckman estimator, we restrict our attention to the solid lines (i.e., the more likely case where the selection equation parameters are unknown to the researcher). In the case that

the selection equation errors have a chi-squared distribution, the graph seems to indicate that the range of bandwidth choices for which the semiparametric estimator is better in terms of RMSE than the parametric estimator decreases when the distribution of the selection index is skewed compared to the normal distribution. The range of bandwidth choices for which the estimator is not subject to over-rejection, on the other hand, seems to be invariant to the specific distribution of the selection index. Little or no difference is seen for the other choices of the distribution of the selection equation errors, suggesting that the distribution of the selection index plays a less important role in the selection of the bandwidth parameter than the amount of censoring.

The estimation of the selection parameters has an important impact on the relative performance of the semiparametric Heckman estimator compared to the parametric two-step Heckman estimator. When the selection errors are non-normal, we notice that for any choice of the bandwidth, the RMSE ratio is higher in the case where the selection parameters are unknown compared to the case where they are known. In contrast, when the selection errors are normal, the reverse is true. The impact on the RMSE ratio arising from the estimation of the selection parameters is the strongest for the case where the selection errors have a distribution with thicker tails than the normal.

The explanation of the findings for non-normal choices of the selection errors, lies primarily in the difference in the bias of the estimates when the true coefficients are used rather than the inconsistent preliminary estimates. For non-normal distributions of the selection index, we know (i) that probit estimation will give us inconsistent preliminary estimates, and (ii) that the inverse Mill's ratio is the inappropriate correction for the selection bias. The simulations show that applying the inverse Mill's ratio with the true coefficients generates a bigger bias in the parametric estimation of the intercept than applying the inverse Mill's ratio with the inconsistent parameter estimates. This effect is stronger for the student t distribution where the bias is reduced from -0.14 to -0.02 when we use the inconsistent preliminary estimates rather than the true coefficients compared to the case of the chi-squared distribution where the bias is reduced from -0.20 to -0.15. The semiparametric estimates are much more robust to the estimation of the selection parameters, also because the average derivative estimator applied will give consistent estimates of the selection parameters regardless of the true distribution of the selection equation errors.

For the normal distribution of the selection errors, the parametric two-step Heckman estimator is consistent, and as expected we do not see a large change in the bias as a result of estimation of the selection parameters. More important here, in explaining the downward move of the RMSE ratio curve when preliminary estimates are used rather than the true coefficients, is the relative stronger increase in the variance of the parametric estimator compared to the semiparametric estimators (the variance-covariance matrix of the intercept and the preliminary estimator is block diagonal in the semiparametric case).

The feasibility of the semiparametric estimators to perform better in terms of RMSE than the parametric estimator when the selection errors have a distribution with thicker tails appears more promising in Schafgans (1997). The use of the inconsistent preliminary probit estimates make it that the parametric estimator ends up performing better in terms of RMSE after all. When the selection errors have a student t distribution it is still feasible for the semiparametric estimator to perform better in terms of RMSE as shown in Graph 2, but than only at much higher levels of censoring.

Correlation of the Errors.

In Graph 4, the effect can be seen of varying the correlation of the selection and outcome equation errors. The close link between the covariance of the errors $\sigma_{\varepsilon U}$ and the magnitude of the inconsistency of the parametric Heckman two-step estimator is discussed in Schafgans (1997). In this graph, the correlations of (ε_i, U_i) considered are 1, $1/\sqrt{2}$, $1/2\sqrt{2}$, and 0 with $\sigma_U^2 = 2$. This corresponds to covariances, $\sigma_{\varepsilon U}$, equal to $\sqrt{2}$, 1, 0.5, and 0 respectively. Although not shown separately, to achieve the same bias for the semiparametric estimators with varying correlations of the errors, one needs to use a lower proportion of the uncensored observations for the estimation of the intercept when the correlation increases. The bias of the semiparametric estimators and the parametric estimator is, as expected, zero when the correlation of the errors equals zero. When we evaluate the impact of the correlation of the errors on the comparison of the semiparametric estimators with the parametric Heckman two-step estimator we notice a difference depending on whether the selection equation errors have a more skew distribution or a more thick tailed distribution than the normal. In the first case (ε_i has a chi-squared distribution), the semiparametric estimators perform better when the correlation of the errors increases. The range of bandwidth choices for which the RMSE ratio is less than one increases with the correlation of (ε_i, U_i). The bandwidth choices for which the hypothesis testing criterion is satisfied simultaneously shifts to the right as $\rho_{\varepsilon U}$ rises. This points to the desire to use a lower proportion of the uncensored observations for the estimation of the intercept when the correlation is higher. In the second case (ε_i has a student t distribution), the semiparametric estimator does not improve with higher correlations relative to the parametric Heckman two-step estimator. This result contrasts to findings in Schafgans (1997) and arises from the fact that the simulation results given in Graph 4, incorporate the estimation of the selection parameters whereas they do not in Schafgans (1997). In Graph 5, the simulations results are shown which assume that the selection parameters are known for the case where the selection errors have a student t distribution. From Graph 5 we learn that, the negative impact on the relative performance of the semiparametric estimator vis-a-vis the parametric Heckman estimator due to the preliminary estimates of the selection parameters (also seen in Graph 3) in fact is stronger when the correlation between the selection and outcome equation errors rises.

Sample Size.
In Graph 6, the effect can be seen of using different sample sizes. Two sample sizes, n, are considered: 500 and 1,000. The number of replications is set equal to 1,000 in both instances. Again the results are presented separately for the semiparametric Heckman estimator (panel A) and the Andrews–Schafgans estimator (panel B). In order for the semiparametric estimator to achieve the expected five percent rejection rate of the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, one needs to choose the bandwidth parameter γ_n larger as the sample size increases. Similarly, the range of bandwidth choices for which the RMSE ratio is lower than one, in the case where the selection errors have a chi-squared distribution, lies somewhat to the right for larger sample sizes. As expected, there is little effect on the bias for any given bandwidth. The standard deviation of the estimators naturally decreases as the sample size increases for given bandwidth since the estimator is based on a larger number of observations. The last two results suggest that one should choose the bandwidth higher as the sample size increases. Recall that for consistency the bandwidth parameter needs to approach infinity as $n \to \infty$.

