Earnings bracket obstacles in household surveys – How sharp are the tools in the shed?

DIETER VON FINTEL

Stellenbosch Economic Working Papers: 08/06

KEYWORDS: LABOUR, HOUSEHOLD SURVEYS, EARNINGS
JEL: C15, C24, C42, J01, C81

DIETER VON FINTEL
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: DIETER@VONFINTEL.COM

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

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

Earnings bracket obstacles in household surveys – How sharp are the tools in the shed?

DIETER VON FINTEL

ABSTRACT

Earnings functions form the basis of numerous labour market analyses. Non-response (particularly among higher earners) may, however, lead to the exclusion of a significant proportion of South Africa's earnings base. Earnings brackets have been built into surveys to maintain sufficient response rates, but also to capture information from those who are unsure about the earnings of fellow household members. This data type gives a rough indication of where the respondent lies in the income distribution, however exact figures are not available for estimation purposes. To overcome the mixed categorical and point nature of the dependent variable, researchers have traditionally applied midpoints to bracket earnings. Is this method too rudimentary? It is important to establish whether the brackets are too broad in South African Household surveys to be able to make reliable inferences. Here, midpoints are imputed to interval-coded responses alongside theoretical conditional means from the Pareto and lognormal distributions. The interval regression is implemented as a basis case, as it soundly incorporates point and bracket data in its likelihood function. Monte-Carlo simulation evidence suggests that interval regressions are least sensitive to bracket size, however midpoint imputation suffers distortions once brackets are too broad. Coefficient differences are investigated to distinguish similar from different results given the chosen remedy, and to establish whether midpoint imputations are credibly similar to applying interval regressions. To this end, testing procedures require adjustment, with due consideration of the heteroskedasticity introduced by Heckman 2-step estimates. Bootstrapping enhances conclusions, which shows that coefficients are virtually invariant to the proposed methods. Given that the bracket structure of South African Household Surveys has remained largely unchanged, midpoints can be applied without introducing coefficient bias.

JEL codes: C15, C24, C42, J01, C81

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 Introduction

The analysis of earnings data in South Africa introduces economists to many insightful conclusions about the structure of the labour market. In the process, however, researchers encounter a number of methodological obstacles, which are data-driven and not directly related to sound economic interpretation. While many respondents in surveys supply useful figures, many others (particularly in higher income groups) are hesitant to divulge their financial positions. Misreporting and underreporting abound. Survey designers have offered the first step in the solution, by building "income bracket options" into questionnaires. Respondents who are unwilling to provide exact amounts are allowed a certain degree of anonymity by being afforded the option to indicate which income band they fall into. This leaves the completion of the task to econometricians, who have to find techniques to maximise information from a mixture of categorical and nominal data. Adler et al (1998: ix) label it "bad data", simply because researchers are not always certain how to analyse such unfamiliar information, particularly in its role as a dependent variable. While it is not the econometrician's task to improve the purity of datasets, it is imperative that sound methods are confirmed and implemented correctly to maximise the "true" information which is extractable.

This study engages both intuitive (applying midpoints) and theory based data simulation (conditional mean imputation from various distributions) for the dependent variable in earnings functions. These variants are compared to interval regressions to introduce a sense of surety that results which are obtained are indeed similar, regardless of the complexity of the entire process - from household responses to economic conclusions. The latter technique's mechanics are designed to handle coarsened data.

A further target for improved accuracy is the reliability of standard errors reported within the Heckman framework. While addressing non-random sample selection (to account for omitted variable bias in earnings equations), impure variance-covariance structures are introduced, which demand attention. Are asymptotic corrections valid, or is it necessary to implement the bootstrap? If resulting confidence intervals are too broad, too many values are regarded to be admissible; overprecision may similarly lead to *non*-rejection of invalid hypotheses (Brownstone & Valetta, 2001: 129). This is of particular relevance to testing procedures when the proposed methods are compared.

The object of this study is therefore to uncover some of the pitfalls encountered in labour market analyses. How do traditional methods compare to new innovations? How does one embark on a process of sifting good information from bad information with complex data? Conclusions are inherently linked to the quality of the chosen data (for this study, the Labour Force Survey, September 2003), but the manner in which they are reached should not be subject to these limitations. How different are the tools researchers have at their disposal?

The rest of this paper is structured as follows: Section 2 motivates the need for interval-coding innovations in questionnaires, while Section 3 outlines accompanying econometric problems related to sample selection bias. Section 4 addresses various methods to overcome the limitations of the dependent variable. Section 5 applies the earnings function literature to the chosen specification, while Section 6 reports the findings of this study. Section 7 concludes.

2 Why Bracket Earnings?

Keswell and Poswell (2004: 855) highlight the need to utilise more than just reported point income data to avoid biased estimation. In many cases the data generating process (DGP) of this group of respondents appears to be different to a simulated lognormal theoretical benchmark. In South Africa, the proportion of reporting decreases as survey years progress, hence the need arises to capture additional information via bracket earnings questions. Respondents reluctant to provide exact income details are presented with the alternative to respond within a category (see Table 1). The way this information is processed is therefore of utmost importance: econometricians no longer have a continuous variable at their disposal, and can therefore not apply well-grounded techniques such as OLS. Classical remedial measures include the imputation of midpoints to interval-coded observations, however Keswell and Poswell (2004: 855) show that this too generally leads to significant differences, when the resultant distribution is compared to the shape of theoretical distributions.

This problem is prominent in later surveys, where underreporting is more prevalent. Midpoint imputations introduce substantial differences from the implied DGP, bar for the 1997 October Household Survey (OHS97) (Keswell & Poswell, 2004: 855). The sole use of point data or the validity of imputations in later surveys should always be questioned. Any earnings study should commence with the comparison of DGP's to theory, in order to establish whether the proposed variant of the dependent variable satisfies the above considerations.

2.1 Which methods have previous studies implemented?

How then does one overcome biases, and are there cases where midpoints are in fact useful indicators of categorical earnings?

Rospabé (2002) and Daniels & Rospabé (2005) capitalise on the innovative "interval regression" to overcome the need to choose between simulation and midpoint imputation. This procedure rests on maximum likelihood estimation within a generalised Tobit model. It therefore bridges the gap from point data to the maximum information provided by respondents, by incorporating interval-coded information into the likelihood function.

Hofmeyr (1999: 8) implements the midpoint method, without imputing a value to the open category. This choice is justified by the fact that brackets are not wide. The 1999 October Household Survey (OHS99) was studied, which Keswell and Poswell (2004: 855) claim to be plagued by biased results following the application of midpoints.

