

Department of Economics
Stellenbosch University

*Sample Selection Bias
and the South African Wage Function*

by

Cobus Burger

Supervisors: Mr. R.P. Burger & Prof. S. van der Berg

March 2007

**Assignment presented in partial fulfillment of the requirements for the degree of Bachelors
of Commerce (Economics) with Honours at the University of Stellenbosch.**

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.

JEL Classification C14 C15 C34 J21

Table of Contents

1. Introduction.....	1
2. Sample Selection Bias.....	1
3. Models and Methodology	2
a. Sample Selection Model	2
b. Heckman’s Maximum Likelihood Estimator.....	4
c. Heckman’s Two-Step Estimator	5
d. Concerns Regarding Sample Selection Models.....	7
e. Semi-Parametric Estimator	9
4. Monte Carlo Simulations	11
a. Identification	11
b. Normality	15
5. Finding Exclusion Restrictions	17
a. Children.....	20
b. Number of Employed Individuals.....	22
c. Household Size	22
6. Testing Normality	23
7. Comparing Results.....	25
8. Conclusion	27
9. References.....	29
Appendix.....	32
a. Figures.....	32
b. Tables.....	36
c. Stata Code	52

1. Introduction

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 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 3 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 4 adds to this discussion, by testing the empirical validity of the competing models, using Monte Carlo simulations. In section 5 and 6 the model is applied to South African LFS data. These sections also cover the problems relating to exclusion restrictions and assuming normally distributed errors. Section 7 outlines the results relating to the four different models. Section 8 concludes.

2. Sample Selection Bias

Using 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 the effect of the observable characteristics can be controlled for by including these variables in the wage function, this is not the case for unobservable characteristics, like ambition and motivation. Yet these variables are likely to play an important role in the decision regarding participation in the labour force as well as affecting wage determination.

In the presence of sample selection bias, conventional wage functions fail to incorporate the role that unobservable attributes could have on the outcome equation. In doing this, these models make themselves susceptible to inconsistent estimators and misleading t-statistics, which in turn may lead to improper results and conclusions.

3. Model and Methodology

a) The Sample Selection Model

The first formal proposal of the sample selection model was by Gronau (1974). The model is essentially an augmented version of the Cragg (1971) model, relaxing the assumption of independence of the participation choice stage and the wage determination stage.

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 other may influence that individual's participation choice or 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 the unobserved productive characteristic, like drive and intelligence, that 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:

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

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

where d_i and y_i are the observed realisations,
 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, or similarly if $\rho \neq 0$. 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 and the Heckman two-step estimation procedure.

b) Heckman's maximum likelihood estimator

Maximum Likelihood (ML) diverges from the method of least squares, by using a likelihood function rather than a probability function to estimate parameters. The likelihood function is used to find the set of parameters that produce the highest likelihood ratio estimate. These parameters are “most likely” or “most probable” to have produced the observed data, most often (Rice, 1995: 254).

The solution is usually attained by maximising the log-likelihood function. Due to the logarithmic function's monotonicity, this log likelihood function delivers the same results as the conventional likelihood function. If the regressors are identically, independently distributed then the log-likelihood function can be expressed as the sum of the logged marginal densities.

$$l(\theta) = \sum \log[f(X_i|\theta)]$$

The maximum likelihood method has been shown to produce consistent estimates under a few plausible conditions. Maximum likelihood estimates have the further advantage of being normal and efficient if sample sizes are large enough (Gujarati, 2003: 113). Unfortunately, the distributional assumptions required to justify the use of the maximum likelihood estimator are no less stringent than is required of OLS: the modeller needs to assume the precise distribution of the error term. In fact, the non-normality assumption is only required to ensure the efficiency of the OLS estimates, not their consistency, whereas the ML estimators are generally not consistent under an incorrect distributional assumption.

Heckman (1974) was first to propose the use of the maximum likelihood method in dealing with the sample selection problem. His assumption that the residuals are bivariate normally distributed is a prerequisite, since it allows one to solve the parameters.

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 .

c) Heckman's two-step estimator

One major drawback of Heckman's maximum likelihood estimator is its procedural complexity. The added computation needed to solve these ML estimates together with people's relative unfamiliarity with the ML technique during the 1970's impeded its use. In another seminal article, Heckman (1979) developed the two-step estimator, a simpler version of his own ML method. The new method could be solved using the more familiar probit function and a conventional OLS regression. The 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. The inverse Mills ratio is a strictly positive, decreasing function that approaches zero on the right and the negative value of the argument to the left (Nelson, 1984: 184).

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 corrects the bias. The inverse Mills ratio forms part of this correction term.

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

If one assumes that e and u are jointly normally distributed, $\begin{bmatrix} e \\ u \end{bmatrix} \sim BN \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix} \right]$,

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 correlated with e . Adding this to the prior model one attains:

$$\begin{aligned}
E(y_i) &= \beta x_{2i} + E\left(\frac{\sigma_{12}}{\sigma_{11}} e + v \mid e > -\alpha x_1\right) \\
&= \beta x_{2i} + E\left(\frac{\sigma_{12}}{\sigma_{11}} e \mid e > -\alpha x_1\right) \\
&= \beta x_{2i} + \frac{\sigma_{12}}{\sigma_{11}} E\left(\frac{e}{\sigma_{11}} \mid \frac{e}{\sigma_{11}} > -\alpha x_1\right) \\
&= \beta x_{2i} + \frac{\sigma_{12}}{\sigma_{11}} \frac{\phi(-\alpha x_1)}{1 - \Phi(-\alpha x_1)} \\
&= \beta x_{2i} + \frac{\sigma_{12}}{\sigma_{11}} \lambda(-\alpha x_1)
\end{aligned}$$

where ϕ denotes the standard normal density function

and Φ denotes the cumulative distribution function.

From the above derivation it is clear that OLS wage function that neglect 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 $\lambda(\cdot) = \phi(\alpha x_1)/1 - \Phi(\alpha x_1)$ 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.

Since the probability of being employed cannot be observed, we have to settle for an estimate. The probit model can be used to estimate the likelihood of employment given a host of observable characteristics and a normally distributed error. The inverse Mills ratio can then be estimated using αx_1 , the linear prediction of the fitted probit model. Finally, the inverse Mills ratio should be added to the wage function as an additional regressor. The function is then solved using conventional OLS analysis. Under this method 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 null hypothesis of no selectivity bias is rejected.

d) Concerns regarding sample selection models

The popularity of the sample selection models introduced in the previous section has grown immensely since the 1970s, also gaining prominence outside economics. The accessibility of these sample selection models has benefited from the introduction of easy-to-implement sample selection procedures, which now come standard with many software programmes. This has promoted the adoption of these techniques by lowering the technical capabilities required for applying these methods.

While the wider use of these models has its benefits, some authors have recently emphasised that they should not be applied indiscriminately. This point was acknowledged by Heckman (1990: 317), who admits that simpler estimation methods may be just as good in answering economic questions under certain circumstances. Johnston and DiNardo (2004: 450) build on this statement, highlighting the sensitivity of these sample selection methods to a range of factors, like the presence of heteroscedasticity, the degree of identification and the validity of the distributional assumptions.

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. In these cases, the outcome equation is identified through the nonlinearity of the inverse Mills ratio alone, a function which has in fact been shown to be quasi-linear. As a result, these models run the risk of obtaining unreliable β 's and inflated standard errors.

Figure 1 in the appendix shows the distribution of the inverse Mills ratio. It can be observed that up to a value of 2, which includes 97.5% of the normally sampled observation, the function is close to being linear. Identification is thus heavily reliant on correctly specifying the upper tail of the error term. 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, given 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.

There have also been questions 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. Both models will produce inconsistent parameter estimates if normality fails. As noted in section 3(b), this is not the case for OLS estimators, which remain consistent even if the errors are distributed non-normally. This means that the consistency of the two-step model is only dependent on the distribution of

the error term e in the selection equation and not that of u , since the outcome equation makes use of the OLS method, which is less sensitive to deviations from normality.

According to Olsen (1982: 236) “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”.

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.² 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.

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

where e is an error term

K specifies the order of the Hermite polynomial,

² 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>

$\phi(\cdot)$ 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 $g(\cdot)$, which makes use of the index restriction, α_{x_1} , to use in the conditional expectation of the outcome equation.

$$E(y \mid d=1) = \beta x_2 + g(\alpha x_1)$$

The conventional two-step's inverse Mills ratio can not be used here, since the function makes use of the parametric assumption, i.e. normality. Several semi-parametric alternatives have been developed to approximate $g(\alpha x_1)$. 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 $g(\cdot)$ and Newey (1988) estimated the selection correction, $g(\alpha x_1)$, using an initial estimate of αx_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 $g(\cdot)$. The semi-parametric procedure is derived from the following two identities, that define the relationship between β , αx_1 and $g(\cdot)$.

$$\begin{aligned} E(y \mid d=1, \alpha x_1) &= \beta x_{2i} + g(\alpha x_1) \\ E(y - \beta x_2 \mid \alpha x_1) &= g(\alpha x_1) \end{aligned}$$

An estimation of $g(\alpha x_1)$ can be obtained by inserting the estimate of αx we obtained from the semi-parametric probit and a preliminary estimate of β into the second equation. The estimate of $g(\alpha x_1)$ 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 $g(\alpha x_1)$, 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.

The working of the iterative process can be viewed in Figure 17 (Appendix). The object function appears to be convex, since the final β estimators are not dependent on the initial β estimates. Given a true value of 2, starting values of 1 and 3 were used and shown to converge after about 20 iterations, while the two-step β estimates converged after about 15 iterations. When applying the iteration of the semi-parametric technique, the two-step model's estimates will be used since they allow us to save on computational time.

4. Monte Carlo Simulations

a) Identification

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)

Nelson (1984) tested the bias and efficiency of the Heckman two-step and the ML estimators and compared them to the original OLS estimates. The sample size was set to 2000 and 30 different sets of specifications were tested. The ρ value ranged from -0.5 to 0.95 and the value of R^2 from 0 to 0.999. Results suggest that the conditions under which the OLS bias is the greatest are also those under which the ML estimator outperforms the Heckman two-step procedure. Nelson proposes that the issue of sample selectivity be dealt with in the following manner: first, the correlation between the residuals in the outcome equation and those in the selection equation should be calculated using the inverse Mills Ratio. If the correlation is found to be small, then the effect of sample selection bias should be permissibly small and one can proceed using simple subsample

OLS. On the other hand, if the correlation is found to be substantial, then one should opt for using the ML estimator. In all those cases where the correlation is only moderate, the Heckman two-step procedure can also be prescribed, although ML estimators proved to perform equally well.