6 Conclusions

In this paper, the finite sample properties of parametric and semiparametric estimates of a censored regression model or sample selection model are compared. The semiparametric estimation techniques of censored regression models, were introduced to deal with the fact that standard estimation techniques for censored regression models based on incorrect distributional assumption were inconsistent. Of concern, however, remains their finite sample performance especially due to the presumable sensitivity of the semiparametric techniques to a bandwidth or smoothing parameter. We focus in particular on estimation of the intercept of the 'outcome' equation of this model when Heckman's normality assumption does not hold.

The finite sample properties of the parametric and semiparametric estimates are compared over a wide range of bandwidth choices, in the hope that the simulations can assist the empirical researcher in choosing a bandwidth or smoothing parameter to 'optimize' the finite sample performance of the semiparametric estimator relative to that of the parametric estimator. The results we find are: (1) For symmetric *distributions of the selection equation errors*, one can only expect to do better in terms of RMSE than the parametric alternative when the tails of the selection equation errors are very thick relative to the normal distribution and *censoring* is very large. The bandwidth should be chosen in such a way that only a small proportion of the uncensored observations are used for the estimation of μ_0. For skew distributions of the selection equation errors, the semiparametric estimator is better than the parametric Heckman two-step estimator for a large range of bandwidths and levels of censoring. The estimation of μ_0 should be based on a smaller proportion of the uncensored observations when the amount of censoring increases, more so the further away the distribu-

tion of the selection equation errors is from the normal. (2) The higher the *correlation of the errors* the larger the range of bandwidth choices for which the RMSE of the semiparametric estimator outperforms the parametric Heckman two-step estimator when the selection errors are skewed. In order for the estimator not to over-reject, however, one needs to base the estimation of μ_0 on a smaller proportion of the uncensored observations when the correlation of the errors is larger. (3) The improvement of the semiparametric estimator over the parametric estimator when the selection errors are skewed holds for a variety of *distributions of* $X_i'\beta_0$. The bandwidth choices for which the simulated rejection rate of the null hypothesis equals five percent however seem to vary little with the actual distribution of $X_i'\beta_0$. And (4) The larger the *sample size* the smaller the proportion of uncensored observations that should be used in the estimation of μ_0.

For any application, however, an empirical researcher only has the actual sample size and the amount of censoring to his or her avail. Given that the bivariate normal distribution underlying the MLE and the parametric Heckman two-step estimator has been rejected by our empirical researcher (e.g., Lee (1984)), in order to choose the 'best' bandwidth for the semiparametric estimator, the researcher will need to do some model pretesting to form an idea about the 'true' distribution functions of $X_i'\beta_0$ and ε_i and the correlation between the selection and outcome equation errors, $\rho_{\varepsilon U}$. One simple approach would be as follows. First, one estimates the selection index $X_i'\widehat{\beta}$ using a semiparametric estimation technique, e.g., Klein and Spady (1993). Inference about the actual distribution of the selection index, $X_i'\beta_0$ could be made by evaluating the statistical properties of this consistently estimated selection index, e.g., variance, skewness, kurtosis, and upper tail index. Second, one estimates the generalized residuals of the selection equation, $\widehat{\varepsilon_i} \equiv D_i - \widehat{E}(D_i|X_i'\widehat{\beta})$, where $\widehat{E}(D_i|X_i'\widehat{\beta})$ is a nonparametrically estimated conditional mean function. Similar calculations on these residuals would give the researcher an indication of the actual distribution of the selection equation errors. To obtain a consistent estimate of the correlation between the selection equation errors and the outcome equation errors one could consider imposing Olsen's (1980) class of dependence models: $U_i = \sigma_{\varepsilon U}\,\varepsilon_i + V_i$, where ε_i and V_i are independent, and estimate the appropriate second Heckman step. Naturally, more elaborate and robust model pretesting could be suggested.

The importance of the simulation study is to realize that it is *feasible* to select a bandwidth parameter at which the semiparametric estimator outperforms the Heckman's two-step estimator, and that its feasibility is particularly revealed in cases where the inconsistency of the Heckman's two-step estimator under non-normality is most severe. For error distributions that are close to the normal, on the other hand, the two-step parametric estimator performs better.

Regarding the relative performance of the semiparametric Heckman (1990) estimator and the Andrews and Schafgans (1998) estimator the simulation results show that the variance

of the Andrews Schafgans estimator increases with the choice of b, for a given bandwidth parameter γ_n; the Heckman estimator ($b = 0$) would appear to be more efficient. Nevertheless, the bias of the Andrews-Schafgans estimator is lower in absolute value than the Heckman estimator. Overall, our simulations tend to favour the finite sampling behaviour of the Andrews-Schafgans (1998) estimator over the Heckman estimator (1990). This is in line with like findings in other nonparametric estimation problems, which show that the trade-off between bias and variance is typically better for smooth "kernels".

References

Amemiya, T. (1984): "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61.

_____ (1985): *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.

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.

Arabmazar, A. and P. Schmidt (1981): "Further Evidence on the Robustness of the Tobit Estimator to Heteroskedasticity," *Journal of Econometrics*, 17, 253–258.

_____ (1982): "An Investigation of the Robustness of the Tobit Estimator to Non-Normality," *Econometrica*, 50, 1055–1063.

Best, D.J. and D.E. Roberts (1975): "Algorithm AS 91: The percentage points of the χ^2 distribution," *Applied Statistics*, 24, 385-388.

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

Goldberger, A.S. (1983): "Abnormal Selection Bias," in S. Karlin, T. Amemiya, and L.A. Goodman (eds.), *Studies in Econometrics, Time Series and Multivariate Statistics*, New York: Wiley.