Work based on earlier surveys, such as the 1993 Project for Statistics on Living Standards and Development (PSLSD) (Mwabu & Schultz, 2000), OHS 1994 (Winter, 1999) and OHS 1995 (Bhorat & Leibbrandt, 2001), does not mention methods implemented to deal with categorical reporting. This may be indicative of sufficient point responses in initial surveys, but also of an uncontroversial implementation of the midpoint methodology.

2.2 The Data Divide

Table 1 LFS September 2003 - Summary of Earnings Data (Employed Respondents)

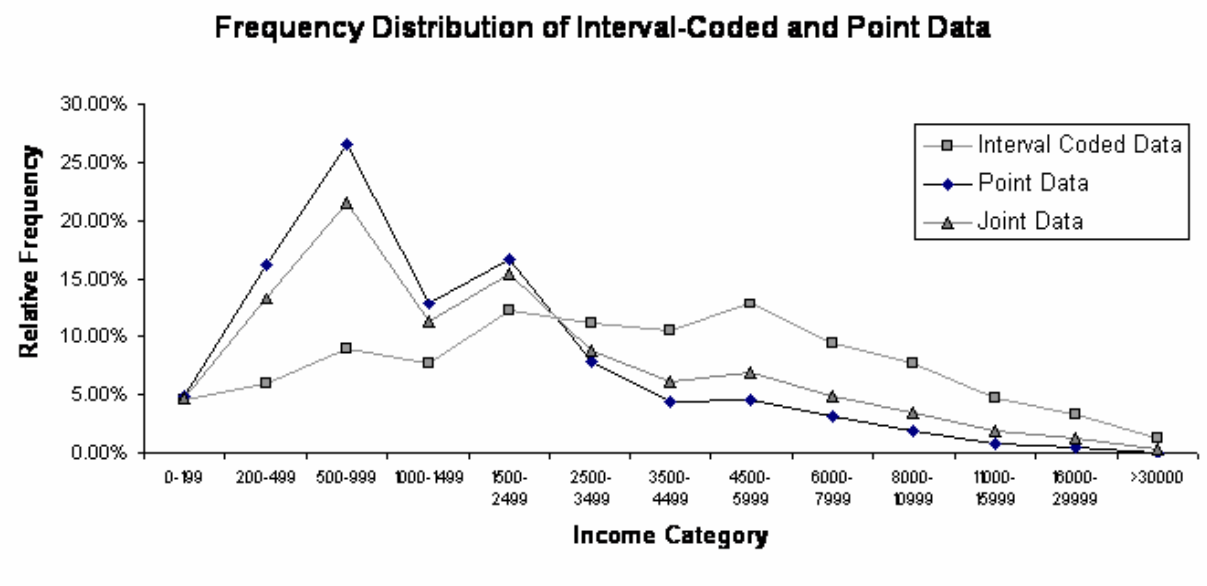
Type of Earnings Data	INTERVAL-CODED RESPONSES		POINT RESPONSES		NO EARNINGS DATA REPORTED
Earnings Range as per LFS bracket question (in Rands)	*Frequency*	*Percent*	*Frequency*	*Percent*	*Frequency*
0	726	10.58%			
0-199	280	4.08%	745	4.79%	
200-499	363	5.29%	2,533	16.27%	
500-999	549	8.00%	4,129	26.53%	
1000-1499	468	6.82%	1,986	12.76%	
1500-2499	745	10.86%	2,603	16.72%	
2500-3499	682	9.94%	1,209	7.77%	
3500-4499	646	9.42%	679	4.36%	
4500-5999	790	11.52%	715	4.59%	
6000-7999	574	8.37%	480	3.08%	
8000-10999	469	6.84%	284	1.82%	
11000-15999	288	4.20%	122	0.78%	
16000-29999	200	2.92%	79	0.51%	
>30000	80	1.17%	2	0.01%	
Other Employed respondents					1,955
Total	**6,860**	**100%**	**15,566**	**100%**	**1,955**
% of Total	**28.137%**		**63.845%**		**8.019%**

The importance of including earnings range questions in the Labour Force Survey (September 2003) is evident in Table 1. While 63.9% of the employed sample provided point data for earnings, a further 28.1% of the sample responded within an income band. This still restricts the analysis to a sub-sample of respondents; however the improved knowledge allows researchers to work with 92 % of those who were reported to be employed. Of particular importance is the number of respondents providing point income data in the lower categories, while those in higher income categories prefer the anonymity of supplying only their earnings brackets. Only two earners reported exact amounts in the open-ended category. If only point data is used, it is clear that a large proportion of South Africa's earnings base is excluded from the analysis. This has implications for distributional questions, but also for the accuracy of coefficients: in addition, the quality of sample estimates is further degraded by those who declined to offer any information or falsely reported zero incomes. It is therefore important to find appropriate techniques to maximise the use of interval-coded data.

Figure 1 illustrates how the inclusion of earnings brackets alters the relative frequency of income: the lower portion of the distribution of joint data is afforded less weight compared to point data,

while the upper portion undergoes an upward adjustment to account for underreporting. In section 4.3 the comparison with the implied lognormal DGP is highlighted.

Figure 1 Distribution of Interval-Coded, Point and Joint Data



The problem of missing data for the remaining employed in the sample (8% do not report any income information – see Table 1) can be addressed in future studies by microdata simulation (see Birkin & Clarke, 1995) and discriminant analysis (see Maddala, 1983: Chapter 4). The latter involves classifying an individual into any number of populations, given other known characteristics. In this case, the methodology would involve allocating an income band to those individuals who refused to answer, grouped according to attributes such as experience, education, unionisation, industry or occupation. This is analogous to obtaining predicted values following the estimation of a regression on the available data, and hence inferring augmented information to the sample. This method, termed multiple imputation, is described and applied in Brownstone & Valetta (2001: 136-139).

It should be noted that this simulated data must be compatible with the DGP, as mentioned above. Should non-reporters have different characteristics point and bracket respondents, this translates to a misleading practice. Multiple imputation is not implemented in this study: the focus remains to test parameter differences across the various solutions to the interval coding obstacle. The question at hand dictates the need for such processes: for example, Keswell and Poswell (2004: 836) show from various previous surveys, that those who reported income did not possess statistically different educational characteristics from those who do not provide income

details. As such, multiple imputation would not have added new information to answer educational questions. Should one find that this group does indeed have different characteristics (given the proposed hypothesis), multiple imputation would be necessary to avoid distortion.

3 Methodological Considerations

Throughout, augmented Mincerian Earnings Functions are estimated for males, females and a joint sample. A parsimonious model is chosen, in accordance with knowledge from previous work. While the expected signs and the relative size of coefficients are well-known for South African data, the object of this study is not to draw new conclusions on the determinants of earnings, but to establish which methods provide the most reliable estimates.