Stolzenberg and Relles (1990) did a similar study, focusing on precision rather than bias. They investigated the role of correlation between the regressors of the selection equation and the outcome equation. A sample size of 500 was chosen and the correlation between the regressors and the correlation between the residuals were both allowed to vary between 0 and 0.75. The variance of the first residual was also allowed to vary between 1/9 and 9. The choice to censor 90 percent of the observations however proved to be most influential in terms of their findings, resulting in peculiar results. Under most specifications the two-step model's estimates were no better than those calculated under the simpler subsample OLS method. A noticeable exception was under the rare circumstance where both the residuals and the regressors are highly correlated. In such cases the Heckman two-step estimator did outperform its OLS counterpart. However, even here the two-step estimator's mean squared error was regularly higher than that of the subsample OLS approach. The authors viewed this as further proof of the two-step model's instability and concluded by discouraging the use of this technique due to its lack of robustness.

From these two Monte Carlo studies, it is clear that the bias and precision of the sample selection models are heavily dependent on the following three factors:

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

For this reason, this study will proceed by repeating these Monte Carlo simulations for a specific range of parameters that corresponds to the South African labour force data.

The following equations were used to model the wage process.

$$\text{Selection equation: } d = \alpha x + \theta z + e$$

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

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 . Note that the values for the parameter ρ , the parameter θ and the ratio of the population for whom $d > \delta$ are determined exiguously. This allows us to compare the results obtained under different sets of specifications. β and σ_e were both set to one, because the efficiency and precision of subsample OLS and sample selectivity estimators are unaffected by the choice of β and behave similarly when the variance of e is either increased or decreased (Nelson, 1984: 190). The parameter α was also set to 1.

Testing on SA household survey data shows that ρ is approximately equal to 0.1 in absolute terms (see Appendix: Table 11, 12 and 13). In our model ρ was allowed to vary between 0, 0.5 and 1. Negative values were not allowed, since the focus here is on the degree of correlation rather than the direction of the relationship. The value of parameter θ serves as a measure of the degree of identification parameter. It denotes the correlation between the regressors that are in both equations and those regressors that are unique to the selection equation. This parameter was also allowed to vary between 0, 0.5 and 1.

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 Monte Carlo simulations. With this in mind, an alternative censoring value was chosen, one that

corresponds to a 66% employment rate. This allows us to test whether the severity of the censoring has a significant impact on the results.

Each simulation contained 10 000 observations and was repeated 100 times. In ideal circumstances the sample size and observations would both have been larger (selecting 69 101 observations to correspond to the number of working age individuals in the September 2005 Labour Force Survey and choosing 1000 repetitions for each set of parameters, for example). The excessive computational power required in handling such large samples made this impossible, however. The conventional OLS subsample method, the ML method, the Heckman two-step and the semi-parametric method were all tested. The mean and mean squared errors of the β 's were obtained by taking the average values over the 100 trials. Estimates of the mean estimates as well as the mean squared errors of these estimates are reported in Table A1 and A2 (Appendix). The standard errors are in parentheses.

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 ρ increase and that there exists no sample selection bias when $\rho = 0$. The role of identification is apparent when the ML estimates are compared with those of the Heckman two-step approach. Although both models succeed in correcting for the sample selection bias, the ML estimates generally appear to be more precise, judging by their lower overall mean squared error values. 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.

The degree of censoring also plays an important role. The results obtained in Table A1 and Table A2 suggest that 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 1/3 to 2/3. While both the two-step and ML method appear to be sufficiently precise under most circumstances, both of them also experience a substantial increase in their mean squared errors when the degree of censoring increases, rendering them less precise. Although the semi-parametric estimator succeeded in correcting sample selection bias, its estimated β 's performed worse than that of the other two sample selection methods, both of which recorded smaller biases and lower mean squared errors.

b) Normality

It is vital to also consider the implications of 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 is followed. The Monte Carlo test is conducted under the normality assumption as well as for a bivariate t and bivariate χ^2 distribution. Results for the normal distribution are reported in Table A1 and A2, while the results for the bivariate t-distribution and bivariate χ^2 -distribution, both with five degrees of freedom, are summarised in Table B1 and B2 and Table C1 and C2 (Appendix).

A t-distribution was introduced to our Monte Carlo simulations (see Table B1 and B2) in an attempt to establish how sensitive the parametric sample selection models are to deviation from normality. The two-step method was shown to be most sensitive to slight deviations from normality. The ML-estimator delivered far better estimates of β than the two-step estimator, especially in the absence of exclusion restrictions.

The opposite was true for the semi-parametric model. Its estimators outperformed the ML model's estimator for more than half of the set of parameter value assumptions. The

general trend appears to be that on average the ML performs better when identification is low and the correlation between the errors is high, while the semi-parametric method is superior when the identification is higher and the correlation between the errors is lower. As expected, the semi-parametric model does better relative to the other models when the normality assumption is relaxed.

The χ^2 -distribution was 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 Appendix (Table C1 and C2). The ML-method, which has performed well up to now, fares considerably worse. Notably, it is now the worst estimator for cases where ρ is high and θ is low, the same combination of conditions for which the ML-method yielded the best results in the previous model. We conclude that the ML-model is very sensitive to the absence of exclusion restrictions or identification when the error term is not symmetrically distributed along the y-axis.

The semi-parametric estimator performed well against both parametric estimators. Its estimates were far better than those of the ML-method. Compared to the two-step models, the semi-parametric models appear to be less dependent on the existence of valid identification, since the two-step model delivered worse β estimates than the semi-parametric estimator in those cases where exclusion restrictions were lacking. Where valid exclusion restrictions did however exist, there was not much difference between the β 's obtained by the semi-parametric and those obtained from the two-step method. The semi-parametric model's mean squared errors remain inferior to that of the two-step method.

Under all three distributions we have found that relative to the other sample selection models, the semi-parametric approach's β estimators are considerably more robust, while the performance of its mean squared errors is disappointing. Since both the β and the mean squared error are estimated over a sample size of 10 000 observations, one would expect each pair of point estimates to be of a similar quality. As Monte Carlo simulations are run and the process is repeated, the estimated value of β is supposed to converge more

closely around β , since the semi-parametric method provides consistent estimates.³ Meanwhile, the mean squared error's expected value remains fixed as the process is repeated. What is seen here (the semi-parametric method's superior β estimates and inferior mean squared errors relative to the other models) is thus in accordance with the estimates' consistency property, which predicts that the semi-parametric model's estimates will become more precise as the size of the sample increases.

5. 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 have little or no impact on his or her wage. Few variables meet this criterion. According to Puhani (2000: 58), household variables are considered to be most apt for use as exclusion restrictions in labour market analysis, since these variables are more likely to effect the participation decision 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 correlated with the wage function.

The following household variables were tested to inquire whether they can be used as exclusion restrictions:

- a) the number of children in the household (where children are considered to be those aged 18 or younger),
- b) the number of elderly in the household (where elderly are considered to be those aged 65 or older),
- c) a dummy variable indicating whether an individual is the head of the household or not,

³ This point has been demonstrated by Lee (Vella, 1998: 143).

- d) the number of people in the household,
- e) a dummy variable denoting an individual's marital status,
- f) the number of employed individuals in the household (apart from the individual being surveyed),
- g) the number of narrowly defined unemployed individuals in the household (apart from the specific individual), and
- h) the number of broadly defined unemployed individuals in the household (apart from the specific individual).

They were tested in turn, by adding each to the selection equation and wage functions and analyzing its t-value. Both the selection and wage equations controlled for education, experience, race, gender, province and whether the individual resides in a rural or urban area. The wage function contains an additional regressor, a dummy, which represents union membership. This variable however does little to improve identification, since the colinearity between the other regressors in the outcome equation and the inverse Mills ratio remains unchanged.

The hope to interpret the relationship between the left-hand and the right-hand variables as one of causation, limits the set of candidate variables to be used for the exclusion restrictions. As statisticians often warn, it is vital to note that the partial correlation of a variable with the employment variable, gives no information about the direction of the causality. It is possible that the direction could work in either or even both ways. It is for instance, quite likely that an individual's employment status could effect 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.

Preliminary tests were performed to enquire whether we can find a household variable that is significant in the selection equation but insignificant in the wage function. The tests were administrated independently, each household variable was added 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 results are summarised in Table 1a and 1b of the Appendix. All the variables had a significant effect on the participation choice, at a 5% level of significance. In the case of the wage function, only three of the household variables were found to be insignificant. These three variables, which contained the number of children, the number of employed persons as well as the total number of people in each household were tested further to examine whether they could be used as exclusion restrictions.

Note that the coefficients that the household variables obtained in the outcome equation under the semi-parametric model were omitted from the tables in the appendix, since the estimated β values failed to converge due to the lack of exclusion restrictions. Although this emphasizes the models sensitivity to an exclusion restriction, it does put us in a difficult position. The validity of a specific exclusion restriction can not be tested unless we make use of another exclusion restriction to allow for identification within the model. Since the choice of an initial exclusion restriction may affect the results, it cannot be chosen at whim.

Although the semi-parametric estimates are omitted, it is interesting to note that the β estimates for the household variables usually converged on the OLS estimates, although the β 's for some of the other control variables variables (education, experience, intercept, etc.) failed to converge at all. In this sense, the OLS estimates offer some information in deciding which variables to use as exclusion restrictions for the semi-parametric model.

a) Children

The variable capturing the number of children per household was broken down according to the age and gender of the children. Boys aged 0 to 5, 6 to 12 and 13 to 18 and girls aged 0 to 5, 6 to 12 and 13 to 18 were classified as separate groups. The effect that a child has on an individual's selection choice was allowed to vary depending on the sex of the parent. The rationale for this is that a young child would be expected to have a greater impact on the participation choice of female parents than on the labour market decisions

of male parents, since traditional intra-household division of work arrangements usually involve that females act as caretakers for young children.