Greene, W.H (1993): *Econometric Analysis*. New York: Macmillan Publishing Company.

Gronau, R. (1974): "Wage Comparisons: A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

Heckman, J.J. (1974): "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42, 679–694.

_____ (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

_____ (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47. 153–161.

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

Hill, G.W. (1970): "Algorithm 395: Student's t distribution," *Communications of the ACM*, 13, 617-619.

Hurd, M. (1979): "Estimation in Truncated Samples when there is Heteroscedasticity," *Journal of Econometrics*, 11, 247–258.

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

Jarque, C.M. and A.K. Bera (1987): "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, 55, 163–172.

Johnson, N.L. and S. Kotz (1970): *Continuous Univariate Distributions I*, Boston: Houghton Mifflin.

Kalbfleisch, J.G. and R.L. Prentice (1980): *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Kennedy, W.J. and J.E. Gentle (1980): *Statistical Computing.* New York: Marcel Dekker, Inc..

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*, 49 , 355–372.

_____ (1984): "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity," *Econometrica*, 52, 843–863.

Lewis, H.G. (1974): "Comments on Selectivity Biases in Wage Comparisons," *Journal of Political Economy*, 82, 1145–1155.

Maddala, G.S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Manski, C.F. (1993): "The Selection Problem in Econometrics and Statistics," in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.), *Handbook of Statistics*, Vol. 11, North-Holland: Elsevier Science Publisher B.V.

Olsen, R.J. (1980): "A Least Squares Correction for Selectivity Bias," *Econometrica*, 48, 1815–1820.

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

Rubin, D. (1987): *Multiple Imputation for Nonresponse in Surveys*. New-York: Wiley.

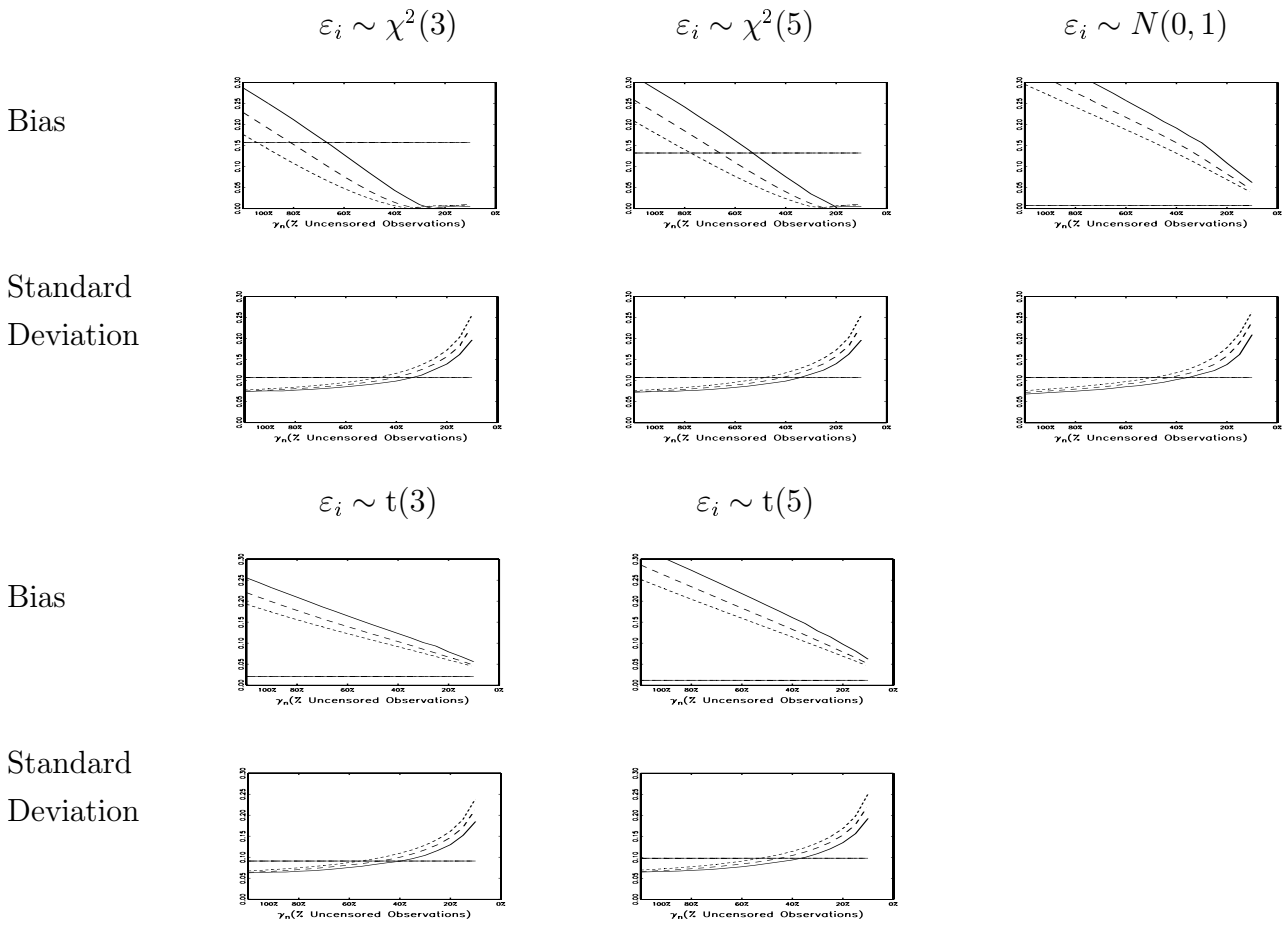
Schafgans, M.M.A. (1997): "Semiparametric Estimation of a Sample Selection Model: A Simulation Study," Sticerd Discussion Paper No. EM/97/326, London School of Economics and Political Science.

