Sample selection bias and the South African wage function

COBUS BURGER

Stellenbosch Economic Working Papers: 18/08

KEYWORDS: SEMIPARAMETRIC AND NONPARAMETRIC METHODS; SIMULATION METHODS; TRUNCATED AND CENSORED MODELS; LABOR FORCE AND EMPLOYMENT, SIZE, AND STRUCTURE
JEL: C14, C15, C34, J21

COBUS BURGER
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: COBUSBURGER@SUN.AC.ZA

UNIVERSITEIT STELLENBOSCH UNIVERSITY

BER
BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

Sample selection bias and the South African wage function

COBUS BURGER[1]

ABSTRACT

Conventional wage analyses suffers from a debilitating ailment: since there are no observable market wages for individuals who do not work, findings are limited to the sample of the population that are employed. Due to the problem of sample selection bias, using this subsample of working individuals to draw conclusions for the entire population will lead to inconsistent estimates. Remedial procedures have been developed to address this issue. Unfortunately, these models strongly rely on the assumed parametric distribution of the unobservable residuals as well as the existence of an exclusion restriction, delivering biased estimates if either of these assumptions is violated. This has given rise to a recent interest in semi-parametric estimation methods that do not make any distributional assumptions and are thus less sensitive to deviations from normality. This paper will investigate a few proposed solutions to the sample selection problem in an attempt to identify the best model of earnings for South African data.

Keywords: Semiparametric and nonparametric methods; Simulation methods; Truncated and censored models; Labour force and employment, Size, and structure
JEL codes: C14, C15, C34, J21

[1] The author is indebted to Servaas van der Berg (SU) and Rulof Burger (SU) for their comments and advice. Any errors are the sole responsibility of the author.

Following the seminal article by Gronau (1974) it is now widely agreed that conventional wage analyses suffer from a debilitating ailment: since there are no observable market wages for individuals who do not work, findings are limited to the sample of the population that are employed. Due to the problem of sample selection bias, using this subsample of working individuals to draw conclusions for the entire population will lead to inconsistent estimates. Some remedial procedures that correct for this bias have been developed by Heckman (1974, 1979). Unfortunately, these models strongly rely on the assumed parametric distribution of the unobservable residuals as well as the existence of an exclusion restriction, delivering biased estimates if either of these assumptions is violated. This has given rise to a recent interest in semi-parametric estimation methods that do not make any distributional assumptions and are thus less sensitive to deviations from normality. This paper will investigate a few of these proposed solutions to the sample selection problem in an attempt to identify the best model of earnings for South African data.

The next section introduces the sample selection problem. Section 2 builds on this discussion by providing a formal model that fits the intuitive problem and discussing and assessing the two most popular sample selection models. Following this, an alternative, but less popular, sample selection model that is less dependent on the parametric assumptions of the residual, is proposed. Section 3 adds to this discussion, by testing the empirical validity of the competing models, using Monte Carlo simulations. In section 4 the models are applied to South African LFS dataset and compared. Section 5 concludes.

1. SAMPLE SELECTION BIAS

The issue of sample selection bias commonly arises not just within econometric subject matters, but also within other social sciences. In this paper, however, the focus is restricted to the effect of sample selection on the wage function, the same framework within which the problem was originally identified by Gronau (1974). Gronau's model showed that the use of an OLS regression that is confined to a certain portion of society to draw inference over the entire population would be flawed if the first group is not a random selection from the population.

While one can control for the effect of the observable characteristics by including these variables in the wage function, this is not the case for unobservable characteristics, like ambition and motivation. Unfortunately, these variables are likely to play an important role both in determining whether one would acquire a job and the wage one ultimately receive. If this is the case, conventional wage functions fail to incorporate the role that unobservable attributes could have on the outcome equation. These models would be susceptible to inconsistent estimators and misleading t-statistics, which in turn may lead to improper results and conclusions.

2. MODEL AND METHODOLOGY

a) The Sample Selection Model

We assume that each individual has a set of characteristics that is specific to him or her. Empirically, it is important to distinguish between features that are observable and those that are not. In terms of the observable attributes, it is assumed that some of these characteristics determine an individual's productivity, x_{2i}, while others may influence that individual's likelihood of attaining work, x_{1i}. The two sets of variables, x_{1i} and x_{2i}, are allowed to overlap (Wooldridge, 2002: 561). The error terms are often conceptualised as representing, or at least including, the unobserved productive characteristic, like drive and intelligence, which are important in determining both employment and wages. Failure of the model to control for these unobservable variables will cause the errors to be correlated and lead to sample selection bias.

Algebraically, the model can be presented as follows:

stage 1:

$$\begin{aligned} d_i^* &= \alpha x_{1i} + e && \text{(selection equation)} \\ d_i &= 1 \quad \text{if } d_i^* \geq 0 \\ d_i &= 0 \quad \text{if } d_i^* < 0 \end{aligned}$$

stage 2:

$$\begin{aligned} y_i^* &= \beta x_{2i} + u && \text{(outcome equation)} \\ y_i &= y_i^* \qquad \text{if } d_i = 1 \\ y_i & \text{ is missing} \quad \text{if } d_i = 0 \end{aligned}$$

where d_i and y_i are the observed realisations, e.g. of employment and wages
d_i^* and y_i^* are their latent counterparts,
x_1 and x_2 are vectors of exogenous variables,
α and β are unknown parameter vectors and
e and u are the corresponding error terms

In the above model, the outcome variable, y, which denotes log of wages, is only observable when some criteria defined in terms of d are met. In our case, d will signify the employment outcome, attaining a value of one if the individual is employed and zero if the individual is not employed. The selection equation is modelled in the first stage. In the second stage, the wage function is estimated by regressing y on a set explanatory variables, x_2, conditional on $d = 1$.

The correlation coefficient between the errors, ϱ, can be interpreted as an indication of the relationship between the unobservable characteristics within the first and second step. The problem of sample selection arises when the errors of the selection equation and the errors of the wage function are correlated, i.e. if $\varrho \neq 0$. If this is the case, simply regressing y on x over the subsample of employed individuals, using standard ordinary least squared estimates will deliver biased estimates of *β*, since it fails to incorporate the relationship between *e* and *u*. The sample selection literature has emerged due to the need to correct for this bias. The two most popular proposed fixes for the problem are the Heckman maximum likelihood estimator method (ML) and the Heckman two-step estimation procedure.

b) Heckman's maximum likelihood estimator

The maximum likelihood estimator (Heckman, 1974) diverges from the method of least squares, by using a likelihood function rather than a probability function to estimate parameters and by assuming that the residuals are bivariate normally distributed[2]. While this model has been shown to produce consistent estimates under a few plausible conditions and normal and efficient estimates if sample sizes are large enough, the distributional assumptions required to justify the use of the maximum likelihood estimator are no less stringent than is required of OLS: With OLS estimation the non-normality assumption is only required to ensure the efficiency of the OLS estimates, but is not required to ensure their consistency, whereas with the ML estimators the *β*'s are generally neither consistent nor efficient under an incorrect distributional assumption.

c) Heckman's two-step estimator

One major drawback of Heckman's maximum likelihood estimator is its procedural complexity and the added computation needed to solve these ML estimates. In response to this critique, Heckman (1979) developed the two-step estimator, a simpler version of his own ML method that could be solved using the more familiar probit function and a conventional OLS regression. This two-step model makes use of a correction term, called the inverse Mills ratio, to correct for any sample selection bias that may have crept into the OLS model.