The results in Table 2a and 2b (Appendix) were obtained by allowing gender and the six children variables to act as interaction variables for all estimation techniques. Ten of the 12 sets of variables were found to have a significant effect in the selection equation at a 5% significance level, providing some proof that the presence of children impact on other household members' likelihood of obtaining work. Table 3 shows that at a 5% level of significance, none of these variables had a significant effect on wages. The finding that the individual wage is not significantly affected by the number of children in the household suggests that the household's number of children does not enter the individual's wage function and that the wage attained does not appear to have an effect on the number of children in a household.

The estimated coefficients of the ML and semi-parametric selection equation are illustrated in Figure 3a and 3b. Judging by the similarity between the two pairs of line-graphs, the gender of the parent plays a much larger role in the employment outcome than the gender of the child. We tested to see whether the coefficient on the two dashed lines in Figures 3a and 3b (indicating the effect of boys and girls on the selection into employment of females) differs significantly. A simple linear Wald-test was applied. The test was repeated for the two other lines - the effect of boys and girl on selection of males. The hypothesis that the coefficients did not differ significantly from one another could not be rejected for any of the six pairs of points at a 5% level of significance when a probit function was used to estimate the employment outcome (Appendix: Table 3a and 3b). This was also the case for the semi-parametric variant of the probit function (Appendix: Table 3e and 3f). For the ML-method (Table 3c, 3d: Appendix) one pair of coefficients, those denoting the effect of boys and girls aged between 6 and 12 on a female's likelihood of obtaining work, differed significantly. The difference is insignificant at a 1% level of significance,

The fact that boys and girls essentially have the same effect on the selection equation allows us to treat the boys and girls within each of the three age groups as homogenous. The coefficients of the simplified model, which groups boys and girls together, can be seen in Figure 4. Table 4a shows the t-statistics. None of these household variables are partially correlated with the wage function and only the first variable, i.e. the effect of young children on the selection of males, is insignificant in the employment decision.

The robustness of the exclusion restriction to the choice of datasets was tested using different datasets. The Labour Force Survey datasets for September 2003 and September 2001 were used. The t-values for both these samples (see Tables 4b and 4c in the appendix) were found to be similar to that of 2005 (Appendix: Table 4a). For all three the samples, all the variables apart from the first (the effect of young children on the selection choices of males) were found to have a significant effect on the employment outcome at a 5% level of significance, regardless of what sample selection method we used. The effect of the children variables proved to be less robust with regard to the wage function, where the last two variables in 2001 (Appendix: Table 4c) and the last variable in 2003 (Appendix: Table 5b) were found to be correlated with the wage function, at a 5% level of significance. This was not the case for the 2005 sample (Appendix: Table 4a), where none of the variable were significantly correlated with the wage function, at a 5% level of significance. The estimated coefficients (see Figure 5a and 5b) look more robust for males than for females, since their line-graphs for males are more closely matched than those of the females. The slope of the female graph is positive for 2005 and 2001, while it is negative for 2003.

It appears as though these six variables may 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. They also behaved consistently under stringent empirical testing.

b) Number of employed individuals

It makes sense that the number of employed individuals in a household may provide us with an important insight into the household circumstances of individuals, serving as an indication of both the degree of labour market attachment as well as the unfulfilled demand of the household. While having fellow household members who work may prove helpful when searching for a job, it is also likely that the additional paycheck may lessen the pressure on other household members to obtain work.

In the preliminary testing, the selection variable was shown to be partially unrelated to the wage function (Table 1 in the Appendix). The variable was continuous in those tests, taking on a value between 0 and 13. The test was repeated, this time allowing the marginal effect of each of the first, second and third employed person in a household to differ. The idea behind this is straightforward: one would expect the benefit of having one employed person within a household rather than none to be greater than the additional benefit of having a second or third breadwinner (Appendix: Table 7a and 7b). The coefficients and t-statistics within the participation choice appear to be robust. Having additional breadwinners appears to have a negative influence on one's participation choice. Unfortunately, the marginal effect of the first employed household member is significantly correlated with the wage function, prohibiting us to use it as an exclusion restriction.

c) Household Size

Another variable which was shown to be highly significant in the selection equation and insignificant in the outcome equation, was the household size variable (see Table 1 or Table 8 in Appendix). The household variable was modified to exclude the number of children in the household as to investigate to what extent the household size variable's effect shown in Table 1 was driven by the children variable. The adjusted household size variable was tested. The estimated coefficients and t-statistics (parenthesis) are in Table 8 in the Appendix. The t-statistics suggest that the adjusted household variable had a significant effect on the selection outcome, regardless of what sample selection process was used.

In Table 9a and 9b, the household size variable was added to our earlier model containing the six children variables. The differences in the coefficient and t-statistics under combined testing is similar to what they were when they were tested separately. The coefficients suggest that the effect that an additional child has on an individual's likelihood to attain work is smaller than the effect of an additional household member who is older than 18.

It is not clear whether the inclusion of an adjusted household size variable would improve estimation. The weak correlation with the wage function, significant at a 5% level, but insignificant at 1%, may be evidence enough to dismiss its use as a exclusion restriction. An argument could, however, be made in favour of the inclusion of the variable, since its t-value in the participation equation is far larger than that of any of the other household variables that we considered. Due to the difficulty in deciding whether or not to include the household size variable, section 7 will compare the effect of including the variable.

6. Testing normality

In section 4, it was established with the aid of Monte Carlo simulations 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) 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.

The first step is to derive the standardised residuals of the probit function. The first four moments, which denote the mean, variance, skewness and kurtosis, are 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:

$$e^{(1)} = -(1-d)\lambda(\alpha x) + d\lambda(-\alpha x)$$

$$\begin{aligned}\hat{e}^{(2)} &= -\alpha x \hat{e}^{(1)} \\ \hat{e}^{(3)} &= (2 + (\alpha x)) \hat{e}^{(1)} \\ \hat{e}^{(4)} &= -(3\alpha x + (\alpha x)^3) \hat{e}^{(1)}\end{aligned}$$

where $d = 1$ if the individual is employed and $d = 0$ if the individual is not employed,
 αx is the linear prediction of the fitted model,
and $\lambda(\cdot)$ 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 $\hat{e}^{(1)}x_l$. One can then proceed in two manners: either constructing a larger matrix L , consisting of $\hat{e}^{(1)}x_l$, $\hat{e}^{(3)}$ 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).

The LM statistics from this test are included in Table 10 (Appendix). The LM statistics 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 are significantly higher than their corresponding critical values, which vary between 35 and 48.

These tests were repeated for certain subsections of the population. Table 11 reports these findings and show 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 exceed their critical values at a 5% level of significance. The last group, Indian females, come closest to being normally distributed. It has a LM-statistic of 21.9 and a critical value of 21.03 when no exclusion restrictions are used.

7. 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. Three sets of exclusion restrictions were implemented: the first model does not make use of any exclusion restrictions, the second model employs the children variables as exclusion restrictions and the third model uses children and household size as exclusion restrictions. The results were tabulated in Tables 12, 13 and 14 in the Appendix. Note that we were unable to attain results for the outcome equation using a semi-parametric estimator in Table 12. As was the case in the Monte Carlo simulation, the non-parametric method was again shown to be highly dependent on the existence of a valid exclusion restriction. Due to the colinearity in the first model, where no exclusion restrictions were used, the estimated β values failed to converge.

The ML estimator and two-step estimators deliver similar estimates of α in the selection equation. This is somewhat surprising since these models use different techniques to derive these estimates. The two-step estimator uses 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 participation 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. Also note that the difference between the parametric and semi-parametric models' estimated outcomes in the selection equation decreased as the amount

of exclusion restrictions increased, displaying much more similarity in Table 14 than in Table 13 and Table 12.

In Table 12, the estimated coefficients of the outcome equation of the two-step model differ considerably from those obtained using the conventional OLS method and the ML method. It seems plausible that this contradicting behavior displayed by the two-step method may be due to the colinearity among between the regressors of the wage and selection functions. The more realistic two-step results obtained in Table 13 and Table 14, where exclusion restrictions are present, appears to support such an argument. The effect of the exclusion restriction can also be seen by comparing the inverse Mills ratios, denoted by λ . They differ significantly, showing much smaller coefficient values with the addition of control variables. The inverse Mills ratio was found to be significant at a 5% level of significance for the first model where no exclusion restrictions were applied and insignificant for the other two models that did include exclusion restrictions.

The semi-parametric β estimates obtained in Table 13 and 14 fail 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 could be 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 $\rho = 0$ (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 obtains 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.

8. 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 their 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.

9. References

- BERK, R.A. and RAY. S.C., 1982. Selection Biases in Sociological Data. *Social Science Research*. 11: 352-398.
- BLAU, D.M., 1985. Self-Employment and Self-Selection in Developing Country Labor Markets. *Southern Economic Journal*. 52: 351-363.
- BREEN, R., 1996. Regression Models: Censored, Sample Selected or Truncated Data. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-111. Thousand Oaks, CA: Sage
- COSSLETT, S., 1983. Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model. *Econometrica*. 51. 3: 765-82.
- 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.
- GUJARATI, D.N., 2003. *Basic Econometrics*. 4th ed. Boston: McGraw-Hill.
- 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.
- NEWY. 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.