3.1 Sample Selection Bias

First, the obstacle of sample selection bias is brought to account. Heckman (1979: 153-154), in his seminal article, outlines that when estimates are based on non-randomly selected samples, population estimates of wage equations are misspecified. In this case, the sample is restricted to those who are employed, and as such the influence of experience and schooling, for example, are misrepresented. Selection may be forced as a result of structural unemployment, which is of particular relevance to South Africa. A broader understanding, however, is based on the notion of self selection: if wages offered (regardless of how high they are) are below reservation wages, labour force participants will *choose* unemployment. Since these individuals (who may be well-educated and experienced) are not incorporated in the earnings equation, coefficients are biased. Typically favourable characteristics in these cases do not improve earnings, and the true value of human capital is not apparent as a result of the selection.

Wooldridge (2002: 552) defines a wage equation only to be valid if it "represents *all* people of working age, whether or not the person is actually working at the time of the survey." (italics in original). As such, a preliminary employment probit equation is consistently estimated to model the selection process. This precedes each earnings equation, from which the Inverse Mills Ratio (λ) for each group member is calculated. This is a function of the probability that each observation is included in the sample (Heckman, 1979: 156):

$$\hat{\lambda}_i = \frac{\phi\left(\underline{w}_i' \hat{\underline{\gamma}}\right)}{\Phi\left(\underline{w}_i' \hat{\underline{\gamma}}\right)} \quad i = 1 \ldots N$$

where

$\underline{w}_i$: m x 1 vector of explanatory observations for individual i in the probit

$\hat{\underline{\gamma}}$: m x 1 vector of estimated parameters from the probit

$\phi(x)$: standard normal density function

$\Phi(x)$: cumulative standard normal distribution function

These ratios are included as regressors in the relevant earnings equations, and therefore correct for both under- and overstatement of each observation's influence on the coefficients. The inclusion of the sample selection term counters inconsistency for all coefficients and the omitted variable bias (albeit a generated variable) which derives from the selection process (Wooldridge, 2002: 563).

Bhorat and Leibbrandt (2001: 113) include a double hurdle selection equation: the purpose is to differentiate between the actual decision to participate and the probability of employment in South Africa. Given the extent of *broad* unemployment in South Africa, and the number of discouraged workers on the periphery of the labour market, such analyses may prove insightful. However, this exercise does not add any value if the sole purpose is to eliminate sample selection bias in earnings functions (Bhorat and Leibbrandt, 2001: 113).

The "Heckit" model's outline follows from Hill et al (2003: 2-3). It highlights the nature of the problem. First, the selection equation is defined, which models the "true" propensity to be employed:

$$z_i^* = \underline{w}_i' \underline{\gamma} + u_i \quad i = 1 \ldots N$$

where

z_i^* : latent variable (underlying propensity to be employed)

$\underline{w}_i'$: m x 1 row vector of m explanatory variables observed for the i^{th} individual

$\underline{\gamma}$: m x 1 vector of population parameters

u_i: random error

This latent variable is not directly observed and represents an underlying propensity to be employed, but the selection variable (employment dummy) is indeed observed, and is the dependent variable in the probit selection equation, where we model the probability that the individual is employed:

$$z_i = \begin{cases} 1 \text{ if offered wage } \geq \text{reservation wage} & (\text{here } z^*_i \text{ is observable, and person is employed}) \\ 0 \text{ if offered wage} < \text{reservation wage} & (\text{here } z^*_i \text{ is unobservable, and person is unemployed}) \end{cases}$$

$$i = 1 \ldots N$$

Now the "true" earnings equation follows as:

$$\log(y_i) = \underline{x}_i' \underline{\beta} + e_i \quad i = 1 \ldots N$$

where

y_i : observations on the earnings variable

$\underline{x}_i$: p x 1 vector of observations on explanatory variables

$\underline{\beta}$: p x 1 vector of parameters

e_i : random error

Now the assumption must hold that the respective error terms are independently distributed. But in general it is true that:

$$\begin{bmatrix} u_i \\ e_i \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} 1 & \rho \\ \rho & \sigma_e^2 \end{bmatrix} \right)$$

The nature of the selection obstacle now becomes clear: if $\rho=0$ and z^*_i is perfectly observable, there is no problem. That implies that shocks to earnings would not influence employment and vice versa. Generally this is not true. If the entire population's true income and employment propensity is observable, one can run OLS, without suffering selection bias.

Hence, conditioning on $z^*_i>0$ is necessary:

$$\log(y_i) = E\left[\log(y_i)\left|z^*_i > 0\right.\right] + e_i \quad i = 1\ldots N$$
$$= \underline{x}'_i \underline{\beta} + (\rho\sigma_e)\hat{\lambda}_i + e_i$$
$$= \underline{x}'_i \underline{\beta} + \beta_\lambda \hat{\lambda}_i + e_i \quad i = 1\ldots N$$

with

$$E\left[e_i \left| z^*_i > 0\right.\right] = 0 \quad i = 1\ldots N$$
$$Var\left[e_i \left| z^*_i > 0\right.\right] = \sigma_e{}^2\left[1 - \rho^2\hat{\lambda}_i\left(\hat{\lambda}_i + \underline{w}_i{}'\underline{\gamma}\right)\right] \quad i = 1\ldots N$$

where $\hat{\lambda}_i$ is the Inverse Mills Ratio, as above.

Wooldridge (2002: 564) indicates that the test for significant selection bias is a conventional T-test of H_0: $\beta_\lambda=0$ within the earnings specification. This derives from the hypothesis that $\rho=0$ for the independence of employment and earnings processes to hold. Under this hypothesis, the standard assumption of homoskedasticity holds, while the introduction of significant sample selection bias causes it to be violated, as is evident in the subscripted, observation-specific variance shown above. This leads to the next point of concern.

3.2 Correct Standard Errors and Confidence Intervals

Given the importance of testing coefficients' comparative precision, correct confidence intervals – unaffected by impure standard errors – are a necessity. Initially, the Heckman covariance matrix (based on the above scenario) is implemented. The first correction applies robust standard errors, according to the Huber-White covariance matrix (Hill et al, 2003: 5):

$$\Sigma_{Huber-White} = (X'X)^{-1} X'DX (X'X)^{-1}$$

where X : n x p is the data matrix
and D: n x n is a diagonal matrix with squared OLS residuals on the diagonal

These are reported when weights correcting for sample design are accounted for in the estimation procedure. Statistics South Africa (StatsSA) (2003a: 2-3) outlines the calculation of these weights as the inverse of the probability that the household and the sampling unit are included in the survey. This adaptation should not be confused with sample selection.