Schafgans, M.M.A. and V. Zinde-Walsh (2000): "On Intercept Estimation in the Sample Selection Model," Sticerd Discussion Paper No. EM/00/380, London School of Economics and Political Science.

Graph 1

The semiparametric Heckman estimator ($b = 0$) and the Andrews–Schafgans estimator vis-a-vis the parametric two-step Heckman estimator.
A comparison using various distributions for the selection errors, ε_i.

A. Absolute bias and standard deviation.

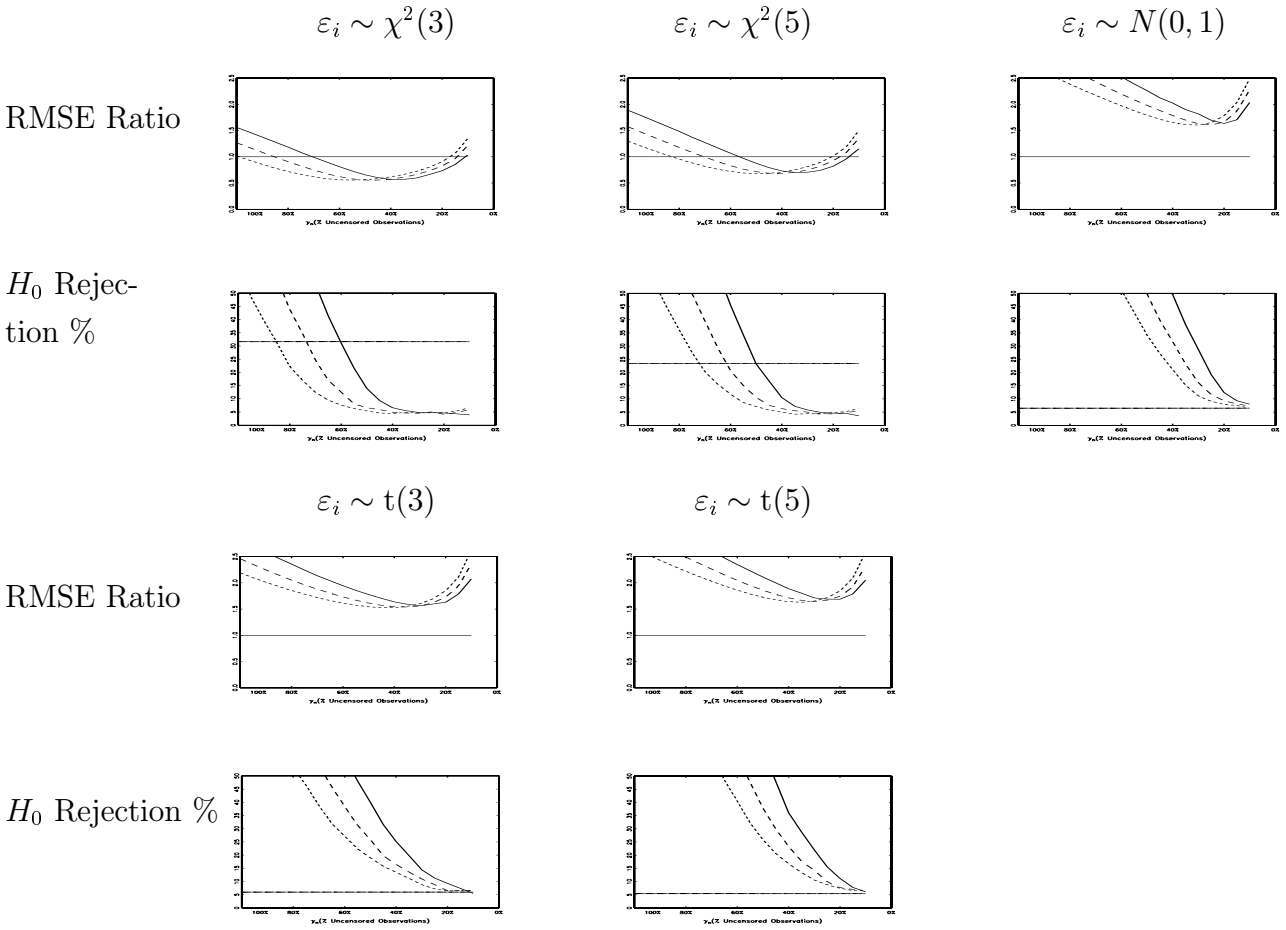


..... $b = 1$ _ _ _ _ $b = 0.5$ ____ $b = 0$

Notes: (1) The bandwidth parameter of the semiparametric estimator is given by γ_n. All horizontal lines correspond, therefore, to the parametric Heckman two-step estimator, the remaining lines to the semiparametric estimators. (2) The average bias and standard deviation of the two estimators over 1,000 replications are given. The simulation results are based on 1,000 observations (3) The distribution of the selection index $X'\beta_0$ is assumed to be $N(0,1)$, and the correlation between the selection and outcome equation errors ($\rho_{\varepsilon U}$) is equal to $1/\sqrt{2}$. (4) Censoring is 50%.

Graph 1 (Continued)