Appendix

a) Figures

Figure 1: The quasi-linearity of the Inverse Mills Ratio

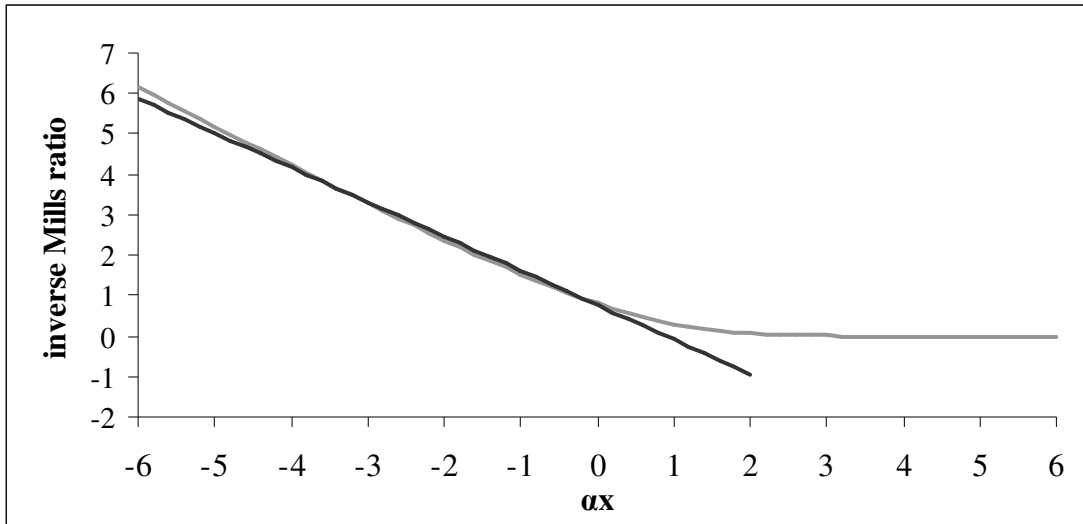


Figure 2: Iterative Procedure's estimates using different starting values

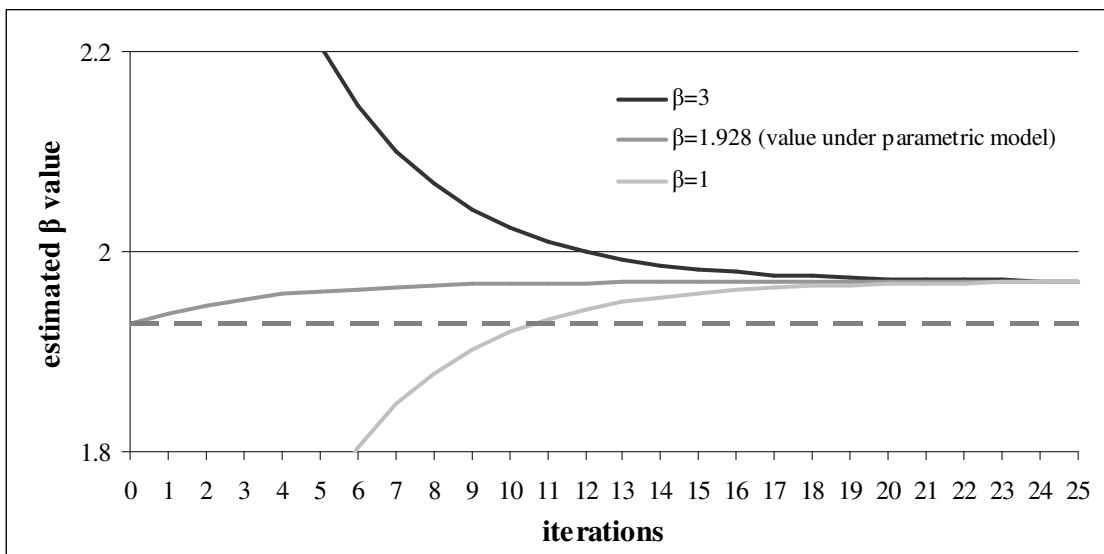


Figure 3a: Effect of children on employment outcome, using ML method

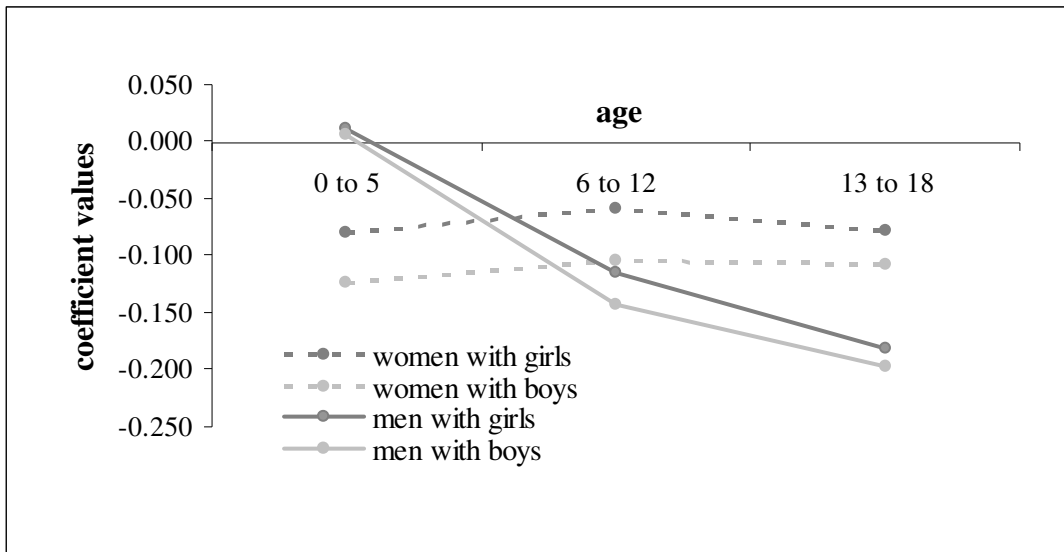


Figure 3b: Effect of children on employment outcome, using SP method

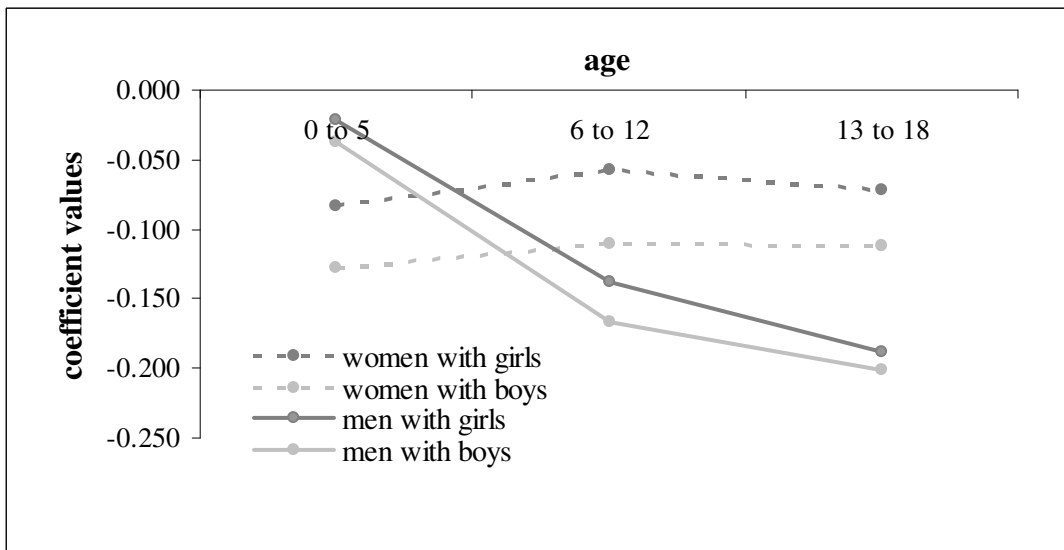


Figure 4: Effect of children on employment outcome, using ML and SL method

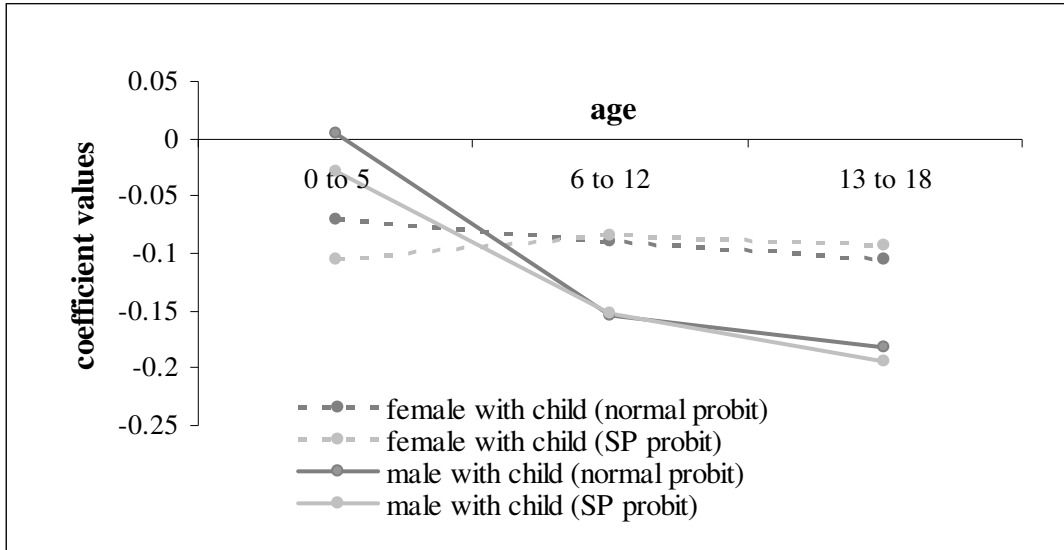


Figure 5a: Effect of children on employment outcome for 3 samples, using ML method