Using some clever arithmetic, Heckman showed that the unbiased expected value of y conditional on *d* = 1 consists of two components: the first contains the conventional regressors, which one would have used in simple subsample OLS regression, while the second contains a term that can be used to correct for the bias. The inverse Mills ratio forms part of this correction term.

$$\begin{aligned} E(y_i) &= E(y_i{*} \mid d_i=1) \\ &= E(\beta x_{2i} + u \mid d_i=1) \\ &= \beta x_{2i} + E(u \mid d_i=1) \\ &= \beta x_{2i} + E(u \mid d_i{*} > 0) \end{aligned}$$

[2] Formally, this would imply that both error terms *u* and *e* are normally distributed, with mean zero, constant variances σ_u^2, σ_e^2 and correlation ϱ.

$$= \beta x_{2i} + E(u \mid \alpha x_{1i} + e > 0)$$

$$= \beta x_{2i} + E(u \mid e > - \alpha x_{1i})$$

If one goes further and assumes that *e* and *u* are jointly normally distributed (i.e. $\begin{bmatrix} e \\ u \end{bmatrix} \sim BN\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix}\right]$) then it follows that $u = \frac{\sigma_{12}}{\sigma_{11}^2} e + v$, where the first term, $\frac{\sigma_{12}}{\sigma_{11}^2} e$, is correlated with *e* and the second term, *v,* is not.

Adding this to the prior model one attains:

$$E(y_i) = \beta x_{2i} + E(\frac{\sigma_{12}}{\sigma_{11}^2} e + v \mid e > - \alpha x_1)$$

$$= \beta x_{2i} + E(\frac{\sigma_{12}}{\sigma_{11}^2} e \mid e > - \alpha x_1)$$

$$= \beta x_{2i} + \frac{\sigma_{12}}{\sigma_{11}} E(\frac{e}{\sigma_{11}} \mid \frac{e}{\sigma_{11}} > - \alpha x_1)$$

$$= \beta x_{2i} + \frac{\sigma_{12}}{\sigma_{11}} \frac{\phi(-\alpha x_1)}{1-\Phi(-\alpha x_1)}$$

$$= \beta x_{2i} + \frac{\sigma_{12}}{\sigma_{11}} \lambda(- a x_1)$$

where ϕ denotes the standard normal density function

and Φ denotes the cumulative distribution function.

From the above derivation it is clear that a OLS wage function that neglects to include the second term will deliver biased estimates of β whenever $\frac{\sigma_{12}}{\sigma_{11}}$ is not equal to zero. As a result, Heckman (1979) defined the sample selection problem as being a special case of the more general omitted variable problem, with $\lambda(.) = \phi(ax_1)/1-\Phi(ax_1)$ being the omitted variable.

He showed that the problem can be overcome by adding the inverse Mills ratio attained in the selection equation as an additional regressor in the outcome equation. This inverse Mills ratio is usually derived by probit model, which estimates the likelihood of employment given a host of observable characters and a normally distributed error. The linear prediction of the fitted probit model, ax_1, should then be added to the wage function as an additional regressor. When the function is then solved using conventional OLS analysis, the inverse Mills ratio coefficient can be regarded as an estimate of $\frac{\sigma_{12}}{\sigma_{11}}$. (Johnston & DiNardo: 449)

Despite the ingenuity and simplicity of the two-step model, Davidson & MacKinnon (1984: 252) warn that it is still inferior to the ML counterpart, since it provides inefficient results. Unlike the two-step method that solves the selection equation and outcome equation in turn, the ML method solves the selection and outcome equations simultaneously. The authors recommend that the two-step Heckman only be used to test for the degree of selection bias, where after the ML method should be applied if the selection bias is significant and a conventional OLS should be applied if the selection bias is not significant.

d) Concerns regarding sample selection models

The popularity of the sample selection models introduced in the previous section has grown immensely since the 1970s. So much so, that sample selection procedures nowadays come standard with many software programmes, helping to lower technical capabilities required for applying these techniques. While the wider use of these models has its benefits, they should not be applied indiscriminately. According to Johnston and DiNardo (2004: 450) sample selection methods are often sensitive to a range of factors, like the presence of heteroscedasticity, the degree of identification and the validity of the distributional assumptions. With this in mind, even Heckman (1990: 317), recognises that simpler estimation methods may be just as good in answering economic questions under certain circumstances.

The problem of identification arises since the set of explanatory variables in the wage function, x_1, and the set of explanatory variables in the selection equation, x_2, tend to overlap and in many cases are even identical. According to Puhani (2000: 57), failure to include exclusion restrictions - regressors that are unique to the selection function - may lead to colinearity problems. Since, in the absence of exclusion restriction, the outcome equation is identified through the nonlinearity of the inverse Mills ratio alone, a function which has been shown to be quasi-linear for a large section of its argument. As a result, these models run the risk of obtaining unreliable *β's* and inflated standard errors. According to Berk and Ray (1982: 386), the identification problem is worsened when the variation of the selection outcome is not properly explained by its regressors, since in this case, the inverse Mills ratio will have little variance and the effect on the outcome equation will be minimal.

Given these difficulties, it should greatly aid identification if the selection equation contains a variable which does not also appear in the wage function. This would induce variation in the inverse Mills ratio, not already contained in the wage regressors, and in doing so allow the inclusion of this variable to absorb the sample selection bias. Unfortunately, this is easier said than done. In practice, due to the problem of omitted variable bias and the complexity of human behaviour, it is often difficult to identify variables that are correlated with the selection without also being correlated with wages.

Questions have also peen posed regarding the validity of the distributional assumptions required of the ML and two-step models. Although the normality assumption allows us to solve these models, it has the unfavourable effect of making estimates overly dependent on the distribution of the residuals. As a result, both models will produce inconsistent parameter estimates if normality fails.

e) Semi-parametric estimator

The problem of non-normality can be addressed in two manners. One method, which was proposed by Lee (1982, 1983), is to transform the random elements in the model so that they can be represented by the bivariate normal distribution. This method however requires knowledge of the marginal distribution of the selection equation's residuals. Alternatively, the reliance on distributional assumption can be avoided by making use of the general estimation strategy proposed by Gallant and Nychka (1987). This semi-parametric method approximates the unknown density of the residuals in the selection equation using a Hermite form.

Stewart (2004) followed an extension of this semi-parametric (SP) method to develop a semi-parametric approximation of the ordered probit function.[3] According to this method, the density distribution of the errors can be attained by multiplying a squared polynomial with a normal density distribution, as is done below.

$$f_K(e) = \frac{(\sum_{k=0}^{K} \alpha_{1k} e^k)^3 \phi(e)}{\int_{-\infty}^{\infty} (\sum_{k=0}^{K} \alpha_{1k} e^k)^3 \phi(e) de}$$

where e is an error term

K specifies the order of the Hermite polynomial,

ϕ (.) is the standard normal density distribution,

and α_{1k} is the estimated parameters of the polynomial function

The second difficulty is to derive the function g(.), which makes use of the index restriction, ax_1, to use in the conditional expectation of the outcome equation.