B. RMSE and rejection rates of the null hypothesis $\mu = \mu_0$ against $\mu \neq \mu_0$.

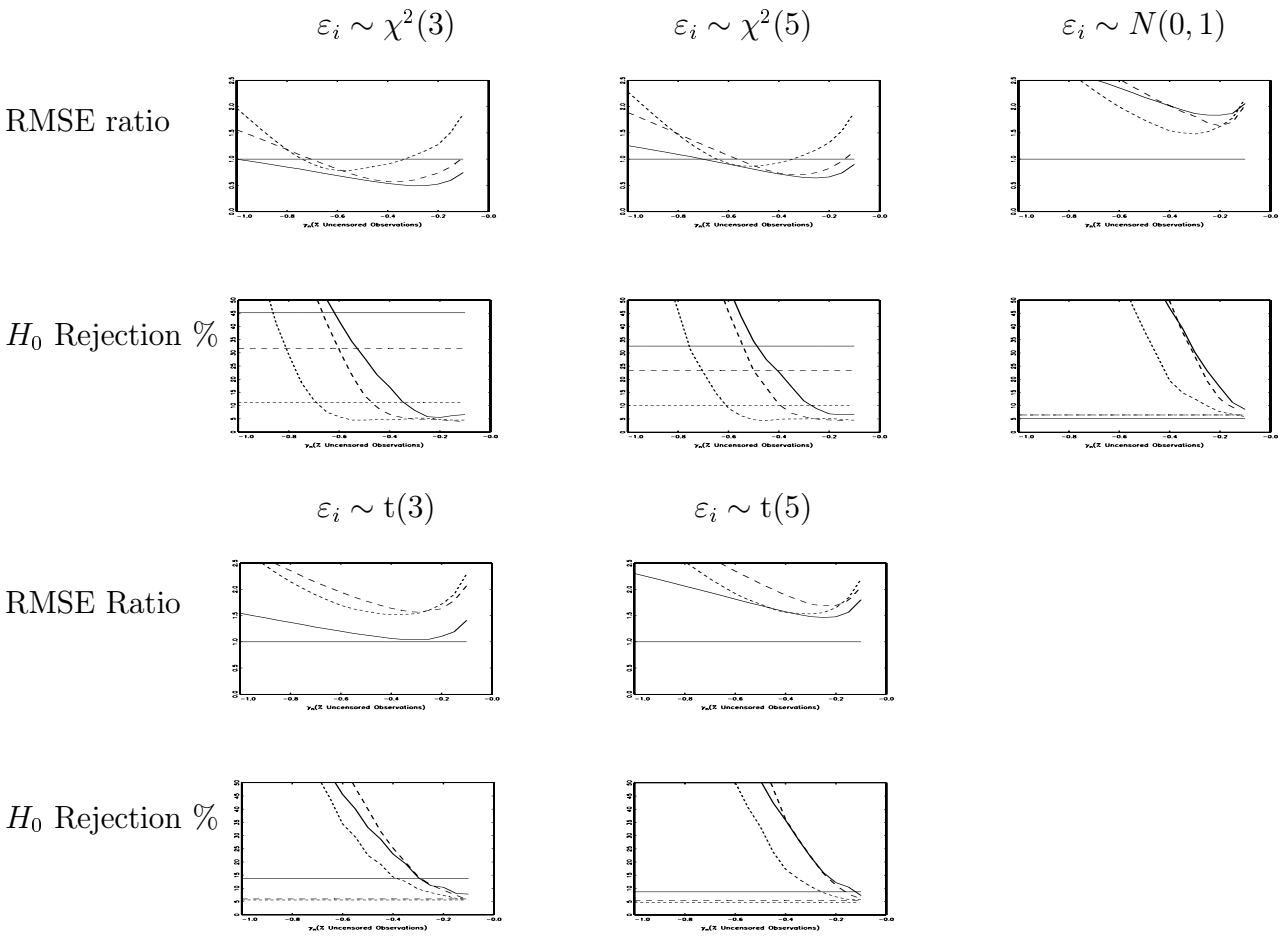


Notes: (1) The RMSE is the abbreviation of the root mean squared error. The RMSE ratio is defined by the RMSE of the semiparametric estimator over the RMSE of the parametric Heckman two-step estimator. (2) The H_0 rejection rates are the simulated rejection rates of the t test for the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the five percent level of significance. The actual bandwidth chosen for the estimation of μ_0 should coincide with a simulated probability of rejection of the null hypothesis equal to five percent.

Graph 2

The semiparametric estimators and the parametric two-step Heckman estimator.
A comparison by various distributions for the selection errors, ε_i, and
the amount of censoring.

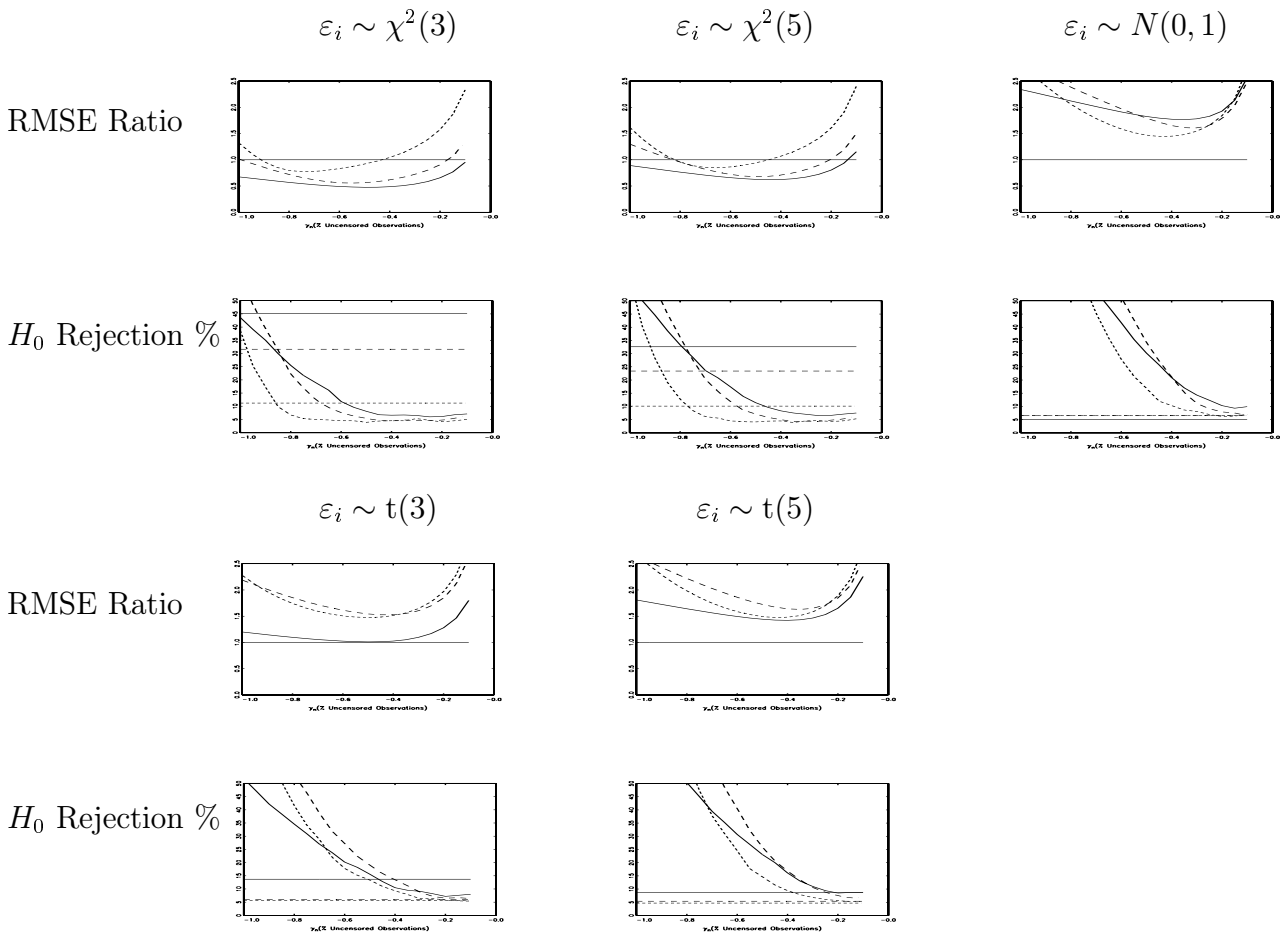
A. The semiparametric Heckman estimator $(b = 0)$