Wooldridge (2002: 564) sounds the warning that robust standard errors may nonetheless be misleading, as β_λ is itself the coefficient of an estimated stochastic quantity. The same problem is encountered when generalised least squares is sought as a solution (Hill et al, 2003: 3).

Hill et al (2003: 4-12) evaluate the adequacy of implementing various proposed asymptotic variance-covariance matrices with Monte-Carlo simulation studies. The cases investigated are: the usual OLS covariance matrix; a heteroskedasticity corrected variance-covariance matrix (which does not take into account the randomness of λ); the White heteroskedasticity consistent estimator, an adjusted form of the latter; the formulation originally proposed by Heckman (1979: 159), later modified by Green; a Murphy-Topel estimator, along with a modification by Hardin for 2-stage estimation. The final approach uses bootstrap estimation: repeated samples are drawn from the data matrix, with replacement, to approach true finite sample measures. For small samples and considerable censoring, none of the variants above perform satisfactorily (Hill et al, 2003: 18). For large samples, estimator variability was reflected well by the bootstrap procedure. Since LFS survey data constitute large samples, this procedure is deemed most appropriate to compare the parameter estimates of the models proposed below.

Henderson (2005: 3) provides a brief overview of the process and benefits of using bootstrap estimation. The basis is repetitive sampling with replacement: an unknown population distribution can be inferred, by deriving properties from the many samples. As such, the parent population is approximated as the number of repetitions is increased. Sprent, in Henderson (2005: 3) claims: "The more vague the supposition about population distributions, the more useful the bootstrap becomes." This underlines the attractiveness of implementing this method, as random repetitive sampling approaches the truth when there is no agreed distributional form to follow as a beacon: accuracy no longer rests on a formula, assumptions may be relaxed, and essentially the data reveals more about itself, without mathematical imposition. Brown (2000: 437) advocates the use of bootstrap for cases where asymptotic variance is impossible or difficult to calculate. In addition, the perception exists that these standard errors are more accurate in finite sample situations. Distributions of parameters are deemed to be closer to the true population approximation than limiting distributions.

Brownstone & Valetta (2001: 131) uncover the mechanics used in bootstrap regression estimation. If the sample size is *n*, then for each repetition implemented via the bootstrap, *n* rows

of the data matrix (along with the associated element of the dependent variable) are sampled *with* replacement. This implies that some rows probably appear more than once in a single repetition's "new" data matrix, while others are excluded. The observations used therefore differ for each of *r* repetitions, and each time a different set of coefficients is estimated. The term "paired bootstrap" is applied. The resulting observed distribution of the *r* sets of coefficients eventually approximates the population distribution. This will deliver consistent results, regardless of the nature of the underlying standard errors in a usual regression. Brownstone & Valetta (2001: 132-133) implement sequentially different methods in their own earnings study: ordinary OLS confidence intervals are narrower than robust confidence intervals, while bootstrapping resulted in the broadest intervals. This is indicative of the difficulties involved with even asymptotic corrections.

Since the purpose of this study is to obtain good confidence intervals, unaffected by impure standard errors or false distributional assumptions, a discussion follows to describe bootstrapped intervals, which follows Henderson (2005: 6-8). Percentile intervals involve first arranging the *r* sets of coefficients in descending order and assigning ranks to each. A 95% confidence interval is constructed by assigning the 2.5% and 97.5% quantiles of this generated distribution to the bounds. As *r* increases, a confidence interval attains accuracy, as continuous estimates are added to the "distribution", and bounds are clearly established. Deficiencies in this method have necessitated the implementation of bias-corrected intervals, both as a result of inaccuracy and asymmetries in the distributions. This method is implemented in STATA via jackknife procedures. Second order accuracy is achieved, as errors decay at a rate of $1/r$. Therefore a large number of replications will lead to satisfactory interval estimates.

The question which remains is how many repetitions are necessary to reach the "truth": since this technique is computer intensive, it can readily be executed many times. Henderson (2005: 5-6) maintains that 200 replications are necessary to approximate standard errors, however in excess of a thousand are necessary for confidence intervals. Given that no distributional assumptions are made, the pivotal statistics and standard errors are not accompanied by tabled percentiles to complete the process of calculating a confidence interval. It is therefore necessary to increase the iterations in bootstrapping to obtain improved distributional knowledge. Brownstone & Valetta (2001: 132-133) implement 1000 repetitions for confidence intervals. Improved computational speed allows the use of 10000 replications in this study.

Hill et al (2003: 9) emphasise the responsibility of the researcher to report not only models implemented, but the software (and version thereof) used to reach results. Each programme may use a specific correction of the variance-covariance structure to achieve "robust" standard errors: not only have programming errors occurred in the past, but the fundamental validity of various forms are drawn into question, as above. As such, it is noted that the software used in this study is STATA/SE 9.0. (Statacorp, 2005). Estimation follows a standard built-in Heckit 2-step procedure, followed by a manual weighted Heckit 2-Step implementing robust standard errors, but also a Heckit 2-step with bootstrap estimation to approximate true standard errors (as in Hill et al, 2003: 26). All methods are implemented for comparative purposes.

4 Dependent Variable Variants

4.1 Generalised Tobit - Interval regression (basis):

As a basis case, an interval regression is implemented. This is a generalised Tobit model and is estimated via pseudo-maximum likelihood procedures when weighting is brought into account. Therefore an understanding of Tobit estimation is first reviewed (following Wooldridge, 2002: 517-525):

Suppose *y* is observed, which represents the underlying variable *y**. A truncation point exists, so that *y* is not observable past or before a particular value of *y**. We consider the model when an upper truncation point arises (which represents the case when of open top category).

$$\log\left(y_i^*\right) = \underline{x}_i'\underline{\beta} + u_i \qquad i = 1\ldots N$$

$$u_i \left| x_i \sim N\left(0;\sigma^2\right) \right. \quad i = 1\ldots N$$

$$y_i = \begin{cases} y_i^* & \text{if the respondent supplies point data} \\ c & \text{if the respondent supplies an earnings bracket} \end{cases} \qquad i = 1\ldots N$$

where c is the lowerbound of the upper category

*y** is therefore restricted to the values observed over the range of *y* even if one is aware that the *potential* value is possibly different. In this case *c* is termed a "corner point". In the interval regression generalisation, interval-coded datapoints have both a lower and an upper "corner point".