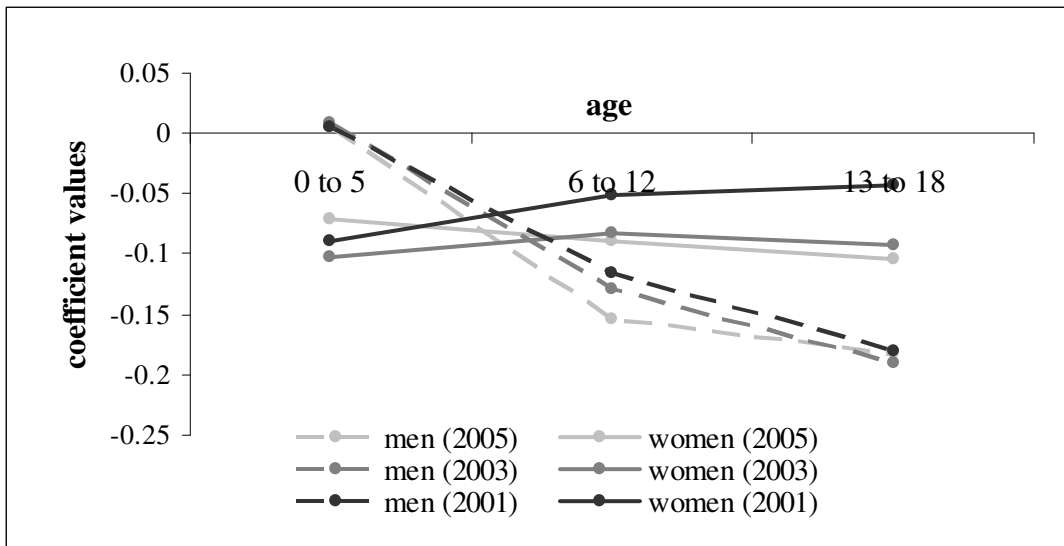
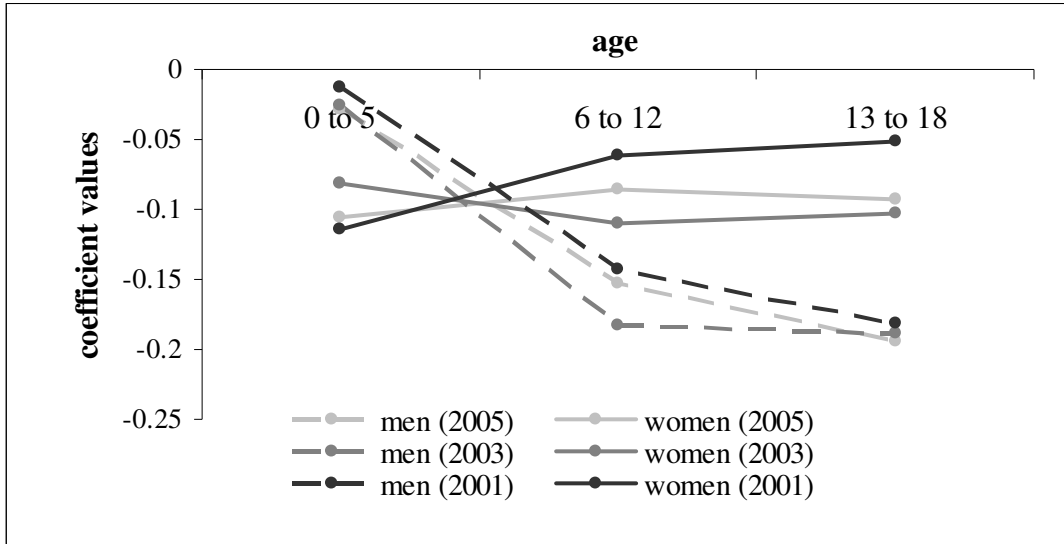


Figure 5b: Effect of children on employment outcome for 3 samples, using SP method



a) Tables

Table 1a: Exclusion Restriction Testing: All Household Variables, coefficients

<i>Variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
children	-0.105	-0.105	-0.122	0.001	0.000	0.001
elderly	-0.385	-0.387	-0.389	-0.079	-0.077	-0.078
head	0.795	0.796	0.925	0.083	0.080	0.083
household size	-0.095	-0.095	-0.101	-0.005	-0.006	-0.005
married	0.215	0.228	0.215	0.113	0.113	0.113
employed	-0.039	-0.390	-0.033	0.007	0.006	0.007
unemployed (narrow def)	-0.168	-0.169	-0.186	-0.040	-0.040	-0.040
unemployed (broad def)	-0.199	-0.201	-0.213	-0.052	-0.052	-0.052

Table 1b: Exclusion Restriction Testing: All Household Variables, t-statistics

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
children	-20.56	-23.82	-19.88	0.15	0.00	0.13
elderly	-24.32	-28.09	-18.58	-4.23	-4.11	-4.22
head	42.09	53.56	22.70	4.05	3.85	4.01
household size	-29.49	-33.53	-21.73	-1.70	-1.90	-1.73
married	11.65	15.38	11.65	5.81	5.80	-5.81
employed	3.86	4.43	3.13	0.78	0.69	0.76
unemployed (narrow def)	-15.97	-18.65	-16.28	-3.42	-3.39	-3.42
unemployed (broad def)	-23.50	-27.11	-20.58	-5.34	-5.32	-5.34

Table 2a: Exclusion Restriction Testing: Children, coefficients

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with boy, aged 0 to 5	0.006	0.007	-0.037	0.015	0.017	0.015
Male with boy, aged 6 to 12	-0.143	-0.143	-0.167	-0.009	-0.010	-0.009
Male with boy, aged 13 to 18	-0.198	-0.198	-0.201	-0.018	-0.022	-0.018
Male with girl, aged 0 to 5	0.011	0.011	-0.021	0.017	0.020	0.017
Male with girl, aged 6 to 12	-0.115	-0.114	-0.138	0.032	0.030	0.032
Male with girl, aged 13 to 18	-0.182	-0.182	-0.188	0.015	0.010	0.014
Female with boy, aged 0 to 5	-0.124	-0.124	-0.127	-0.030	-0.031	-0.030
Female with boy, aged 6 to 12	-0.105	-0.105	-0.110	-0.037	-0.037	-0.037
Female with boy, aged 13 to 18	-0.107	-0.107	-0.112	-0.017	-0.018	-0.017
Female with girl, aged 0 to 5	-0.080	-0.080	-0.083	-0.014	-0.016	-0.014
Female with girl, aged 6 to 12	-0.058	-0.058	-0.057	0.023	0.021	0.023
Female with girl, aged 13 to 18	-0.078	-0.078	-0.072	-0.019	-0.020	-0.019

Table 2b: Exclusion Restriction Testing: Children, t-statistics

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with boy, aged 0 to 5	0.30	0.38	-1.68	0.60	0.71	0.62
Male with boy, aged 6 to 12	-7.94	-9.65	-9.27	-0.43	-0.47	-0.44
Male with boy, aged 13 to 18	-10.64	-12.55	-10.44	-0.92	-1.12	-0.95
Male with girl, aged 0 to 5	0.52	0.65	-0.99	0.68	0.79	0.69
Male with girl, aged 6 to 12	-6.44	-7.91	-7.83	1.61	1.52	1.60
Male with girl, aged 13 to 18	-9.10	-10.80	-9.05	0.60	0.41	0.57
Female with boy, aged 0 to 5	-6.08	-7.08	-6.00	-1.04	-1.10	-1.05
Female with boy, aged 6 to 12	-5.89	-6.76	-5.69	-1.74	-1.73	-1.74
Female with boy, aged 13 to 18	-6.09	-7.18	-5.52	-0.80	-0.85	-0.81
Female with girl, aged 0 to 5	-4.01	-4.60	-4.01	-0.50	-0.59	-0.52
Female with girl, aged 6 to 12	-3.45	-3.98	-3.09	1.07	1.01	1.06
Female with girl, aged 13 to 18	-4.41	-5.24	-3.92	-0.78	-0.82	-0.78

Table 3a: Probability that child's gender plays a role in employment outcome of males

<i>variable</i>	<i>Child's age</i>		
	<i>0 to 5</i>	<i>6 to 12</i>	<i>13 to 18</i>
Chi-squared-statistic	0.02	1.28	0.35
Probability	0.8831	0.2587	0.5534

Table 3b: Probability that child's gender plays a role in employment outcome of females

<i>variable</i>	<i>Child's age</i>		
	<i>0 to 5</i>	<i>6 to 12</i>	<i>13 to 18</i>
Chi-squared-statistic	2.33	3.41	1.42
Probability	0.1271	0.0648	0.233

Table 3c: Probability that child's gender plays a role in employment outcome of males

<i>variable</i>	<i>Child's age</i>		
	<i>0 to 5</i>	<i>6 to 12</i>	<i>13 to 18</i>
Chi-squared-statistic	0.03	2.06	0.51
Probability	0.8577	0.1508	0.4771

Table 3d: Probability that child's gender plays a role in employment outcome of females

<i>variable</i>	<i>Child's age</i>		
	<i>0 to 5</i>	<i>6 to 12</i>	<i>13 to 18</i>
Chi-squared-statistic	3.17	4.51	1.99
Probability	0.0748	0.0337	0.1585

Table 3e: Probability that child's gender plays a role in employment outcome of males

<i>variable</i>	<i>Child's age</i>		
	<i>0 to 5</i>	<i>6 to 12</i>	<i>13 to 18</i>
Chi-squared-statistic	0.26	1.38	0.21
Probability	0.6134	0.2402	0.6467

Table 3f: Probability that child's gender plays a role in employment outcome of females

<i>variable</i>	<i>Child's age</i>		
	<i>0 to 5</i>	<i>6 to 12</i>	<i>13 to 18</i>
Chi-squared-statistic	2.36	3.63	2.22
Probability	0.1248	0.0567	0.1363

Table 4a: Exclusion Restriction Testing: Children, coefficients for 2005

<i>Variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.005	0.005	-0.029	0.017	0.019	0.017
Male with child, aged 6 to 12	-0.153	-0.152	-0.153	0.011	0.010	0.011
Male with child, aged 13 to 18	-0.181	-0.181	-0.194	-0.002	-0.007	-0.003
Female with child, aged 0 to 5	-0.070	-0.070	-0.105	-0.020	-0.022	-0.021
Female with child, aged 6 to 12	-0.089	-0.089	-0.085	-0.007	-0.007	-0.007
Female with child, aged 13 to 18	-0.105	-0.107	-0.092	-0.019	-0.020	-0.019

Table 4b: Exclusion Restriction Testing: Children, t-statistics for 2005

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.38	0.46	-1.89	1.01	1.17	1.03
Male with child, aged 6 to 12	-13.35	-15.77	-11.85	0.81	0.71	0.80
Male with child, aged 13 to 18	-14.63	-17.26	-13.69	-0.15	-0.44	-0.19
Female with child, aged 0 to 5	-5.60	-6.40	-6.86	-1.13	-1.23	-1.14
Female with child, aged 6 to 12	-8.20	-9.31	-6.62	-0.48	-0.52	-0.49
Female with child, aged 13 to 18	-9.01	-10.45	-6.51	-1.16	-1.22	-1.17