Note: The distribution of the selection index $X'\beta_0$ is assumed to be $N(0,1)$, and the correlation between the selection and outcome equation errors ($\rho_{\varepsilon U}$) is equal to $1/\sqrt{2}$. The amount of censoring defined to equal $\Pr(X'\beta_0 < c)$ is obtained by choosing an appropriate c.

Graph 2 (Continued)

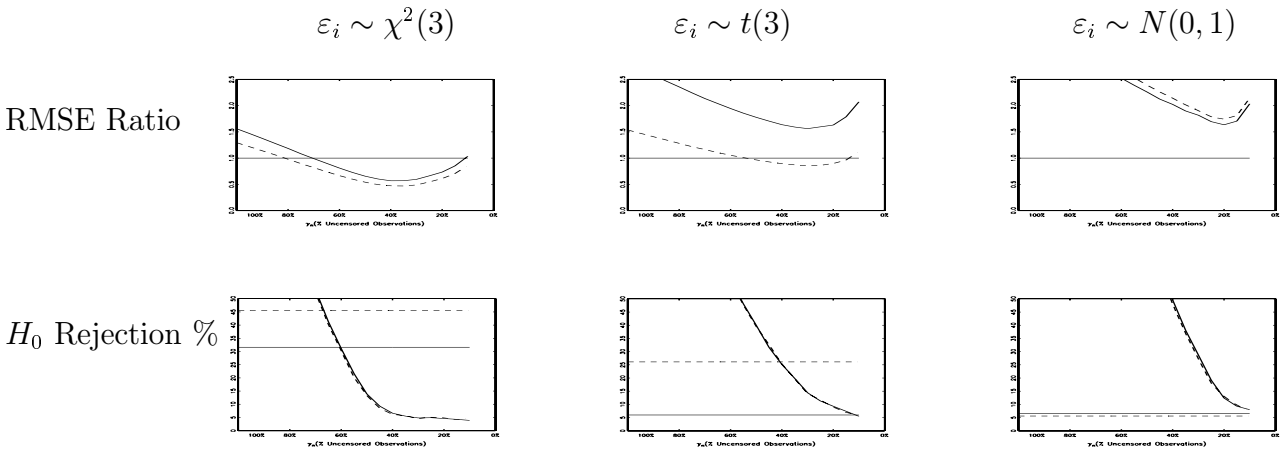
B. the Andrews–Schafgans estimator ($b = 1$)



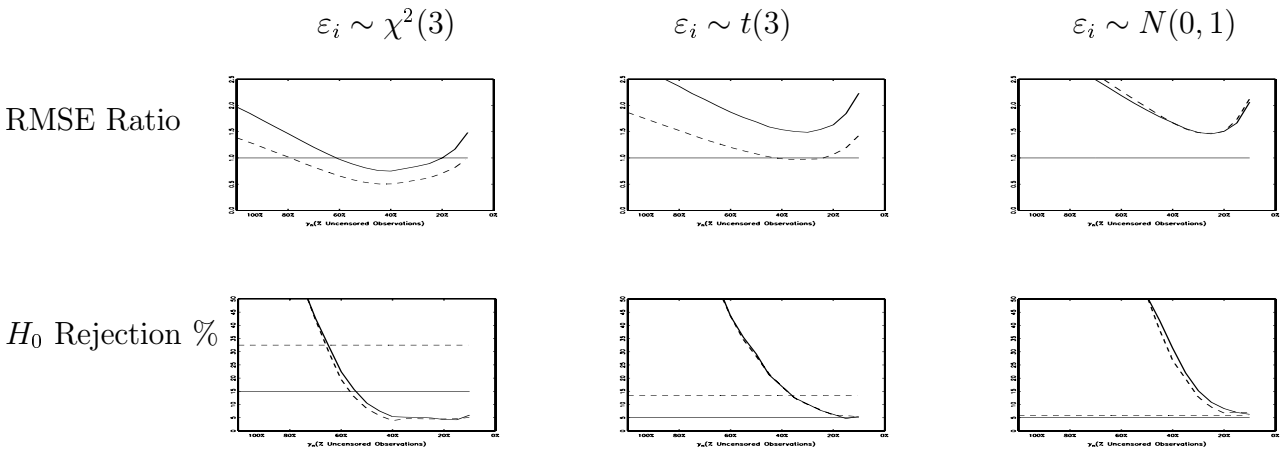
Graph 3

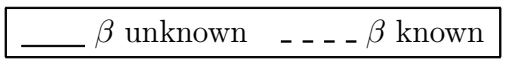
The semiparametric estimator ($b = 0$) and the parametric two-step Heckman estimator.
A comparison by selection index distribution and availability of the selection parameters.