$E(y \mid d=1) = \beta x_2 + g(ax_1)$

[3] The Stata ado file which was written by M. B. Stewart can be attained from the Stata Journal website at the following address: http://www.stata-journal.com/software/sj4-1/st0056.pkg

The conventional two-step inverse Mills ratio cannot be used here, since the function makes use of the parametric assumption, i.e. normality. Several semi-parametric alternatives have been developed to approximate $g(ax_1)$. Heckman and Robb (1985) made use of a Fourier expansion around the probability that a person is employed, Costlett (1983) used intervals and indicator variables to approximate g(.) and Newey (1988) estimated the selection correction, $g(ax_1)$, using an initial estimate of ax_1 and an approximation series which was allowed to grow as the sample size increased.

In this paper the author will employ an iterative technique suggested by Ichimura and Lee (1991) to estimate g(.). The semi-parametric procedure is derived from the following two identities, that define the relationship between β, ax_1 and g(.):

$$E(y \mid d=1, ax_1) = \beta x_{2i} + g(ax_1)$$
$$E(y - \beta x_2 \mid ax_1) = g(ax_1)$$

An estimation of $g(ax_1)$ can be obtained by inserting the estimate of ax we obtained from the semi-parametric probit and a preliminary estimate of β into the second equation. The estimate of $g(ax_1)$ is then inserted into the first equation to derive a new approximation of β. This new estimate of β can now be used to derive another estimate of $g(ax_1)$, which in turn can be used to derive a new β. The iterative process is repeated until the estimated values of β converge. Ichimura and Lee showed that the estimated parameters that one obtains through this method are consistent and asymptotically normal. In essence, the semi-parametric method is an augmentation of the standard two-step Heckman model, the main difference being that the augmented model uses a semi-parametric binary function in place of the conventional parametric probit function and an iterative approximation process rather than a conventional OLS regression.

3. Monte Carlo Simulations

The importance of including an identification variable, a variable that is unique to the selection equation, has been hotly debated among statisticians and economists. While some downplay its importance, others claim that two-step methods that do not contain adequate exclusion restrictions are inherently flawed. The discussion has benefited from insights gained through the use of Monte Carlo simulations. Two studies that are widely cited in this regard are those of Nelson (1984) and Stolzenberg and Relles (1990), who showed that the bias and precision of the sample selection models are heavily dependent on the following three factors:

- the value of ϱ (denoting the correlation between the two error terms, e and u);
- the correlation between the explanatory variables in the selection equation, x_1, and the outcome equation, x_2, denoted by θ; and
- the degree of censoring (i.e. the proportion of the working age population that is not employed, in the case of a wage equation), denoting the degree of identification.

a) The Model

Using a similar model as that of Stolzenberg and Relles (1990) we replicated their Monte Carlo simulations for a specific range of parameters that correspond to the South African labour force data.

The following equations were used to model the wage process.

Selection equation: $d = ax + \theta z + e$

Outcome equation: $y = \beta x + u, \qquad \text{if } d > \delta$

where d is the selection variable, y is the outcome variable, x is a regressor in both equations and z is a regressor that is unique to the first equation. e and u are bivariate normally distributed errors with correlation coefficient ϱ.

The values for the parameter ϱ, the parameter θ and the ratio of the population for whom $d > \delta$ were allowed to vary, permitting us to compare the results obtained under different sets of specifications. In our model the correlation between the two error term, denoted by ϱ, and the degree of identification, denoted by θ, where both allowed to vary between 0, ¼, ½, ¾ and 1. a, β and σ_e were both set to one, since it has been shown that the efficiency and precision of subsample OLS and sample selectivity estimators are unaffected by the choice of a and β and behave similarly when the variance of e is either increased or decreased (Nelson, 1984: 190)

We allowed for two different selection rates, namely 33% and 66%. The first value was chosen to roughly correspond with the estimated South African employment rate of 40.3% (calculated over the whole working age population). The proportion of the sample judged to be employed drops to 35.4% when we omit those individuals for whom we also have no observable market wage. This value is significantly lower than that of most developed countries and consequently also lower than the default values used in previous Monte Carlo simulations. With this in mind, an alternative censoring value was chosen, one that corresponds to a 66% employment rate, allowing us to test whether the severity of the censoring has a significant impact on the results.

For each simulation, we generated a sample of 10 000 observations using the true parameters and an error term. These "true" parameters were then approximated using the four different techniques: the conventional OLS subsample method, the ML method, the Heckman two-step and the Ichimura-Lee semi-parametric method. This process was repeated 1000 times. The average of these beta-approximations and the average of the mean squared error were then calculated over the 1000 trials.

b) Results under normality

Assuming the two error components, e and u were bivariate normally distributed, the following beta estimates and mean squared error estimates (in parenthesis) were obtained:

Table 1: Average beta's and standard errors obtained from Monte Carlo simulations with normally distributed errors and 66% censoring

	θ = 0		θ = 0.25		θ = 0.5		θ = 0.75		θ = 1	
Subsample OLS Estimate										
ϱ =1	1.000	*(0.000)*	0.854	*(0.021)*	0.708	*(0.086)*	0.562	*(0.192)*	0.415	*(0.343)*
ϱ =0.75	1.000	*(0.000)*	0.861	*(0.019)*	0.723	*(0.077)*	0.584	*(0.173)*	0.445	*(0.308)*
ϱ =0.5	1.000	*(0.000)*	0.879	*(0.015)*	0.758	*(0.058)*	0.638	*(0.131)*	0.517	*(0.233)*
ϱ =0.25	1.000	*(0.000)*	0.901	*(0.010)*	0.802	*(0.039)*	0.702	*(0.089)*	0.603	*(0.158)*
ϱ =0	1.000	*(0.000)*	0.921	*(0.006)*	0.842	*(0.025)*	0.763	*(0.056)*	0.684	*(0.100)*
Maximum Likelihood Estimate										
ϱ =1	1.000	*(0.000)*	0.997	*(0.001)*	0.999	*(0.001)*	0.999	*(0.001)*	0.998	*(0.000)*
ϱ =0.75	1.000	*(0.000)*	0.999	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*
ϱ =0.5	1.000	*(0.000)*	1.000	*(0.000)*	0.999	*(0.000)*	1.001	*(0.000)*	1.000	*(0.000)*
ϱ =0.25	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*
ϱ =0	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*
Two-Step Estimate										
ϱ =1	1.000	*(0.000)*	0.999	*(0.002)*	1.000	*(0.002)*	0.998	*(0.002)*	0.996	*(0.003)*
ϱ =0.75	1.000	*(0.000)*	0.999	*(0.001)*	1.001	*(0.001)*	1.000	*(0.001)*	0.999	*(0.001)*
ϱ =0.5	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.001	*(0.000)*	1.000	*(0.000)*
ϱ =0.25	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*
ϱ =0	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*	1.000	*(0.000)*
Semi-Parametric Estimate										
ϱ =1	1.001	*(0.000)*	0.960	*(0.004)*	0.922	*(0.009)*	0.881	*(0.017)*	0.839	*(0.030)*
ϱ =0.75	1.001	*(0.000)*	0.989	*(0.001)*	0.981	*(0.001)*	0.969	*(0.002)*	0.958	*(0.003)*
ϱ =0.5	1.000	*(0.000)*	0.996	*(0.000)*	0.993	*(0.000)*	0.991	*(0.000)*	0.987	*(0.001)*
ϱ =0.25	1.000	*(0.000)*	0.998	*(0.000)*	0.997	*(0.000)*	0.995	*(0.000)*	0.994	*(0.000)*
ϱ =0	1.000	*(0.000)*	0.999	*(0.000)*	0.998	*(0.000)*	0.997	*(0.000)*	0.996	*(0.000)*
Overall										
Subsample OLS Estimate									0.7664	*(0.0856)*
Maximum Likelihood Estimate									1.026	0.9998
Two-Step Estimate									0.9997	*(0.0007)*
Semi-Parametric Estimate									0.9775	*(0.0028)*