Table 5a: Exclusion Restriction Testing: Children, coefficients for 2003

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.009	0.009	-0.026	0.005	0.005	0.005
Male with child, aged 6 to 12	-0.129	-0.129	-0.182	0.006	0.006	0.006
Male with child, aged 13 to 18	-0.190	-0.190	-0.188	-0.016	-0.016	-0.016
Female with child, aged 0 to 5	-0.102	-0.102	-0.081	0.002	0.001	0.002
Female with child, aged 6 to 12	-0.082	-0.082	-0.111	-0.005	-0.004	-0.005
Female with child, aged 13 to 18	-0.093	-0.093	-0.104	-0.036	-0.042	-0.042

Table 5b: Exclusion Restriction Testing: Children, t-statistics for 2003

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.65	0.81	-1.60	0.33	0.35	0.34
Male with child, aged 6 to 12	-10.20	-12.15	-12.70	0.48	0.47	0.48
Male with child, aged 13 to 18	-14.02	-16.60	-13.47	-1.15	-1.17	-1.16
Female with child, aged 0 to 5	-7.21	-8.24	-5.91	0.10	0.09	0.10
Female with child, aged 6 to 12	-6.97	-8.12	-9.06	-0.31	-0.30	-0.31
Female with child, aged 13 to 18	-7.39	-8.75	-7.95	-2.93	-2.82	-2.81

Table 6a: Exclusion Restriction Testing: Children, coefficients for 2001

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.006	0.007	-0.013	0.009	0.006	0.007
Male with child, aged 6 to 12	-0.116	-0.116	-0.143	-0.010	-0.009	-0.010
Male with child, aged 13 to 18	-0.180	-0.180	-0.181	-0.017	-0.013	-0.014
Female with child, aged 0 to 5	-0.089	-0.087	-0.114	0.009	0.010	0.010
Female with child, aged 6 to 12	-0.051	-0.053	-0.062	-0.035	-0.035	-0.035
Female with child, aged 13 to 18	-0.043	-0.046	-0.051	-0.036	-0.035	-0.035

Table 6b: Exclusion Restriction Testing: Children, t-statistics for 2001

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.46	0.62	-0.98	0.67	0.43	0.56
Male with child, aged 6 to 12	-10.59	-12.43	-12.58	-0.85	-0.78	-0.83
Male with child, aged 13 to 18	-15.25	-17.70	-13.87	-1.23	-0.98	-1.02
Female with child, aged 0 to 5	-8.14	-9.26	-8.77	0.60	0.67	0.73
Female with child, aged 6 to 12	-5.37	-6.46	-5.39	-3.03	-3.06	-3.06
Female with child, aged 13 to 18	-4.25	-5.27	-4.46	-2.93	-2.85	-2.87

Table 7a: Exclusion Restriction Testing: Number of Employed People, coefficients

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
<i># of employed people = 1</i>	-0.136	-0.143	-0.169	0.089	0.088	0.089
<i># of employed people = 2</i>	-0.118	-0.121	-0.138	0.035	0.034	0.035
<i># of employed people >= 3</i>	-0.061	-0.051	0.003	-0.081	-0.084	-0.081

Table 7b: Exclusion Restriction Testing: Number of Employed People, t-statistics

<i>Variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
<i># of employed people = 1</i>	-7.72	-10.12	-9.07	4.71	4.64	4.70
<i># of employed people = 2</i>	-4.34	-5.48	-4.97	1.26	1.21	1.26
<i># of employed people >= 3</i>	-1.38	-1.33	0.07	-1.79	-1.86	-1.80

Table 8: Exclusion Restriction Testing: Household Size, coefficients and t-statistics

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
adjusted household size	-0.165	-0.165	-0.162	-0.014	-0.015	-0.014
	-28.82	-32.67	-22.98	-2.52	-2.63	-2.53

Table 9a: Exclusion Restriction Testing: Children & Household Size, coefficients

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	0.072	0.073	0.051	0.023	0.025	0.023
Male with child, aged 6 to 12	-0.091	-0.091	-0.100	0.015	0.014	0.015
Male with child, aged 13 to 18	-0.135	-0.135	-0.133	0.005	0.000	0.004
Female with child, aged 0 to 5	-0.061	-0.062	-0.065	-0.015	-0.017	-0.015
Female with child, aged 6 to 12	-0.057	-0.057	-0.053	-0.003	-0.004	-0.004
Female with child, aged 13 to 18	-0.052	-0.052	-0.048	-0.014	-0.015	-0.014
adjusted household size	-0.144	-0.144	-0.136	-0.016	-0.016	-0.016

Table 9b: Exclusion Restriction Testing: Children & Household Size, t-statistics

<i>variable</i>	<i>selection equation</i>			<i>outcome equation</i>		
	<i>2 Step</i>	<i>ML</i>	<i>SP</i>	<i>OLS</i>	<i>2 Step</i>	<i>ML</i>
Male with child, aged 0 to 5	4.88	6.08	3.26	1.34	1.51	1.36
Male with child, aged 6 to 12	-7.11	-8.41	-7.81	1.11	1.01	1.10
Male with child, aged 13 to 18	-9.86	-11.68	-9.79	0.30	0.01	0.26
Female with child, aged 0 to 5	-4.16	-4.81	-3.56	-0.81	-0.91	-0.82
Female with child, aged 6 to 12	-4.62	-5.40	-3.91	-0.24	-0.28	-0.25
Female with child, aged 13 to 18	-4.01	-4.78	-3.39	-0.85	-0.90	-0.86
adjusted household size	-23.89	-27.41	-18.66	-2.57	-2.59	-2.56

<i>Control Variables</i>	<i>none</i>	<i>Children</i>	<i>Children & HHSize</i>
MSE	1453.19	1459.40	1376.94
df	19	25	26
R ²	0.021	0.021	0.020
observations	68735	68735	68735

Table 11: Normality Testing: Subsamples of Working Age Population

<i>variable</i>	<i>Nochildren</i>		<i>children</i>		<i>children&hhsiz</i>	
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.

Table 12: Model Testing, no exclusion restriction

<i>Variable</i>	<i>wage equation</i>				<i>participation equation</i>		
	<i>OLS</i>	<i>ML</i>	<i>2 Step</i>	<i>SP</i>	<i>ML</i>	<i>2 Step</i>	<i>SP</i>
Experience	0.026 (14.55)	0.034 (7.36)	0.085 (3.73)	.	0.119 (72.33)	0.119 (61.95)	0.126 (38.03)
Experience2	-0.107 (-3.18)	-0.243 (-3.04)	-1.078 (-2.89)	.	-1.893 (-55.2)	-1.892 (-47.56)	-1.889 (-29.6)
Education	0.133 (64.76)	0.138 (37.28)	0.164 (14.14)	.	0.066 (29.65)	0.066 (24.42)	0.075 (26.29)
Female	-0.273 (-23.35)	-0.303 (-13.89)	-0.484 (-5.92)	.	-0.440 (-33.95)	-0.440 (-27.69)	-0.474 (-25.2)
Rural	-0.222 (-14.13)	-0.238 (-10.95)	-0.333 (-7.1)	.	-0.218 (-13.82)	-0.218 (-11.72)	-0.254 (-12.68)
White	1.001 (54.65)	0.998 (29.6)	0.983 (27.61)	.	-0.022 (-0.78)	-0.022 (-0.65)	0.262 (3.96)
Coloured	0.269 (12.1)	0.282 (8.25)	0.360 (7.96)	.	0.200 (7.6)	0.200 (5.99)	0.204 (4.75)
Indian	0.763 (22.61)	0.759 (15.14)	0.738 (14.69)	.	-0.048 (-1.05)	-0.046 (-0.84)	0.074 (0.99)
Union	0.656 (48.06)	0.656 (35.99)	0.657 (36.03)	.			
λ			0.747 (2.68)				
ρ		0.121 (2.20)					
Constant	0.184 (5.08)	-0.028 (-0.28)	-1.336 (-2.36)	.	-1.858 (-48.68)	-1.854 (-39.94)	-1.853 (fixed)

(pseudo) R ²	0.478		0.478	.		0.170	
Observations	22960	22960	22960	22960	68735	68735	68735

Table 13: Model Testing, Children as exclusion restriction

<i>Variable</i>	<i>wage equation</i>				<i>participation equation</i>		
	<i>OLS</i>	<i>ML</i>	<i>2 Step</i>	<i>SP</i>	<i>ML</i>	<i>2 Step</i>	<i>SP</i>
Experience	0.026 (14.55)	0.025 (5.77)	0.025 (4.41)	0.042 (23.81)	0.119 (72.55)	0.119 (61.87)	0.122 (41.28)
Experience2	-0.107 (-3.18)	-0.101 (-1.33)	-0.099 (-1.02)	-0.393 (-11.93)	-1.923 (-56.01)	-1.923 (-47.98)	-1.847 (-30.73)
Education	0.133 (64.76)	0.133 (36.27)	0.133 (33.02)	0.129 (65.06)	0.061 (27.86)	0.061 (22.54)	0.068 (25.34)
Female	-0.273 (-23.35)	-0.272 (-12.74)	-0.271 (-10.87)	-0.318 (-27.36)	-0.411 (-21.59)	-0.411 (-17.54)	-0.470 (-18.49)
Rural	-0.222 (-14.13)	-0.222 (-10.5)	-0.222 (-10.1)	-0.208 (-14.92)	-0.143 (-9.04)	-0.143 (-7.59)	-0.160 (-8.27)
White	1.001 (54.65)	1.001 (29.7)	1.001 (29.62)	1.019 (49.26)	-0.086 (-3.01)	-0.086 (-2.53)	0.200 (2.89)
Coloured	0.269 (12.1)	0.269 (7.9)	0.269 (7.87)	0.239 (11.81)	0.213 (8.21)	0.213 (6.35)	0.208 (4.85)
Indian	0.763 (22.61)	0.763 (15.41)	0.763 (15.39)	0.733 (19.4)	-0.097 (-2.13)	-0.097 (-1.76)	0.001 (0.01)
Union	0.656 (48.06)	0.656 (35.99)	0.656 (35.96)	0.731 (52.61)			
m1					0.009 (0.81)	0.009 (0.65)	-0.029 (-1.89)
m2					-0.129 (-12.15)	-0.129 (-10.2)	-0.153 (-11.85)
m3					-0.190 (-16.6)	-0.190 (-14.02)	-0.194 (-13.69)
f1					-0.102 (-8.24)	-0.102 (-7.21)	-0.105 (-6.86)
f2					-0.082 (-8.12)	-0.082 (-6.97)	-0.085 (-6.62)
f3					-0.093 (-8.75)	-0.093 (-7.39)	-0.092 (-6.51)
λ			-0.007 (-0.11)	-0.017 (-0.25)			
ρ		-0.005 (0.01)					
constant	0.184 (5.08)	0.193 (1.96)	0.198 (1.49)	0.207 (1.99)	-1.672 (-42.85)	-1.672 (-34.71)	-1.672 (fixed)
(pseudo) R ²	0.478		0.478	0.483		0.183	
observations	22960	22960	22960	22960	68735	68735	68735