A. Distribution selection index: $X'\beta_0 \sim N(0,1)$.



B. Distribution selection index: $X'\beta_0 \sim \chi^2(4)$.



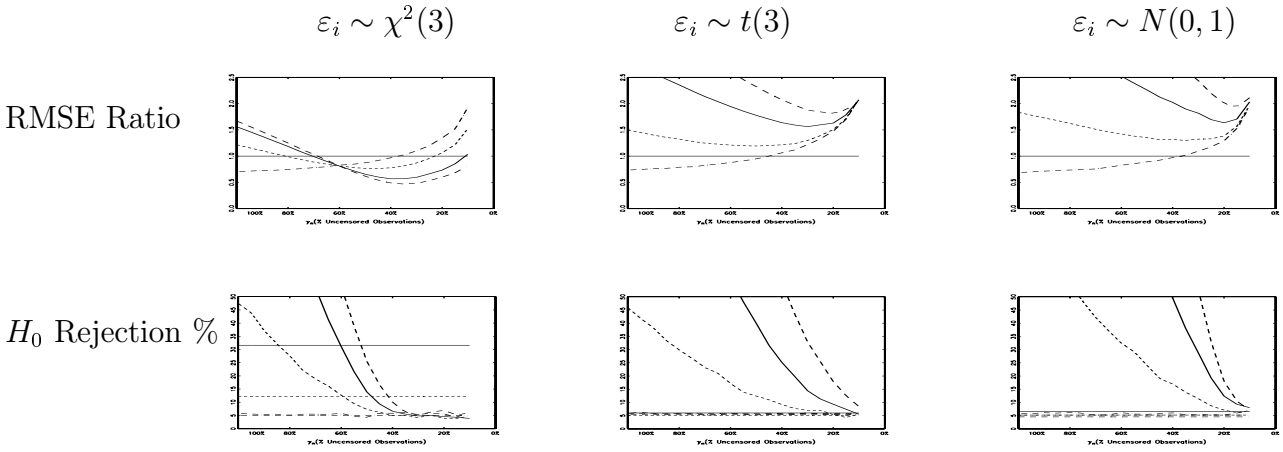


Note: The selection equation errors are denoted by ε_i and β are the selection parameters. The correlation between the selection and outcome equation errors ($\rho_{\varepsilon U}$) is equal to $1/\sqrt{2}$, and the amount of censoring is 50%.

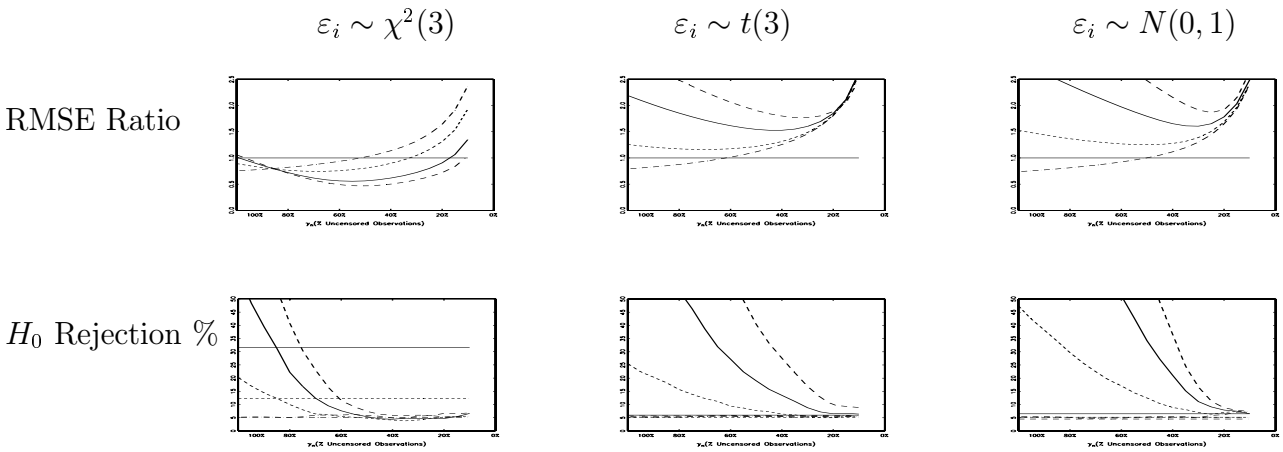
Graph 4

The semiparametric estimators and the parametric two-step Heckman estimator.
A comparison by correlation of the errors and distribution of the selection errors.

A. The semiparametric Heckman estimator ($b = 0$)



B. the Andrews–Schafgans estimator ($b = 1$)



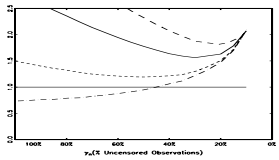
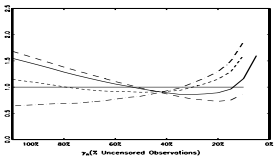
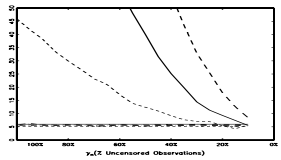
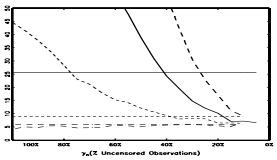
$\_.\_.\_.\_$ $\rho_{\varepsilon U} = 0$ ..... $\rho_{\varepsilon U} = 1/2\sqrt{2}$ ____ $\rho_{\varepsilon U} = 1/\sqrt{2}$ $- - - -$ $\rho_{\varepsilon U} = 1$

Note: The selection equation errors are denoted by ε_i and $\rho_{\varepsilon U}$ is the correlation between the selection and outcome equation errors. The distribution of the selection index $X'\beta_0$ is assumed to be $N(0,1)$, and the amount of censoring is 50%.