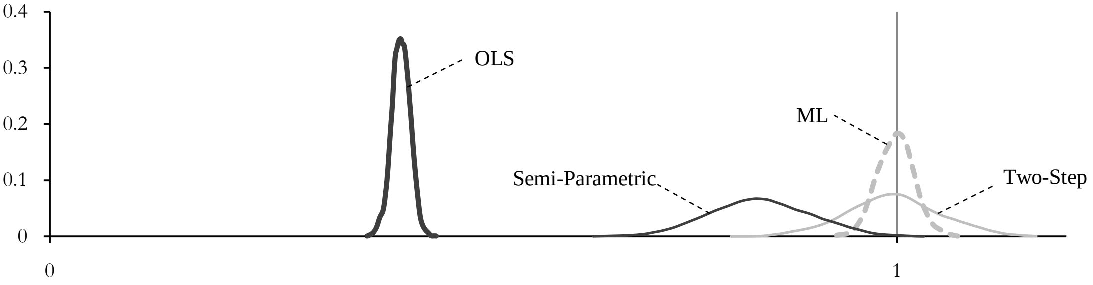
Note: For each of the estimation methods, the degree of correlation between the errors, ϱ, decreases as one reads the table from the top downwards and the degrees of correlation between the exogenous variables in the two equations, θ, increases from left to right.

Since sample selection bias works through the correlation between the unobservable characteristics, *e* and *u*, it is unsurprising that subsample OLS estimates grow more biased as the value of *θ* increases and that there exists no sample selection biased when *θ = 0.* The role of identification is also apparent; the subsample OLS estimates become more bias as the correlation between the regressors in the selection equation and those in the outcome equation increases. The degree of censoring also played an important role. Although we do not include the results here, we found that the OLS estimates become less biased as the size of the subsample relative to the full sample increases. Mean squared errors dropped by about 50% on average as censoring decreased from 33% to 66%.

Although both the ML and two-step models succeed in correcting for the sample selection bias, the ML estimates generally appear to be more precise, judging by the lower overall mean squared error values of 0.0004 rather than 0.0007. The mean squared error of the ML estimator was lower than that of the two-step model, regardless of which set of parameters were used. This difference in precision (mean squared error) between the ML and two-step models was greatest where θ was lowest, corresponding to the case of weak identification. This serves as further proof of the two-step model's inferiority in dealing with sample selection problems when exclusion restrictions are lacking.

This point is made more vivid below, where we graphed the four competing models under the assumption that $\theta = 1$ and $\varrho = 0$. From the graph it clear that although both the two-step and ML models are unbiased, the ML is more efficient, since its estimates are more narrowly distributed around the true value.

Figure 1: Kernel density curve of beta's obtained by different models using Monte Carlo simulation with normally distributed errors, 66% censoring, $\theta = 1$ and $\varrho = 0$.



The above figure also illustrates that the semi-parametric estimator succeeded in correcting for some of the effects of sample selection bias, obtaining an average estimate somewhere between the OLS estimates and the true value of 1. The semi-parametric estimates were however far worse than both the ML and two-step models, both of which recorded smaller biases and lower mean squared errors.

While both the two-step and ML method appeared to be sufficiently accurate under most circumstances, both experienced a substantial increase in their mean squared errors when the degree of censoring increased, rendering them less precise. This was not the case for OLS estimators. As was predicted in section 2(b), the estimators remained consistent even when the errors were distributed non-normally. This means that the consistency of the two-step model was only dependent on the distribution of the error term *e* in the selection equation and not on that of *u*, since the outcome equation made use of the OLS method, which is less sensitive to deviations from normality.

c) Results under t-distribution and under χ^2-distribtution

It is vital to also consider the implications of deviations from the normality assumptions. Zuehlke and Zeman (1991) conducted Monte Carlo simulations to test the sensitivity of sample selection models to the normality assumption. They compared the results under the conventional bivariate normality distribution to that of a bivariate t-distribution with five degrees of freedom and a bivariate χ^2-distribution with five degrees of freedom. Their results were inconclusive, for although the Heckman two-step reduced the bias, its parameter estimates had higher standard errors than that of the subsample OLS models.

In this study, a similar approach was followed. The Monte Carlo test was conducted under the normality assumption as well as for a bivariate t and bivariate χ^2 distribution. Results for the normal distribution have already been reported above in table 1. The results for simulations generating error values with a bivariate t-distribution and bivariate χ^2-distribution, both with five degrees of freedom, are summarised below in tables 2 and 3.

Table 2: Average beta's and standard errors obtained from Monte Carlo simulations with t-distributed errors and 66% censoring