Table 14: Model Testing, Children & Household Size as exclusion restrictions

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

Table A1: Monte Carlo simulation, with normal distribution and 33% employment

	β -estimate			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.693 (0.027)	0.528 (0.024)	0.436 (0.025)	$\rho = 1$
<i>MLE:</i>	1.001 (0.029)	0.992 (0.036)	1.000 (0.05)	
<i>2Step:</i>	1.003 (0.032)	0.991 (0.046)	1.007 (0.114)	
<i>SP:</i>	1.001 (0.033)	0.991 (0.055)	1.005 (0.114)	
<i>OLS:</i>	0.846 (0.02)	0.770 (0.023)	0.715 (0.021)	$\rho = 0.5$
<i>MLE:</i>	0.998 (0.023)	1.008 (0.035)	0.992 (0.061)	
<i>2Step:</i>	0.999 (0.023)	1.006 (0.038)	0.995 (0.092)	
<i>SP:</i>	0.999 (0.025)	1.007 (0.047)	0.994 (0.092)	
<i>OLS:</i>	0.998 (0.018)	1.002 (0.021)	0.999 (0.023)	$\rho = 0$
<i>MLE:</i>	0.998 (0.023)	1.005 (0.035)	1.012 (0.097)	
<i>2Step:</i>	0.998 (0.023)	1.005 (0.035)	1.012 (0.101)	
<i>SP:</i>	0.995 (0.026)	0.999 (0.043)	1.013 (0.101)	

	Mean Squared Error			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.0948 (0.0166)	0.2236 (0.0227)	0.3189 (0.0284)	$\rho = 1$
<i>MLE:</i>	0.0009 (0.0012)	0.0013 (0.0019)	0.0024 (0.0029)	
<i>2Step:</i>	0.0010 (0.0015)	0.0021 (0.003)	0.0130 (0.0162)	
<i>SP:</i>	0.0011 (0.0016)	0.0030 (0.0038)	0.0130 (0.0158)	
<i>OLS:</i>	0.0241 (0.0063)	0.0536 (0.0108)	0.0819 (0.0121)	$\rho = 0.5$
<i>MLE:</i>	0.0005 (0.0009)	0.0013 (0.0019)	0.0037 (0.0049)	
<i>2Step:</i>	0.0005 (0.0008)	0.0015 (0.0021)	0.0084 (0.01)	
<i>SP:</i>	0.0006 (0.0009)	0.0023 (0.0032)	0.0084 (0.0101)	
<i>OLS:</i>	0.0003 (0.0005)	0.0004 (0.0008)	0.0005 (0.0007)	$\rho = 0$
<i>MLE:</i>	0.0005 (0.0008)	0.0012 (0.0017)	0.0095 (0.0121)	
<i>2Step:</i>	0.0005 (0.0008)	0.0013 (0.0018)	0.0103 (0.013)	
<i>SP:</i>	0.0007 (0.0009)	0.0018 (0.0027)	0.0102 (0.0131)	

Table A2: Monte Carlo simulation, with normal distribution and 66% employment

	β -estimate			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.776 (0.016)	0.684 (0.019)	0.643 (0.018)	$\rho = 1$
<i>MLE:</i>	1.000 (0.018)	0.999 (0.027)	1.004 (0.023)	
<i>2Step:</i>	0.999 (0.021)	0.999 (0.032)	0.999 (0.045)	
<i>SP:</i>	1.001 (0.024)	0.996 (0.042)	0.997 (0.045)	
<i>OLS:</i>	0.884 (0.017)	0.845 (0.016)	0.820 (0.016)	$\rho = 0.5$
<i>MLE:</i>	0.997 (0.018)	0.999 (0.023)	1.000 (0.028)	
<i>2Step:</i>	0.996 (0.019)	1.000 (0.025)	1.005 (0.045)	
<i>SP:</i>	0.995 (0.021)	0.995 (0.033)	1.004 (0.046)	
<i>OLS:</i>	0.999 (0.013)	1.000 (0.015)	1.000 (0.012)	$\rho = 0$
<i>MLE:</i>	0.998 (0.016)	0.998 (0.024)	0.996 (0.035)	
<i>2Step:</i>	0.998 (0.016)	0.998 (0.024)	0.996 (0.036)	
<i>SP:</i>	1.000 (0.019)	1.000 (0.031)	0.995 (0.036)	

	Mean Squared Error			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.0504 (0.0073)	0.1000 (0.0121)	0.1281 (0.013)	$\rho = 1$
<i>MLE:</i>	0.0003 (0.0004)	0.0007 (0.0012)	0.0005 (0.0007)	
<i>2Step:</i>	0.0005 (0.0006)	0.0010 (0.0013)	0.0020 (0.0025)	
<i>SP:</i>	0.0006 (0.0008)	0.0018 (0.0025)	0.0020 (0.0026)	
<i>OLS:</i>	0.0136 (0.004)	0.0243 (0.005)	0.0325 (0.0057)	$\rho = 0.5$
<i>MLE:</i>	0.0003 (0.0005)	0.0005 (0.0009)	0.0008 (0.0012)	
<i>2Step:</i>	0.0004 (0.0005)	0.0006 (0.001)	0.0021 (0.0031)	
<i>SP:</i>	0.0005 (0.0006)	0.0011 (0.0018)	0.0021 (0.0032)	
<i>OLS:</i>	0.0002 (0.0003)	0.0002 (0.0003)	0.0002 (0.0002)	$\rho = 0$
<i>MLE:</i>	0.0002 (0.0003)	0.0006 (0.0007)	0.0012 (0.0015)	
<i>2Step:</i>	0.0002 (0.0003)	0.0006 (0.0007)	0.0013 (0.0017)	
<i>SP:</i>	0.0003 (0.0005)	0.0009 (0.0014)	0.0013 (0.0017)	

Table B1: Monte Carlo simulation, with t-distribution and 33% employment

	β -estimate			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.682 (0.026)	0.496 (0.028)	0.379 (0.029)	$\rho = 1$
<i>MLE:</i>	1.004 (0.03)	1.020 (0.046)	1.025 (0.091)	
<i>2Step:</i>	1.015 (0.029)	1.041 (0.048)	1.184 (0.125)	
<i>SP:</i>	1.001 (0.033)	1.007 (0.058)	1.180 (0.124)	
<i>OLS:</i>	0.840 (0.025)	0.744 (0.022)	0.684 (0.024)	$\rho = 0.5$
<i>MLE:</i>	1.000 (0.03)	1.009 (0.039)	1.033 (0.087)	
<i>2Step:</i>	1.007 (0.028)	1.010 (0.038)	1.096 (0.102)	
<i>SP:</i>	0.998 (0.029)	0.998 (0.046)	1.094 (0.103)	
<i>OLS:</i>	0.999 (0.02)	0.999 (0.021)	1.001 (0.023)	$\rho = 0$
<i>MLE:</i>	0.999 (0.023)	0.996 (0.034)	1.006 (0.155)	
<i>2Step:</i>	0.999 (0.023)	0.997 (0.033)	0.998 (0.099)	
<i>SP:</i>	1.000 (0.027)	0.997 (0.038)	0.998 (0.099)	

	Mean Squared Error			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.1015 (0.0164)	0.2549 (0.0279)	0.1241 (0.3867)	$\rho = 1$
<i>MLE:</i>	0.0009 (0.0012)	0.0025 (0.0035)	0.0362 (0.0089)	
<i>2Step:</i>	0.0011 (0.0014)	0.0040 (0.0056)	0.0121 (0.0493)	
<i>SP:</i>	0.0011 (0.0015)	0.0033 (0.0053)	0.0533 (0.0477)	
<i>OLS:</i>	0.0263 (0.0079)	0.0661 (0.0112)	0.1030 (0.1005)	$\rho = 0.5$
<i>MLE:</i>	0.0009 (0.0014)	0.0016 (0.0028)	0.0153 (0.0086)	
<i>2Step:</i>	0.0008 (0.0013)	0.0015 (0.0022)	0.0121 (0.0194)	
<i>SP:</i>	0.0008 (0.0011)	0.0021 (0.0025)	0.0246 (0.0194)	
<i>OLS:</i>	0.0004 (0.0005)	0.0004 (0.0005)	0.0994 (0.0005)	$\rho = 0$
<i>MLE:</i>	0.0005 (0.0007)	0.0011 (0.0017)	0.0008 (0.0238)	
<i>2Step:</i>	0.0005 (0.0007)	0.0011 (0.0015)	0.0441 (0.0097)	
<i>SP:</i>	0.0007 (0.0009)	0.0014 (0.0018)	0.0111 (0.0098)	

Table B2: Monte Carlo simulation, with t-distribution and 66% employment

	β -estimate			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.791 (0.019)	0.709 (0.02)	0.664 (0.019)	$\rho = 1$
<i>MLE:</i>	0.998 (0.021)	1.000 (0.029)	1.009 (0.032)	
<i>2Step:</i>	1.005 (0.021)	1.018 (0.028)	1.061 (0.052)	
<i>SP:</i>	0.995 (0.023)	0.998 (0.039)	1.060 (0.052)	
<i>OLS:</i>	0.897 (0.015)	0.857 (0.013)	0.831 (0.016)	$\rho = 0.5$
<i>MLE:</i>	0.999 (0.019)	1.000 (0.022)	0.997 (0.038)	
<i>2Step:</i>	1.005 (0.02)	1.010 (0.021)	1.027 (0.04)	
<i>SP:</i>	1.000 (0.022)	1.001 (0.03)	1.026 (0.04)	
<i>OLS:</i>	1.001 (0.012)	0.999 (0.014)	0.998 (0.015)	$\rho = 0$
<i>MLE:</i>	1.001 (0.015)	0.997 (0.023)	0.991 (0.043)	
<i>2Step:</i>	1.001 (0.015)	0.997 (0.023)	0.994 (0.037)	
<i>SP:</i>	1.001 (0.018)	0.998 (0.031)	0.994 (0.037)	