Graph 5

The semiparametric estimator ($b = 0$) and the parametric two-step Heckman estimator.
A comparison by $\rho_{\varepsilon U}$ and the availability of the selection parameters, β.

Distribution selection equation error, $\varepsilon_i \sim$ t(3)**.**

	β unknown	β known
RMSE Ratio		
H_0 Rejection %		

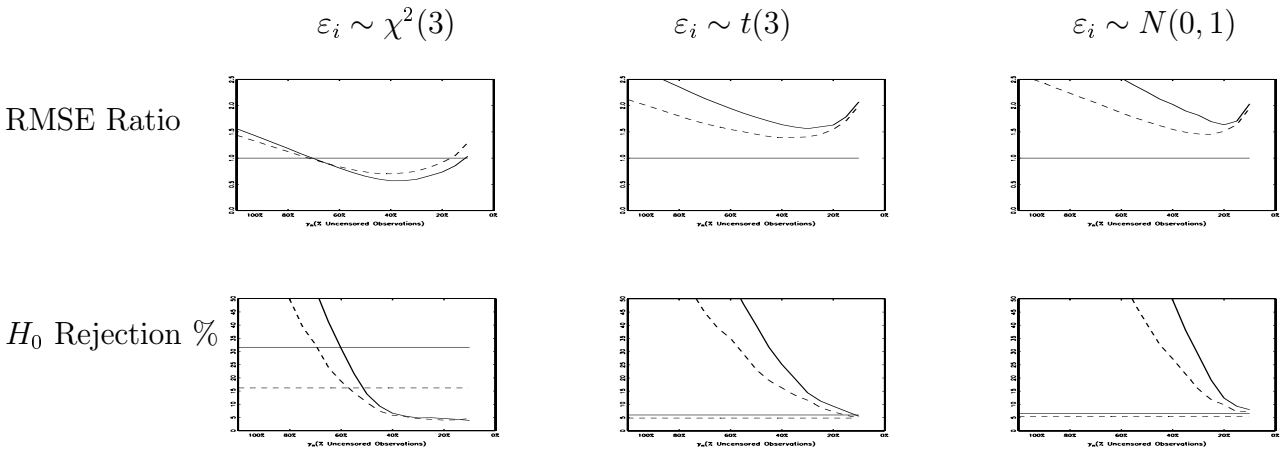
$\_.\_.\_.\_$ $\rho_{\varepsilon U} = 0$ ..... $\rho_{\varepsilon U} = 1/2\sqrt{2}$ $\_\_\_\_$ $\rho_{\varepsilon U} = 1/\sqrt{2}$ $\_\ \_\ \_\ \_$ $\rho_{\varepsilon U} = 1$

Note: The distribution of the selection index $X'\beta_0$ is assumed to be $N(0, 1)$, and the amount of censoring is 50%.

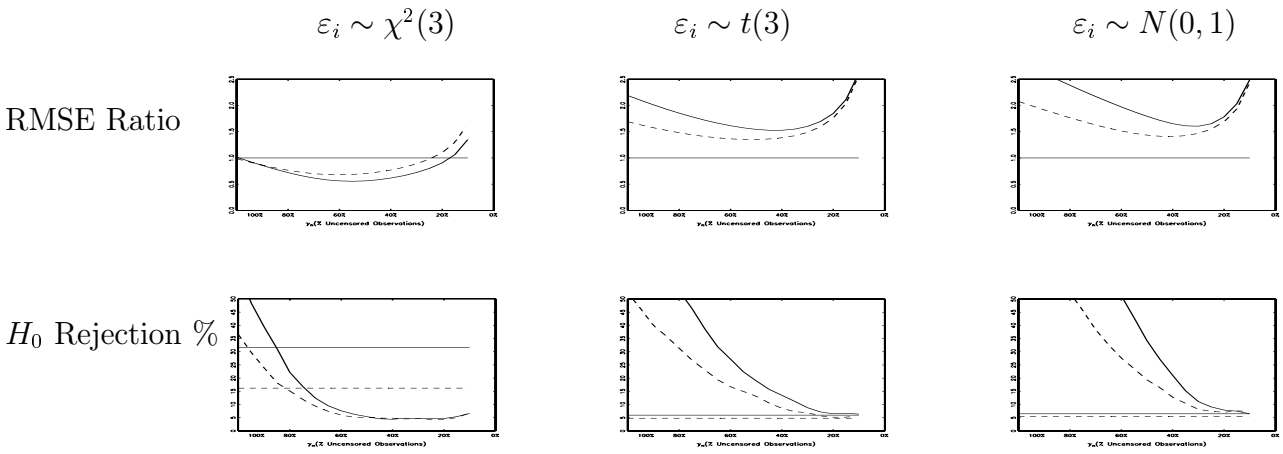
Graph 6

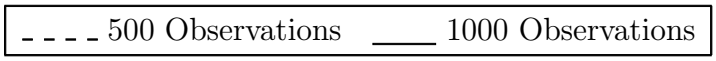
The semiparametric estimators and the parametric two-step Heckman estimator.
A comparison by sample size and distribution of the selection errors, ε_i.

A. The semiparametric Heckman estimator $(b = 0)$



B. the Andrews–Schafgans estimator $(b = 1)$





Note: The distribution of the selection index $X'\beta_0$ is assumed to be $N(0,1)$, and the correlation between the selection and outcome equation errors ($\rho_{\varepsilon U}$) is equal to $1/\sqrt{2}$. The amount of censoring is 50%.