	θ = 0		θ = 0.25		θ = 0.5		θ = 0.75		θ = 1	
Subsample OLS Estimate										
ϱ =1	1.001	*(0.001)*	0.822	*(0.033)*	0.643	*(0.129)*	0.465	*(0.288)*	0.290	*(0.506)*
ϱ =0.75	0.999	*(0.001)*	0.828	*(0.031)*	0.657	*(0.119)*	0.487	*(0.265)*	0.314	*(0.472)*
ϱ =0.5	1.001	*(0.001)*	0.848	*(0.024)*	0.695	*(0.094)*	0.541	*(0.212)*	0.386	*(0.379)*
ϱ =0.25	1.001	*(0.001)*	0.869	*(0.018)*	0.741	*(0.068)*	0.609	*(0.154)*	0.480	*(0.272)*
ϱ =0	1.002	*(0.001)*	0.892	*(0.012)*	0.786	*(0.047)*	0.679	*(0.104)*	0.571	*(0.185)*
Maximum Likelihood Estimate										
ϱ =1	1.010	*(0.078)*	1.048	*(0.059)*	1.057	*(0.032)*	1.059	*(0.023)*	1.085	*(0.023)*
ϱ =0.75	0.995	*(0.028)*	1.048	*(0.021)*	1.040	*(0.017)*	1.047	*(0.014)*	1.073	*(0.017)*
ϱ =0.5	0.999	*(0.004)*	1.015	*(0.005)*	1.022	*(0.006)*	1.028	*(0.007)*	1.043	*(0.008)*
ϱ =0.25	1.001	*(0.002)*	1.005	*(0.002)*	1.012	*(0.003)*	1.011	*(0.003)*	1.019	*(0.004)*
ϱ =0	1.002	*(0.001)*	1.004	*(0.002)*	1.005	*(0.002)*	1.008	*(0.002)*	1.012	*(0.003)*
Two-Step Estimate										
ϱ =1	1.017	*(0.037)*	1.055	*(0.040)*	1.119	*(0.055)*	1.187	*(0.083)*	1.243	*(0.120)*
ϱ =0.75	0.996	*(0.010)*	1.022	*(0.011)*	1.034	*(0.014)*	1.060	*(0.018)*	1.071	*(0.024)*
ϱ =0.5	1.000	*(0.004)*	1.007	*(0.004)*	1.015	*(0.005)*	1.023	*(0.006)*	1.028	*(0.007)*
ϱ =0.25	1.001	*(0.002)*	1.003	*(0.002)*	1.012	*(0.003)*	1.012	*(0.003)*	1.016	*(0.004)*
ϱ =0	1.002	*(0.001)*	1.003	*(0.001)*	1.007	*(0.002)*	1.012	*(0.002)*	1.014	*(0.003)*
Semi-Parametric Estimate										
ϱ =1	1.018	*(0.042)*	0.898	*(0.056)*	0.813	*(0.083)*	0.730	*(0.140)*	0.649	*(0.203)*
ϱ =0.75	0.996	*(0.011)*	0.985	*(0.011)*	0.962	*(0.014)*	0.944	*(0.017)*	0.923	*(0.024)*
ϱ =0.5	0.999	*(0.004)*	0.994	*(0.004)*	0.990	*(0.005)*	0.985	*(0.005)*	0.978	*(0.007)*
ϱ =0.25	1.001	*(0.002)*	0.997	*(0.002)*	0.999	*(0.002)*	0.993	*(0.003)*	0.992	*(0.003)*
ϱ =0	1.002	*(0.001)*	0.999	*(0.001)*	1.000	*(0.002)*	1.000	*(0.002)*	0.998	*(0.003)*
Overall										
Subsample OLS Estimate									0.704	*(0.137)*
Maximum Likelihood Estimate									1.026	*(0.015)*
Two-Step Estimate									1.038	*(0.018)*
Semi-Parametric Estimate									0.954	*(0.026)*

Note: For each of the estimation methods, the degree of correlation between the errors, ϱ, decreases as one reads the table from the top downwards and the degrees of correlation between the exogenous variables in the two equations, θ, increases from left to right.

The t-distribution was introduced to the Monte Carlo simulations in an attempt to establish how sensitive the parametric sample selection models are to deviation from normality. Both the ML and two-step models' estimates performed worse. On average, the ML method performed better when identification was low, while the two-step and semi-parametric methods were superior when the identification was higher.

The semi-parametric model was less sensitive to the slight deviations from normality. Surprisingly, its estimates actually fared better under the t-distribution than it did under the normality-distribution. The semi-parametric method however appeared to be even more dependent on the existence of proper exclusion restrictions than both the ML and two-step methods. While it outperformed both when the identification was high, it came apart when there were no exclusion restrictions (when ϱ = 0).

The χ^2-distribution was also simulated to investigate how the rival sample selection approaches fare when skewness is also introduced into the model. The Monte Carlo results are presented in the table 3, below.

Table 3: Average beta's and standard errors obtained from Monte Carlo simulations with X^2-distributed errors and 66% censoring

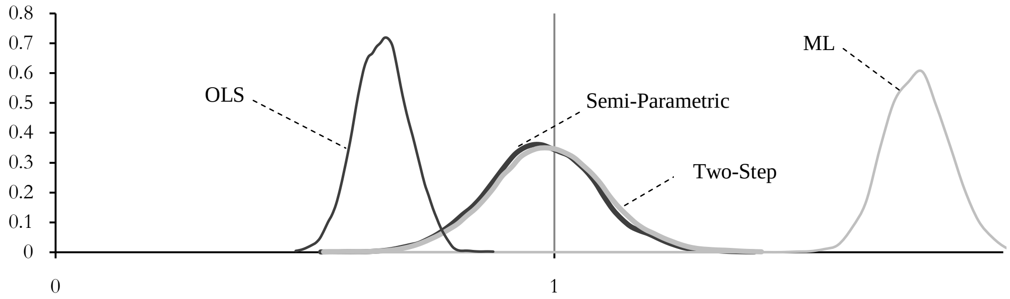
	$\theta = 0$		$\theta = 0.25$		$\theta = 0.5$		$\theta = 0.75$		$\theta = 1$	
Subsample OLS Estimate										
ϱ =1	1.000	*(0.002)*	0.826	*(0.033)*	0.647	*(0.127)*	0.468	*(0.286)*	0.290	*(0.508)*
ϱ =0.75	1.000	*(0.002)*	0.825	*(0.033)*	0.650	*(0.125)*	0.474	*(0.280)*	0.299	*(0.495)*
ϱ =0.5	1.000	*(0.002)*	0.828	*(0.032)*	0.655	*(0.122)*	0.486	*(0.268)*	0.315	*(0.473)*
ϱ =0.25	0.999	*(0.002)*	0.838	*(0.029)*	0.668	*(0.113)*	0.504	*(0.250)*	0.337	*(0.444)*
ϱ =0	0.998	*(0.002)*	0.842	*(0.027)*	0.686	*(0.102)*	0.528	*(0.226)*	0.375	*(0.395)*
Maximum Likelihood Estimate										
ϱ =1	1.481	*(0.534)*	1.522	*(0.545)*	1.505	*(0.487)*	1.476	*(0.447)*	1.461	*(0.439)*
ϱ =0.75	1.495	*(0.552)*	1.706	*(0.661)*	1.719	*(0.560)*	1.666	*(0.459)*	1.632	*(0.41)*
ϱ =0.5	1.465	*(0.507)*	1.808	*(0.702)*	1.725	*(0.531)*	1.651	*(0.429)*	1.613	*(0.382)*
ϱ =0.25	1.298	*(0.316)*	1.810	*(0.669)*	1.682	*(0.469)*	1.607	*(0.373)*	1.571	*(0.332)*
ϱ =0	1.143	*(0.147)*	1.756	*(0.578)*	1.629	*(0.400)*	1.552	*(0.310)*	1.523	*(0.279)*
Two-Step Estimate										
ϱ =1	0.972	*(0.088)*	0.961	*(0.100)*	0.891	*(0.117)*	0.835	*(0.169)*	0.758	*(0.229)*
ϱ =0.75	1.000	*(0.027)*	0.992	*(0.029)*	0.978	*(0.036)*	0.952	*(0.043)*	0.955	*(0.061)*
ϱ =0.5	1.004	*(0.011)*	0.995	*(0.011)*	0.988	*(0.013)*	0.987	*(0.017)*	0.985	*(0.019)*
ϱ =0.25	0.997	*(0.006)*	1.003	*(0.006)*	0.992	*(0.008)*	0.993	*(0.009)*	0.988	*(0.012)*
ϱ =0	0.996	*(0.004)*	0.999	*(0.004)*	0.996	*(0.005)*	0.994	*(0.006)*	0.995	*(0.009)*
Semi-Parametric Estimate										
ϱ =1	0.975	*(0.061)*	0.940	*(0.070)*	0.852	*(0.096)*	0.777	*(0.148)*	0.684	*(0.219)*
ϱ =0.75	1.001	*(0.024)*	0.978	*(0.026)*	0.952	*(0.033)*	0.913	*(0.043)*	0.901	*(0.060)*
ϱ =0.5	1.004	*(0.010)*	0.989	*(0.010)*	0.976	*(0.013)*	0.969	*(0.017)*	0.962	*(0.019)*
ϱ =0.25	0.997	*(0.006)*	1.000	*(0.006)*	0.984	*(0.008)*	0.982	*(0.009)*	0.973	*(0.012)*
ϱ =0	0.996	*(0.004)*	0.996	*(0.004)*	0.990	*(0.005)*	0.985	*(0.007)*	0.984	*(0.009)*
Overall										
Subsample OLS Estimate									0.662	*(0.175)*
Maximum Likelihood Estimate									1.580	*(0.461)*
Two-Step Estimate									0.968	*(0.042)*
Semi-Parametric Estimate									0.950	*(0.037)*