	Mean Squared Error			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.0439 (0.0077)	0.0850 (0.0117)	0.1130 (0.0129)	$\rho = 1$
<i>MLE:</i>	0.0004 (0.0007)	0.0008 (0.0011)	0.0011 (0.0016)	
<i>2Step:</i>	0.0005 (0.0008)	0.0011 (0.0015)	0.0064 (0.008)	
<i>SP:</i>	0.0006 (0.0008)	0.0015 (0.0021)	0.0063 (0.0078)	
<i>OLS:</i>	0.0107 (0.0032)	0.0207 (0.0037)	0.0287 (0.0055)	$\rho = 0.5$
<i>MLE:</i>	0.0004 (0.0005)	0.0005 (0.0008)	0.0015 (0.0021)	
<i>2Step:</i>	0.0004 (0.0006)	0.0005 (0.0007)	0.0023 (0.0028)	
<i>SP:</i>	0.0005 (0.0007)	0.0009 (0.0012)	0.0023 (0.0027)	
<i>OLS:</i>	0.0001 (0.0002)	0.0002 (0.0003)	0.0002 (0.0003)	$\rho = 0$
<i>MLE:</i>	0.0002 (0.0003)	0.0005 (0.0007)	0.0019 (0.0037)	
<i>2Step:</i>	0.0002 (0.0003)	0.0005 (0.0007)	0.0014 (0.0019)	
<i>SP:</i>	0.0003 (0.0005)	0.0009 (0.0014)	0.0014 (0.0018)	

Table C1: Monte Carlo simulation, with X^2 distribution and 33% employment

	β -estimate			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
OLS:	0.583 (0.06)	0.367 (0.06)	0.270 (0.061)	$\rho = 1$
MLE:	1.206 (0.107)	2.512 (0.164)	3.328 (0.124)	
2Step:	1.018 (0.073)	0.978 (0.11)	1.058 (0.292)	
SP:	1.003 (0.08)	0.964 (0.121)	1.055 (0.288)	
OLS:	0.773 (0.054)	0.683 (0.06)	0.631 (0.063)	$\rho = 0.5$
MLE:	1.031 (0.075)	2.289 (0.9)	3.258 (1.09)	
2Step:	0.989 (0.062)	0.985 (0.11)	0.994 (0.282)	
SP:	0.982 (0.067)	0.980 (0.12)	0.994 (0.283)	
OLS:	1.004 (0.057)	1.010 (0.061)	0.977 (0.059)	$\rho = 0$
MLE:	0.998 (0.066)	1.054 (0.236)	2.332 (1.621)	
2Step:	0.997 (0.066)	1.026 (0.107)	1.001 (0.349)	
SP:	0.997 (0.072)	1.026 (0.113)	0.999 (0.345)	

	Mean Squared Error			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
OLS:	0.1773 (0.0509)	0.4047 (0.075)	0.5373 (0.0888)	$\rho = 1$
MLE:	0.0540 (0.0473)	2.3118 (0.48)	5.4354 (0.5776)	
2Step:	0.0056 (0.0074)	0.0125 (0.0149)	0.0876 (0.1352)	
SP:	0.0063 (0.0075)	0.0158 (0.0173)	0.0852 (0.1316)	
OLS:	0.0543 (0.0244)	0.1043 (0.0388)	0.1403 (0.0457)	$\rho = 0.5$
MLE:	0.0066 (0.0082)	2.4642 (1.7515)	6.2775 (2.5357)	
2Step:	0.0039 (0.0046)	0.0121 (0.0172)	0.0789 (0.1238)	
SP:	0.0048 (0.006)	0.0147 (0.0228)	0.0793 (0.1226)	
OLS:	0.0032 (0.0044)	0.0037 (0.0052)	0.0039 (0.0058)	$\rho = 0$
MLE:	0.0043 (0.0051)	0.0581 (0.4475)	4.3762 (4.8587)	
2Step:	0.0043 (0.0051)	0.0121 (0.0147)	0.1208 (0.177)	
SP:	0.0052 (0.006)	0.0134 (0.0187)	0.1182 (0.1703)	

Table C2: Monte Carlo simulation, with X^2 distribution and 66% employment

	β -estimate			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.750 (0.053)	0.635 (0.048)	0.583 (0.05)	$\rho = 1$
<i>MLE:</i>	1.138 (0.082)	1.861 (0.109)	2.306 (0.081)	
<i>2Step:</i>	1.011 (0.061)	1.013 (0.072)	1.052 (0.093)	
<i>SP:</i>	1.003 (0.063)	0.998 (0.098)	1.051 (0.092)	
<i>OLS:</i>	0.873 (0.041)	0.819 (0.04)	0.790 (0.041)	$\rho = 0.5$
<i>MLE:</i>	1.037 (0.055)	1.664 (0.456)	2.474 (0.34)	
<i>2Step:</i>	1.005 (0.045)	1.012 (0.062)	1.021 (0.097)	
<i>SP:</i>	1.001 (0.055)	0.991 (0.08)	1.019 (0.097)	
<i>OLS:</i>	1.004 (0.044)	0.998 (0.045)	1.003 (0.051)	$\rho = 0$
<i>MLE:</i>	1.004 (0.051)	1.003 (0.07)	1.215 (0.612)	
<i>2Step:</i>	1.004 (0.051)	1.000 (0.07)	0.992 (0.099)	
<i>SP:</i>	1.000 (0.058)	1.008 (0.091)	0.993 (0.102)	

	Mean Squared Error			
	$\theta = 1$	$\theta = 0.5$	$\theta = 0$	
<i>OLS:</i>	0.0652 (0.028)	0.1352 (0.0353)	0.1766 (0.0416)	$\rho = 1$
<i>MLE:</i>	0.0258 (0.0231)	0.7527 (0.1858)	1.7124 (0.211)	
<i>2Step:</i>	0.0038 (0.005)	0.0053 (0.0071)	0.0113 (0.014)	
<i>SP:</i>	0.0039 (0.0051)	0.0095 (0.0152)	0.0109 (0.0134)	
<i>OLS:</i>	0.0177 (0.0107)	0.0341 (0.0149)	0.0457 (0.0173)	$\rho = 0.5$
<i>MLE:</i>	0.0043 (0.0053)	0.6467 (0.5147)	2.2864 (0.4968)	
<i>2Step:</i>	0.0021 (0.0027)	0.0040 (0.0067)	0.0097 (0.0133)	
<i>SP:</i>	0.0030 (0.0038)	0.0065 (0.0112)	0.0098 (0.0133)	
<i>OLS:</i>	0.0019 (0.0028)	0.0020 (0.0023)	0.0026 (0.0036)	$\rho = 0$
<i>MLE:</i>	0.0026 (0.0037)	0.0049 (0.0071)	0.4169 (1.073)	
<i>2Step:</i>	0.0025 (0.0037)	0.0048 (0.0072)	0.0099 (0.0143)	
<i>SP:</i>	0.0033 (0.0056)	0.0083 (0.0118)	0.0103 (0.0145)	

c) Stata Code

The following is a brief excerpt from one of the stata .do files that were used in Monte Carlo simulation for section 4. The code can be broken up as follows:

- line 4 – 20: generate model
- line 22-24: calculating OLS estimate
- line 25-27: calculating ML estimate
- line 28-30: calculating 2-step estimate
- line 31-32: deriving initial beta's for iterative part of semi-parametric process
- line 33-36: acquiring α using semi-parametric variant of probit function
- line 37-60: iterative part of semi-parametric process

```
1 clear;
2 set mem 500M;
3
4 global n=10000;
5 global rho=1;
6 global weak_beta=0;
7 global erate = 0.3652;
8
9 capture program drop m_error;
10 program define MonteCarlo, rclass;
11   drop _all;
12   set obs $n;
13   gen x=invnorm(uniform());
14   gen z=invnorm(uniform());
15   gen e=(invttail(5,uniform())/1.293);
16   gen u=$rho*e+(invttail(5,uniform())/1.293);
17   gen d=x+$weak_beta*z+e;
18   gen empl=0;
19   replace empl=1 if d>invnorm(1-
20   $erate)*(3^(1/2));
21
22   reg y x if empl==1;
23   return scalar bhat1=_b[x];
24   return scalar bvar1=( _b[x]-1)^2;
25   heckman y x, select(empl= x z);
26   return scalar bhat2=_b[x];
27   return scalar bvar2=( _b[x]-1)^2;
28   heckman y x, twostep select(empl= x z);
29   return scalar bhat3=_b[x];
30   return scalar bvar3=( _b[x]-1)^2;
31   gen twostep1=_b[_cons];
32   gen twostepb=_b[x];
33   gen empl2=empl+1;
34   sneop empl2 x z, order(3);
35   predict ax;
36   replace ax=ax-1;
```

```

37   gen xb=twostep1+twostepb*x;
38
39   regr y ax;
40   gen k1 = _b[_cons]+_b[ax]*ax;
41   regr xb ax;
42   gen k2 = _b[_cons]+_b[ax]*ax;
43   gen g = k1 - k2;
44   gen y1 = y - g;
45   regr y1 x;
46   replace xb = _b[_cons]+_b[x]*x;
47
48   while i<20{
49     replace i=i+1
50     regr y ax;
51     replace k1 = _b[_cons]+_b[ax]*ax;
52     regr xb ax;
53     replace k2 = _b[_cons]+_b[ax]*ax;
54     replace g = k1 - k2;
55     replace y1 = y - g;
56     regr y1 x;
57     replace xb = _b[_cons]+_b[x]*x;
58   }
59   return scalar bhat4=_b[x];
60   return scalar bvar4=(_b[x]-1)^2;
61
62 end;

```