It can be shown that

$$E\left[\log\left(y_i\right)\middle|\underline{x}_i, y_i < c\right] = \underline{x}_i'\underline{\beta} + \sigma\hat{\lambda}_i\left(\underline{x}_i'\underline{\beta}/\sigma\right) \quad i = 1\ldots N$$

where

$$\hat{\lambda}_i\left(\underline{x}_i'\underline{\beta}/\sigma\right) = \frac{\phi\left(\underline{x}_i'\underline{\beta}/\sigma\right)}{\Phi\left(\underline{x}_i'\underline{\beta}/\sigma\right)} \quad i = 1\ldots N$$

$\underline{\beta}$: p x 1 is the true population parameter vector

σ is the standard error of regression

$\hat{\lambda}_i$ is the Inverse Mills Ratio for each observation

This shows that the expected value of the true variable in its observed range is larger than the OLS estimates ($\underline{x}'\underline{\beta}$) using the data points which are indeed available. For this reason it is postulated that simply applying OLS to available data points is not a satisfactory method, which introduces inconsistency: the omitted $\hat{\lambda}_i$ is clearly correlated with the other regressors of the known range. The Tobit and consequently the generalised Tobit models therefore provide better estimates to base findings on.

Daniels and Rospabé (2005: 2) provide a log-likelihood function adjusted to make provision for point, left-censored (unused in this setting, since the first earnings group contains only zero values, which are counted as missing when logged), right-censored (top income category with only a lower bound) and interval data:

$Y \equiv Earnings$

The likelihood function accommodates $X = \log(Y)$:

$$\log L = -\frac{1}{2}\sum_{i\in C} w_i\left[\left(\frac{\log\left(y_i\right) - \underline{\beta}'\underline{x}_i}{\sigma}\right)^2 + \log 2\pi\sigma^2\right] + \sum_{i\in L} w_i \log\Phi\left(\frac{\log\left(y_{Li}\right) - \underline{\beta}'\underline{x}_i}{\sigma}\right)$$

$$+\sum_{i\in R} w_i \log\left[1 - \Phi\left(\frac{\log\left(y_{Ri}\right) - \underline{\beta}'\underline{x}_i}{\sigma}\right)\right] + \sum_{i\in I} w_i \log\left[\Phi\left(\frac{\log\left(y_{2i}\right) - \underline{\beta}'\underline{x}_i}{\sigma}\right) - \Phi\left(\frac{\log\left(y_{1i}\right) - \underline{\beta}'\underline{x}_i}{\sigma}\right)\right]$$

$i \in C$ = point data
$i \in L$ = left-censored data
$i \in R$ = right-censored data
$i \in I$ = interval-censored data

and
w_i are the sampling weights

$\log(y_i) = \log(y_{Li}) \quad if \quad y_i^* \leq y_{Li}$ $\quad y_{Li}$ is the upperbound of the first category

$\log(y_i) = \log(y_{Ri}) \quad if \quad y_i^* \geq y_{Ri}$ $\quad y_{Ri}$ is the lowerbound of the top category

$\log(y_i) = \log(y_i^*) \quad if \quad y_{1i} \leq y_i^* \leq y_{2i}$ $\quad y_{1i}$ is the lowerbound of the ith category

$\quad y_{2i}$ is the upperbound of the ith category

A similar log-likelihood is applied in Wik et al (2004: 2447).

This method does not require the assignment any value within the ranges of interval-coded data, and can therefore be regarded as a reliable starting point, which removes the process of educated guessing. It does, however, rest on the assumption of the normality of logged earnings, and consequently a lognormal distribution for the untransformed earnings variable. Supplementary interval information is soundly incorporated into the procedure, and broadens the base of research from only point observations. Further work will be judged in the light of this specification.

It is evident in the framework above, that the inverse Mills ratio inherent to the Tobit family, already accounts for a bias. Daniels & Rospabé[1] (2005) maintain that this correction accounts for sample selection bias. The interval regressions in this study are nonetheless specified with Inverse Mills Ratios, which prove to be significant and point to the fact that the bias is not completely overcome.

4.2 Alternatives – Imputation

Whiteford & McGrath (1994: 28-29) list, among others, two methods to approximate the income distribution: the Midpoint method and the Midpoint-Pareto method.

[1] While this is not explicitly referred to in their paper, this assertion was confirmed upon communication with the authors

4.2.1 Midpoints

This method is conceptually simple and widely implemented by researchers. It is assumed that each person who supplies his/her income bracket earns the category mean - its midpoint. Since no upper bound exists for the top category, it is assumed that the mean exceeds the lower bound by 10%. The pitfall of this method seems to be its lack of theoretical backing (Whiteford & McGrath, 1994: 28), but at the same time it may be attractive due to the limited knowledge of statistics required. If this method is approximately close to that of an interval regression in all cases, it confirms much of past research. Keswell & Poswell (2004: 854) point to a practical problem (ignoring any statistical properties which may be violated in the process): as survey years progress, income brackets will invariably change with inflation. In effect, the midpoints vary over time, and coefficients are not comparable.

Survey design and the size of brackets introduce sensitivity in estimation. In particular, the broad lowest category in the 1995 October Household Survey afforded too much weight to the upper portions of that bracket (Keswell & Poswell, 2004: 855); other surveys broke down the band into smaller intervals, and midpoint estimates fared better in comparison. Seiver (1979: 230, 232) maintains that the true mean of any interval will always be below its midpoint, and that income distribution results are influenced by the number of intervals chosen to span the range – fewer, wider brackets distort the picture. This methodology is included, but the sensitivity of results can only be tested for this particular interval structure: the benefits are pronounced, given that the specific ranges and interval sizes of South African household surveys have been maintained since OHS 1996.

4.2.2 Midpoint-Pareto Method

Given that lower income categories are narrow, the distribution of income at the bottom end is not markedly influenced by midpoint imputation (Whiteford & McGrath, 1994: 29). However, a parametric approach is necessary for higher income categories, as greater skewness *within* groups becomes evident. Crato (2000:1239) emphasises the need to "model situations in which extreme values are observed with a relatively high probability" with the use of heavy-tailed distributions such as the Pareto. As such, a "Pareto Mean" is estimated for the open upper category (but also for selected bounded categories in the upper tail) and is assigned to each interval-coded datapoint. This value will deviate from the midpoint, according to the heaviness of the tail.

Pareto was the first to concern himself with the heavy tail evident in empirical income distributions, but concluded that the distribution which resulted from his work only clarified the distribution of the right tail with any precision. Additionally, the fact that no closed form exists for evaluation purposes, was a deterrent before computer intensive implementation became possible (Dagsvik & Vatne, 1999: 4-5). Mandelbrot (in Dagsvik & Vatne, 1999: 6) broadened the scope of this work by investigating the so-called class of "stable distributions" of which both the Normal and Pareto distributions are members. These have the property that a linear combination of several stable distributions remains a stable distribution. Mandelbrot applied this to income distributions relating to various sources: for instance the distinction between wage and capital income. He found that these sub-grouped distributions had approximately the same shape, and that by use of stable distribution properties their sum would maintain these characteristics.