Note: For each of the estimation methods, the degree of correlation between the errors, ϱ, decreases as one reads the table from the top downwards and the degrees of correlation between the exogenous variables in the two equations, θ, increases from left to right.

The ML-method, which had performed well up to this point, fared considerably worse when the error term was not symmetrically distributed along the y-axis. Notably, it now became the worst estimator regardless of the values that ϱ and θ took on. This was in line with Olsen's (1982: 236) finding that "maximum likelihood methods have the little appreciated attribute that they are extremely sensitive to the assumption made about the population distribution of the regression residuals".

There was no real difference in the performance of the two-step and semi-parametric methods. Taken over all the values, the two-step method provided a better β estimate (0.968) than the semi-parametric method (0.950). The mean squared error value for the later method was, however, lower than for the two-step method.

Figure 2: Kernel density curve of beta's obtained by different models using Monte Carlo simulation with X^2-distributed errors, 66% censoring, $\theta = 0.5$ and $\varrho = 0.5$.



4. Applying Methods to SA Dataset

a) Finding Exclusion Restrictions

To adequately and accurately correct for the impact of sample selection, some measure is required to adjust for the colinearity between the regressors in the outcome equation and the correction term, called g(x_1) in section 3.e. The most effective way of doing this is to add at least one variable to the selection equation that is not contained in the outcome equation. This variable needs to influence the individual's likelihood of being employed, but should have little or no impact on his or her wage. According to Puhani (2000: 58), household variables are most appropriate for use as exclusion restrictions, since they are most likely to fit this criterion - affecting participation decisions (and likelihood of being employed) without also affecting the wage an individual would attain. This is not the case for most other variables, especially those that denote personal characteristics, since these are usually also correlated with the wage function.

As statisticians often warn, it is vital to note that the partial correlation alone of a variable with the employment variable is not enough, since it gives no information about the direction of the causality, which could be in either or even both directions. It is for instance, quite likely that an individual's employment status could affect her or his decision to marry, but it is also conceivable that an individual's marital status can affect his or her decision to look for work. If this is the case, it would be incorrect to include the marital status dummy in the selection equation, because of the variable's endogeneity.

Using the 2005 September Labour Force Survey of South African dataset, the preliminary tests were performed on a host of household variables in an effort to find a variable that was significant in the selection equation but insignificant in the wage function. The tests were administrated independently, by adding each household variable to the selection equation to test its significance on the employment decision, after which it was added to the outcome equation to determine whether it has an effect on the wage determination process. The test showed that most of the household variables had a significant effect on the likelihood of being employed, at a 5% level of significance. In the case of the wage function, however, only three of the household variables were found to be insignificant.

These three variables, which were the number of children, the number of employed persons as well as the total number of people in each household, were tested further to ensure that preliminary results did not themselves suffer from the lack of a proper exclusion restriction. This was done by adding all three of these variables to the selection equation and adding each variable in turn to the outcome equation. This method allows the two variables that are not included in the outcome equation to act as temporary exclusion restrictions, while testing the validity of the third exclusion restriction. This procedure was followed for all three sample selection models. The t-statistics attained under these test, as well as those attained under the conventional OLS method, are tabulated below.

Table 4: t-statistics of household variables

	OLS	2-step	ML	SP
# employed individuals in household*	1.87	1.97	1.57	0.06
# children in household	-1.97	0.27	-0.24	-0.26
household size	-2.59	4.30	0.10	1.07

Source: September 2005 LFS

Ignoring the OLS results, one finds that all for all four these samples all the variables in the outcome equation were found to be insignificant at a 5% level, apart from the household size variable which was found to be significant under the two-step model. These three variables thus appear to be adequate for use as exclusion restrictions. They were shown to be partially correlated with the selection equation, without being partially correlated with the outcome equation.. Despite this empirical validation we decided to drop the variable containing the amount of employed individuals (apart from the individual itself) in the household as an exclusion restriction on theoretical grounds. We feared that these variables may bias the results if it captures the common immeasurable attributes that employed individuals within a households share rather than the effect of having an additional breadwinner.

b) Testing Normality

In section 4, we established that both the two-step and ML methods yield biased estimates of the *β*'s if the errors are not normally distributed. Several normality tests exist, but most of these test the normality assumption against some alternative distributional assumption. Chesher & Irish (1987), however, developed a normality test that can be performed without having to compare it to any other specific distribution. This is done by comparing the residual moments with what they would have been if the errors were normally distributed.

Using the standardised residuals of the probit function, the first four moments, which denote the mean, variance, skewness and kurtosis, were calculated for all *n* observations using the selection variable *d*, the *k* explanatory variables, labeled *x,* and the estimated parameter of α. Chesher & Irish (1987: 41) proposed that the four moments be derived using the following formulas:

$$\hat{e}^{(1)} = -(1-d)\lambda(ax) + d\lambda(-ax)$$

$$\hat{e}^{(2)} = -\, ax\hat{e}^{(1)}$$

$$\hat{e}^{(3)} = (2+(ax))\hat{e}^{(1)}$$

$$\hat{e}^{(4)} = -(3ax+(ax)^{3})\hat{e}^{(1)}$$

where *d = 1 if the individual is employed and d = 0 if the individual is not employed,*
ax is the linear prediction of the fitted model, and
λ(.) is the standard normal hazard function, $\phi(z)/(1-\Phi(z))$

Once the moments are calculated, we multiply the first moment with each of the regressors contained in the selection equation to derive a matrix $\hat{e}^{(1)}x_1$. One can then proceed in two manners: either constructing a larger matrix *L*, consisting of $\hat{e}^{(1)}x_1$, $\hat{e}^{(3)}$ and $\hat{e}^{(4)}$ and obtaining the Lagrange Multiplier (LM) statistic by solving $t'L[L'L]^{-1}L't$, where *t* is a vector of ones, or equivalently, regress a vector of ones on the *k*+2 columns contained in the matrix *L*. In the latter case, the LM statistic would be equal to the explained sum of squares. In both cases the LM-statistic follows a chi-squared distribution, which is used to calculate the critical value for the test (whether or not the null hypothesis of normality can be rejected).

Table 5: Normality Testing: Entire Working Age Population under different exclusion restrictions

Control Variables	*none*		*Children*		*Children & HHSize*	
$\hat{e}_1$	0.295	(4.93)	0.261	(5.08)	0.153	(3.36)
$\hat{e}_1$•Experience	0.014	(2.79)	0.029	(6.90)	0.024	(7.37)
$\hat{e}_1$•Experience2	0.000	(-1.85)	0.000	(-5.75)	0.000	(-6.37)
$\hat{e}_1$•Education	0.016	(4.82)	0.024	(8.86)	0.019	(7.56)
$\hat{e}_1$•Female	-0.045	(-2.06)	-0.152	(-7.35)	-0.122	(-6.44)
$\hat{e}_1$•Rural	0.154	(9.97)	0.119	(8.70)	0.119	(8.84)
$\hat{e}_1$•White	0.268	(12.94)	0.248	(11.76)	0.239	(11.32)
$\hat{e}_1$•Coloured	0.000	(0.00)	0.040	(1.89)	0.020	(0.96)
$\hat{e}_1$•Indian	0.219	(6.00)	0.176	(4.79)	0.185	(5.08)
$\hat{e}_1$•Province2	-0.047	(-3.87)	-0.058	(-4.93)	-0.050	(-4.27)
$\hat{e}_1$•Province3	-0.075	(-9.18)	-0.078	(-9.46)	-0.072	(-8.71)
$\hat{e}_1$•Province4	-0.068	(-10.21)	-0.065	(-9.63)	-0.066	(-9.76)
$\hat{e}_1$•Province5	-0.042	(-9.31)	-0.037	(-8.16)	-0.035	(-7.66)
$\hat{e}_1$•Province6	-0.050	(-11.2)	-0.049	(-10.95)	-0.046	(-10.24)
$\hat{e}_1$•Province7	-0.018	(-5.05)	-0.015	(-4.36)	-0.015	(-4.18)
$\hat{e}_1$•Province8	-0.023	(-6.55)	-0.019	(-5.60)	-0.019	(-5.44)
$\hat{e}_1$•Province9	-0.030	(-9.66)	-0.031	(-9.94)	-0.029	(-9.29)
$\hat{e}_1$•m1			-0.015	(-1.43)	0.000	(0.02)
$\hat{e}_1$•m2			-0.021	(-2.09)	-0.011	(-1.19)
$\hat{e}_1$•m3			-0.021	(-1.87)	-0.009	(-0.90)
$\hat{e}_1$•f1			-0.008	(-0.81)	0.000	(-0.05)
$\hat{e}_1$•f2			-0.005	(-0.58)	-0.002	(-0.22)
$\hat{e}_1$•f3			-0.001	(-0.12)	0.000	(-0.04)
$\hat{e}_1$•hhsize					-0.016	(-2.84)
$\hat{e}_3$	-0.192	(-7.71)	-0.264	(-11.82)	-0.147	(-7.4)
$\hat{e}_4$	0.052	(3.95)	0.095	(8.84)	0.061	(6.87)
MSE	1453.19		1459.40		1376.94	
df	19		25		26	
R^2	0.021		0.021		0.020	
observations	68735		68735		68735	

Note: t-statistics in parenthesis

The LM statistics obtained from Table 5 ranged between 1377 and 1459 (depending on the exclusion restrictions used). The assumption of normality was rejected in all three cases, since all three the statistics were significantly higher than their corresponding critical values, which varied between 35 and 48.

Table 6: Normality Testing: Subsamples of Working Age Population

variable	*Nochildren*		*children*		*children&hhsize*	
White Males	642.8	(15)	661.1	(18)	664.8	(19)
White Females	238.8	(15)	229.9	(18)	224.6	(19)
Black Males	444.5	(15)	404.7	(18)	350.9	(19)
Black Females	421.1	(15)	342.4	(18)	337.6	(19)
Coloured Males	66.6	(15)	66.6	(18)	60.0	(19)
Coloured Females	60.9	(15)	67.0	(18)	68.0	(19)
Indian Males	98.8	(14)	103.7	(17)	108.7	(18)
Indian Females	21.9	(12)	30.6	(15)	36.4	(16)

Note: The degrees of freedom for the last two groups are lower than for the rest. This is due to the shortage of Indian Males in the Free State and Indian Females in the Eastern Cape, North West Province and Free State.

These tests were repeated for certain subsections of the population. Table 6 reports these findings and shows that non-normality is consistently worse among men than among women. The LM-statistic also differs significantly between races; the value of whites being the highest, followed by blacks, coloureds and then Indians. For all six these groups their LM statistics exceeded their critical values at a 5% level of significance. The last group, which consisted of Indian females, came closest to being normally distributed. It had a LM-statistic of 21.9 and a critical value of 21.03 when no exclusion restrictions were used.

c) Comparing Results

All four the models (the subsample OLS model, the Heckman ML model, the Heckman two-step model and the semi-parametric model) were applied to a September 2005 Labour Force Survey dataset. The variables number of children and household size were used as exclusion restrictions. The results are tabulated below.

Table 7: Model Testing, Children & Household Size as exclusion restrictions

	wage equation				*employment equation*		
Variable	*OLS*	*ML*	*2 Step*	*SP*	*ML*	*2 Step*	*SP*
Experience	0.026	0.021	0.021	0.036	0.111	0.111	0.111
	(14.55)	(5.26)	(4.63)	(20.72)	(67.61)	(57.36)	(28.86)
Experience2	-0.107	-0.035	-0.021	-0.314	-1.773	-1.773	-1.701
	(-3.18)	(-0.49)	(-0.28)	(-9.55)	(-52.37)	(-44.46)	(-28)
Education	0.133	0.131	0.131	0.126	0.062	0.062	0.068
	(64.76)	(38.11)	(37.27)	(63.43)	(28.17)	(22.61)	(20.03)
Female	-0.273	-0.256	-0.253	-0.297	-0.403	-0.405	-0.435
	(-23.35)	(-12.33)	(-11.66)	(-25.58)	(-21.19)	(-17.18)	(-16.81)
Rural	-0.222	-0.215	-0.213	-0.198	-0.144	-0.144	-0.139
	(-14.13)	(-10.18)	(-10)	(-14.22)	(-9.1)	(-7.59)	(-7.02)
White	1.001	1.003	1.003	0.999	-0.098	-0.098	0.051
	(54.65)	(29.71)	(29.73)	(48.31)	(-3.48)	(-2.9)	(0.91)
Coloured	0.269	0.264	0.263	0.234	0.247	0.247	0.284
	(12.1)	(7.7)	(7.67)	(11.56)	(9.59)	(7.29)	(7.4)
Indian	0.763	0.766	0.767	0.727	-0.056	-0.057	-0.011
	(22.61)	(15.54)	(15.47)	(19.24)	(-1.24)	(-1.03)	(-0.17)
Union	0.656	0.656	0.656	0.731			
	(48.06)	(36)	(35.97)	(52.62)			
m1					0.073	0.072	0.051
					(6.08)	(4.88)	(3.26)
m2					-0.091	-0.091	-0.100
					(-8.41)	(-7.11)	(-7.81)
m3					-0.135	-0.135	-0.133
					(-11.68)	(-9.86)	(-9.79)
f1					-0.062	-0.061	-0.065
					(-4.81)	(-4.16)	(-3.56)
f2					-0.057	-0.057	-0.053
					(-5.4)	(-4.62)	(-3.91)
f3					-0.052	-0.052	-0.048
					(-4.78)	(-4.01)	(-3.39)
Household Size					-0.144	-0.144	-0.136
					(-27.41)	(-23.89)	(-18.66)
λ			-0.068	-0.059			
			(-1.53)	(-1.18)			
ϱ		-0.066					
		(-1.70)					
constant	0.184	0.293	0.313	0.279	-1.261	-1.263	1.263
	(5.08)	(3.35)	(3.23)	(2.97)	(-30.56)	(-24.69)	(fixed)
(pseudo) R^2	0.478		0.478	0.471		0.198	
observations	22960	22960	22960	22960	68735	68735	68735