For the purposes of this study, the methods employed by Whiteford & McGrath (1994: 81-84) and Gustavsson (2004: 20) are utilised. The probability density function of the Pareto distribution is given as:

$$f_Y(y) = \begin{cases} \alpha k^{\alpha} y^{-(\alpha+1)} & \text{for } y \geq k \geq 0 \text{ and } \alpha > 0 \\ 0 & \text{otherwise} \end{cases}$$

with α a shape parameter, which needs to be estimated. This can also be expressed in the log-linear form:

$$\log P = k - \alpha \log Y$$

Where Y represents any given level of income and P is the proportion of the sample earning that amount or more.

The latter equation underlines the intuition which Pareto used in the derivation of the distribution: Pareto's "Law of Distribution" postulates, on the ground of empirical observation, that a log-linear relationship exists between an income level and the number of people who earn greater or equal that amount (Whiteford & McGrath , 1994: 81).

The above equation can be implemented by OLS on the point data in the sample to obtain an estimate of α within each cohort. The next task is to establish the range over which the data does indeed match the Pareto distribution. Parker and Fenwick (1983: 874) assert that this relationship is only linear in the upper tail. Gustavsson (2004:20) proposes various proportions of the upper tail for which the data are maintained to fit the equation well; Whiteford & McGrath (1994: 29) suggest using usual midpoints below the category which contains the median income, and Pareto

means for all income brackets above that. Following the procedure set out in their appendix (Whiteford & McGrath, 1994: 81-82), the equation is estimated with all categories. Successively the lowest income band is excluded from the estimation. That equation which exhibits the best fit in terms of the R^2 of the regression is deemed to contain the most reliable estimate of α, but also serves as an indicator of the portion of the tail for which the distribution holds. The lowest category included in this "best" equation is therefore deemed to be a suitable starting point to impute Pareto Means.

Estimates conducted in this study for LFS September 2003 suggest imputing Pareto means for the following categories (with midpoints below these). For females the regression method implies including all categories with earners above R6000 per month. This is the range over which brackets increased from an interval covering R1500 to a wider range of R2000: this result therefore confirms the broad interval problem as outlined above. While the male estimation procedure suggests the inclusion of categories above R1500 (where intervals grow from a range of R500 to R1000), the fit is only marginally better compared to the scenario established for females. To remain consistent with the "broadening interval" criteria, as well as to maintain uniformity between groups, it was decided to choose the cut-off for mean estimation and imputation of R6000 and above.

Crato (2000: 1251-1252) concludes that the regression estimator of α has a smaller bias than the proposed Hill-Hall estimator. A modified version of the latter, however, has a smaller variance than the regression estimator: results are, however, similar, and this simple conceptual method is maintained for this study. It is, however, clear that procedures such as these still depend largely on survey design and the size of intervals.

Consequently, the Pareto means (conditional on the range of each applicable category) can be calculated. Appendix 2 reveals that the imputed Pareto means can be obtained as follows, where *a* and *b* are the lower and upper bound of the *bounded* category concerned and $\hat{\alpha}$ is the regression estimator of the Pareto shape parameter:

$$\overline{y}_{pareto|a\le Y\le b} = \frac{\hat{\alpha}}{1-\hat{\alpha}} \frac{b^{-\hat{\alpha}+1} - a^{-\hat{\alpha}+1}}{a^{-\hat{\alpha}} - b^{-\hat{\alpha}}} \quad \hat{\alpha} > 1$$

These respective means are imputed to the interval-coded observations above the threshold referred to, given the applicable bounds. Below this point, the usual practice of midpoint imputation is followed.

For the *open-ended* interval, each right-censored value in that category is assigned the following mean:

$$\overline{y}_{pareto|Y\geq a} = \frac{\hat{\alpha}}{\hat{\alpha}-1}a \quad \hat{\alpha}>1$$

a is the lowerbound of the open interval, while $\hat{\alpha}$ is the regression estimator of the Pareto parameter.

4.2.3 Lognormal Means

Gustavsson (2004: 20-21) explains the implementation of a lognormal distribution over earnings data. This distribution also has a heavy tail, and justifies the assumption in its use as a distribution to fit income data. When data is expressed in log form, a normal distribution is fitted, and as a result the untransformed data will be lognormally distributed. The standard procedure is to use maximum likelihood estimation on the log of earnings to find the mean and variance for the distribution of the data available, and use these as parameters of a normal distribution to simulate the rest of the data. Maximum Likelihood procedures are complicated by iterative computations, which may prove to be time-inefficient. Integrals do not possess a closed form, and therefore various estimates do not converge to the same value when different techniques are used. The introduction of censored and interval-coded data adds further complications in the maximum-likelihood iterations. (See Sultan, 1997 and Hajivassilou, 2000 and Hajivassilou et al, 1996 for attempts to simplify and find satisfactory maximum likelihood estimates in the presence of Limited Dependent Variables).

It is first necessary to find the mean and standard error of the log-transformed variable. This study implements an interval regression on the log of point and categorical earnings data without regressors, bar for the constant. From this computation, an estimate of the distribution's mean (the constant) and its standard error (standard error of regression) is obtained. Appendix 3 elaborates the imputation of normal means to the intervals in log format by the following formula:

$$\overline{y}_{\text{normal}|a\le y\le b} = \hat{\mu} - \hat{\sigma}\frac{\phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - \phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}{\Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}$$

where

a: lowerbound of category

b: upperbound of category

$\hat{\mu}$: Estimator of Normal mean of logged data

$\hat{\sigma}$: Estimator of Normal standard deviation of logged data

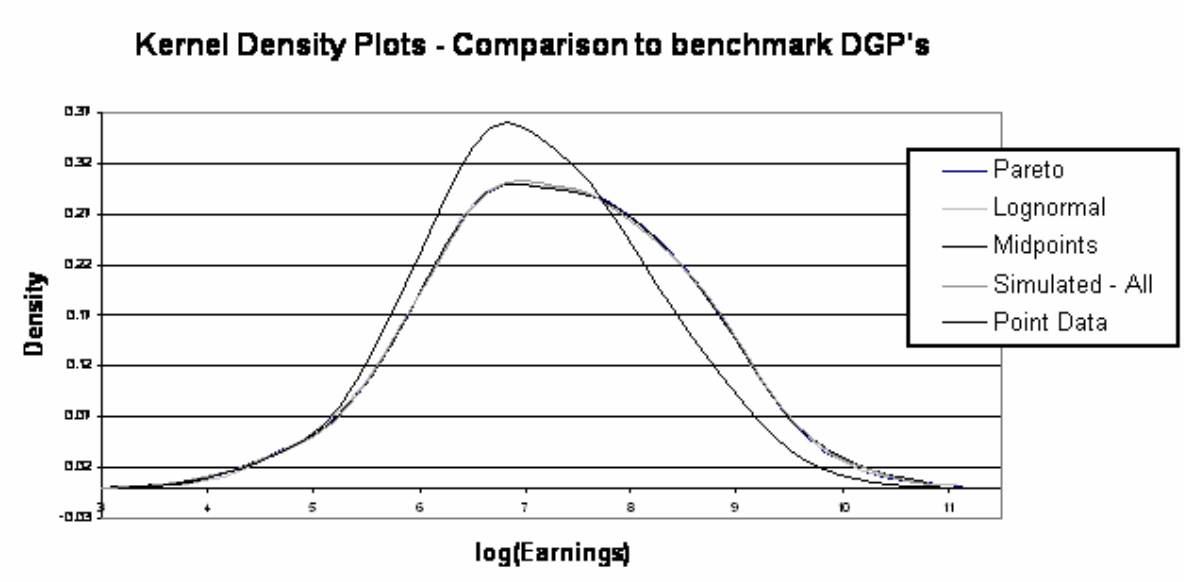
$\phi(x)$: Standard Normal Density Function

$\Phi(x)$: Standard Normal Cumulative Distribution Function

This can be applied to interval-coded data, as well as the bottom category (which has a lower bound tending to negative infinity) and the open-ended category (which has an upper bound tending towards positive infinity). When normal conditional means are imputed to the logged data, the variable in raw format by default assumes lognormal imputation.

4.3 Preliminary Evaluation of Imputations for LFS – September 2003

Figure 2 Kernel Density Plots - Comparison to benchmark DGP's



This section follows suggested and adapted methods of Keswell and Poswell (2004: 854-856), which are outlined further in Appendix 5.

The resulting densities in Figure 2 reveal the stark differences between the point data DGP and the joint sample DGP. It is clear that greater weight is assigned to the lower earnings groups when only point data is used. This therefore confirms the usefulness of including income brackets in survey question design, as it captures the lost information in the upper tail of the distribution. For this dataset it is evident that the distribution of the joint simulated data (which is assumed to represent the process underlying the interval regression), the midpoint, Pareto mean and the lognormal mean imputations are virtually indistinguishable. The continued analysis of all techniques is therefore justified, as the shape of each is similar. However, even mild potential deviations highlight the biases which the different techniques possibly accord to the data, and call for parametric estimates to separate the good methods from the poor.

5 Variables and Specification

The standard Mincerian Earnings Function (Mincer, 1974: 130) can be expressed as

$$\log Y_{s,t} = \log Y_0 + r_s s + r_p k_0 t - \frac{r_p k_0}{2T} t^2$$

where

$Y_{s,t}$ is the income earned by an individual with s years of education and t years of experience

r_s is the rate of return to s years of education

r_s is the rate of return to t years of experience

k_0 is the ratio of accrued investment to earnings when a worker enters the labour market

T is the positive net investment period

This formulation attempts to capture the full influence of human capital development on earnings: both within the educational system, but also the additional skills acquired following entry into the labour market.

Earnings of a person who has undergone s years of schooling and has been in the labour market for t years is indicated by $Y_{s,t}$ and Y_0 therefore represents the earnings which would accrue to a person who has neither any training nor work experience: it serves as the constant in a regression setup. The estimated returns which s years of schooling has in terms of earnings is captured by r_s, and similarly r_p represents returns on any post-schooling investments (on the job training, accumulated knowledge). Experience is incorporated into the model via t. The ratio of accrued

investment to earnings when a worker enters the labour market is k_0: jointly $r_p k_0$ - the coefficient of experience - represent the returns on a combination of schooling and labour market inputs. This term is also evident in the coefficient of the quadratic form of experience which enters the equation. The motivation for the negative sign is derived from the inclusion of T (the positive net investment period) in the denominator of the coefficient: this implies that as the timeframe over which workers increase their expertise progresses, the marginal returns of experience declines.

This specification simplifies in the current framework, with the resulting coefficients as implied from the above description:

$$\log(\text{Earnings}_i) = \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{experience}_i + \beta_3 \text{experience}_i^2 + \underline{\beta'}_{other}\,\underline{x}_i + \varepsilon_i \quad i = 1 \ldots N$$
$$\varepsilon_i \sim N(0; \sigma^2)$$

Now $\underline{\beta}_{other}'\underline{x}$ represents coefficients which are specific to the South African labour market, as well as advances in the understanding of the determinants of earnings. The inclusion of racial dummy variables, as well as controlling for union membership are (not exclusively) specific to South Africa, while the inclusion of a quadratic term for education has recently enjoyed attention in a wider spectrum of the literature. Elaborations of the specification and some characteristics of the sample used are highlighted below:

5.1 Earnings

The survey allows respondents to supply figures on the basis of weekly, monthly and annual earnings. Overtime, allowances and bonuses, before taxation and deductions, are accounted for (StatsSA, 2003a: 49-50). Monthly earnings are used in this analysis. The choice of this magnitude, as opposed to the hourly wage, should not affect the outcomes of the study significantly, given the assumption that the workers' life-time behaviour (in terms of their choice of working hours) is determined exogenously (Keswell & Poswell, 2004: 838). It is further questionable to convert interval data into other frequency domains (Daniels & Rospabé, 2005: 6). It would be preferable to use hourly wages to remove the effects of longer working weeks on earnings, however the inclusion of *log(hours worked per month)* as a regressor partially accounts for this discrepancy, as revealed in the arithmetic below. This term is expected to play a positive role in the determination of earnings.

$$\log(\text{Monthly Earnings}_i) = \underline{x}'_i \underline{\beta} + \beta_H \log(\text{hours per month}_i) + \varepsilon_i \quad i = 1 \ldots N \quad \ldots(1)$$

$$\Rightarrow \log(\text{Monthly Earnings}_i) - \log(\text{hours per month}_i) = \underline{x}'_i \underline{\beta} + (\beta_H - 1)\log(\text{hours per month}_i) + \varepsilon_i$$

$$\Rightarrow \log(\text{Hourly Wages}_i) = \log\left(\frac{\text{Monthly Earnings}_i}{\text{hours per month}_i}\right) = \underline{x}'_i \underline{\beta} + (\beta_H - 1)\log(\text{hours per month}_i) + \varepsilon_i$$

$$i = 1 \ldots N$$

Now specification (1) is a more flexible version of the earnings function, and nests the Mincerian function if $(\beta_H\text{-}1) = 0$.