The ML estimator and two-step estimators delivered similar estimates of α in the selection equation. This is somewhat surprising since these models used different techniques to derive these estimates. The two-step estimator used a standard probit function that ignores the outcome equation, while the ML derives its estimate of α by solving the selection and outcome equation simultaneously. The similarity of the α estimates provides evidence that the effect of the outcome equation on the selection equation within the ML model is minimal.

The non-parametric estimates of α differs from those attained using the parametric ML and two-step methods. Judging by the coefficients attained, it appears as though the effect of education, experience, gender, race and type of area one resides in all play a larger role in the semi-parametric employment equation than in its parametric counterpart. It is not just the magnitudes of the coefficients that differ between parametric and semi-parametric method, the effect of being white and Indian rather than being black turned from negative to positive.

The semi-parametric β estimates obtained in Table 7 fails to agree with those obtained by the parametric sample selection models and the conventional subsample OLS procedure. The returns to education appear to be lower. The impact on experience is deceptive, although both coefficients are larger; the overall effect of experience (within the feasible range of 0 to 50 years) is much smaller than it was for the three parametric models. The effect of gender is greater, while the effect of race and the type area one resides in is smaller. The effect of union membership is also larger under the non-parametric assumption.

There are two ways to test whether the problem of sample selection merits intervention. If the normality assumption is valid, either the ML or 2-step models allow testing of the validity of $\varrho = 0$ (i.e. no sample selection bias). If the normality assumption, however, fails, as appears to be the case with the South Africa data, then the best we can do is to compare the results obtained from the OLS and sample selection techniques to see if they differ in an economically significant manner. In this study they do and as a result, intervention is required.

5. Conclusion

This paper tried to establish whether sample selection is indeed a problem in South African labour market analysis and if it is, how it can be addressed optimally. Our findings suggest that the questions should be addressed in reverse order, since one’s choice of selection correction method ultimately determines whether or not the problem is significant.

The results obtained from sample selection methods did not differ from those that did not use sample selection methods under parametric testing. When differences did occur it was due to the lack of proper exclusion restrictions rather than the effect of selection bias. This provides further evidence that the sample selection models can be misleading, when they are not handled with the necessary caution. This is not the case for semi-parametric methods. The semi-parametric estimates differed greatly from those obtained from conventional OLS analysis.

Despite the advantage that semi-parametric estimates offer over there parametric counterparts, they are rarely used in applied work. According to Vella (1998, 144), the wide-scale implementation of these methods has been hindered by the degree of technicality associated with these techniques and the perception that parametric models perform adequately as long as the conditional mean is correctly specified. This is regrettable since labour market analysis can benefit a great deal from the use of these methods.

6. References

BERK, R.A. and RAY. S.C., 1982. Selection Biases in Sociological Data. *Social Science Research*. 11: 352-398.

CHESHER, A. and IRISH, M. 1987. Residual Analysis in the Grouped and Censored Normal Linear Model. *The Journal of Econometrics*. 34: 33-61.

DAVIDSON, R. and MACKINNON, J.G., 1984. Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics*. 25: 241-262.

GALLANT, A.R. and NYCHKA, D.W., 1987. Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*. 55, 2: 363-390.

GRONAU, R. 1974. Wage Comparisons – A selectivity Bias. *Journal of Political Economy*. 82. 6: 1119-1144.

HECKMAN, J., 1974. Shadow Prices, Market Wages and Labor Supply. *Econometrica*. 42, 4: 697-694.

HECKMAN, J., 1979. Sample Selection Bias as a Specification Error. *Econometrica*. 46, 1: 153-161.

HECKMAN, J., 1990. Selectivity Bias: New Developments. Varieties of Selection Bias. *American Economic Review*. 80, 2: 313-318.

HECKMAN, J. and RICHARD, R., 1985. Alternative Methods for Evaluating the Impact of Interventions. *Longitudinal Analysis of Labor Market Data*. Ed. Heckman and Singer. Cambridge: Cambridge University Press.

ICHIMURA, H. and LEE. L.F., 1991. Semiparametric Least Squares of Multiple Index Models: Single Equation Estimation. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Ed. Barnett, Powell and Tauchen. Cambridge: Cambridge.University Press.

JOHNSTON, J. and DINARDO, J., Econometric Methods. 4th ed. New York: McGraw-Hill

NELSON, F.D., 1984. Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection. *The Journal of Econometrics*. 24: 181–196.

NEWEY. W., 1988. Two-Step Estimation of Sample Selection Models. Unpublished.

OLSEN, R.J., 1982. Distributional Test for Selectivity Bias and a More Robust Likelihood Estimator. *International Economic Review*. 23: 223-240.

PUHANI, P.A., 2000. The Heckman Correction for Sample Selection and its Critique. *Journal of Economic Surveys*. 14, 1: 53-68.

RICE, A.R., 1995. *Mathematical Statistics and Data Analysis.* 2nd ed. Belmont, CA: Duxbury Press.

STEWART, M.B., 2004. Semi-nonparametric Estimation of Extended Ordered Probit Models. *The Stata Journal.* 4, 1: 27-39.

STOLZENBERG, R.M. and RELLES, D.A., 1990. Theory Testing in a World of Constrained Research Design. The Significance of Heckmans' Censored Sampling Bias Correction for Nonexperimental Research. *Sociological Methods and Research*, 18, 4: 395–415.

VELLA, F., 1998. Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources.* 33, 1: 127–168.

WOOLDRIDGE, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

ZUEHLKE, T.W. and ZEMAN, A.R., 1990. A Comparison of Two-Stage Estimators of Censored Regression Models. *The Review of Economics and Statistics.* 72: 185-188.