The nature of the earnings data in the survey is central to this study and is discussed above. Mincer (1974: 130) motivates the use of logged earnings as a dependent variable as opposed to obtaining untransformed level estimates: experience and education can be expressed in time units, and need not be approximated by monetary equivalents. The sample is restricted to those typically assumed to be in the labour force, namely workers between the ages of 16 and 64. These bounds are reflected in the legal working age for the youth (who are assumed to be in education) and legislated pension ages.

5.2 Returns to education

Van der Berg and Burger (2003: 496) commence a study on educational inequalities by posing the question whether the intended rectification of human capital inequalities in South Africa is indeed being achieved via the education system. Bantu education policies and separate schooling systems are known to have introduced stark differences not only in the skill levels within the economy, but indirectly translated to the differentiated earnings achievable by various racial groups. The change of dispensation and the accompanying unified education system should (by inference) result in an equal footing within the labour market. The augmented Mincerian Earnings function proves to be a workhorse to measure progress or indeed a lack of success in this endeavour. While this study does not include interaction coefficients to measure effects on earnings of the particular education which Blacks, Whites, Coloureds and Indians receive, the methods employed here will improve any estimates indeed obtained, given the different imputations employed. Should education coefficients prove to be significant, and show increasing influence, pre-labour market human capital development can be deemed effective in South Africa. As this is the first step in eliminating disparities, it is important to monitor the progress of the educational institutions of this country. Experience in post-entry positions can only be effective in human

capital and earnings progression if a person has attained a suitable qualification to improve prospects in the first place. This is reflected in the Mincerian framework above, where the coefficient of experience combines factors relevant to both pre- and post-labour market entrance. Chamberlain & van der Berg (2002: 26) show that differential quality of education accounts for much of post-entry wage discrimination in the labour market: this underlines that this facet of human capital investment is perpetuated and influences the success of subsequent investments. Successful education will result in a better capacity to assimilate valuable experience, which in turn translates to improved earnings potential. Experience cannot be divorced from the basis of education.

5.2.1 Quadratic Specification – new evidence

Dacuycuy (2005: 2) implements a semiparametric procedure to establish the nature of the earnings function without assuming previous knowledge of its specification. That is, it does not assume that experience enters in both the quadratic and linear forms, and education only in the linear form. Estimation rather follows the following procedure:

$$E\left[\log\left(\text{Earnings}_i\right)\middle|\underline{x}_i\right] = \beta_0 + f_{X1}\left(\text{experience}_\text{i}\right) + f_{X2}\left(\text{education}_\text{i}\right) + \varepsilon_i$$
$$i = 1 \ldots N$$

The given $f_{Xi}(\bullet)$ are evaluated via kernel density estimation and an integration procedure. Dacuycuy (2005: 5) found that the relationship between earnings and schooling is indeed non-linear for the Philippines, and that when interactions are omitted, convexity exists.

This finding is confirmed for South Africa by Keswell and Poswell (2004: 844), who show that when controlling for potential experience, the returns to education are positive for the first 12 years of education. The additional positive quadratic term causes predicted income to rise even more sharply following this attainment. Tertiary qualifications benefit an entrant into the labour market with greater magnitude than a matric certificate. Chamberlain & van der Berg (2002: 26) conclude, with reference to a study of Mwabu and Schultz, that returns for a specific level of education decline as the proportion of the population attaining that level increases. This inverse relationship is evident in the high premium attached to tertiary qualifications. As a result, large imbalances exist, with secondary education (which can be regarded as some workers' most realistic opportunity to improve their earnings potential) adding little value compared to higher education (which is acquired by few).

5.3 Experience – approximation and relevance

Mincer (1974: 129) ascribes the primary reason for the inclusion of experience in the human capital framework to the fact that the completion of a schooling career does not conclude the investment in human capital. Further, these investments generally occur at a young age, with diminishing rates of new learning as people age – this translates to the declining additional earnings return available from current investments later in life and justifies the inclusion of a quadratic term in the specification. An approximation of experience is used (*exp=age-years in education – 6*) to separate as far as possible the returns of education from the returns from on the job training.

Keswell and Poswell (2004: 836) motivate their use of age instead of potential experience, due to specific factors in the South African labour market: the large number of learners who repeat years at school, the fact that a substantial proportion of learners do not complete the full number of years of education and that those within the labour market are not likely to be employed during all the years outside of formal education. Mincer (1974: 129-130) himself warns against the difficulties of approximating the variable in this fashion: in particular, it is evident for females that actual experience data is relevant instead of a variable defined largely by the individual's age. Indeed, the estimates in this study for the potential experience co-efficient do not reflect theory in the particular cases where females are included in the sample. This could be a result of unequal labour markets which still prevail in South Africa: females traditionally stay at home for longer years, hence *potential* experience does not count in their favour as much as the *actual* experience which males accrue.

Age, however, does not account for the actual on the job training which increases the earnings potential of workers – hence its use is not implemented in the earnings equations as such. Daniels & Rospabé (2005) employ a *tenure* variable, which is directly available from survey data. This accounts for the length of time respondents have spent at their *current* jobs. While this proxy is a well-defined quantity, and serves the cause of empirical accuracy, it is not deemed to represent a person's lifetime accumulated knowledge and expertise. It may well capture firm-specific skill acquisition. The longer an employee stays with a firm, the better are the prospects for internal promotion.

Further, job reservation in the previous dispensation and Black Economic Empowerment in the current may undercut the relevance of raw experience in earnings determination. Rather, earnings may be reflected by the influence of labour market regulations. These peculiarities may form interesting corollaries within future studies.

Since this study does not attempt to establish directly the determinants of earnings in South Africa, these difficulties are noted. Coefficients are tested for stability, and not used to confirm or refute their theoretical bases.

5.4 Racial Dummies

The nature of South Africa's labour market and the historical context dictates that racial dummies are still significant in earnings equations. A number of studies incorporate the imbalance by estimating separate equations for Black and White cohorts (see Chamberlain & van der Berg, 2002; Mwabu & Schultz, 2000; Rospabé, 2002), which has useful applications for decomposition of differences and discrimination in wages among racial groups. Rospabé (2002: 210) concludes that while there has been a reduction in earnings and employment gaps among the races, differences remain substantial. Earnings functions therefore have important applications in establishing whether restitutive legislation has proposed effects. A further option, implemented here, includes dummies within a joint equation (in terms of race). The Black cohort is chosen as a basis (excluded from the analysis), with relative estimates obtained for Whites, Indians and Coloureds.

5.5 Union Membership Dummy

Hofmeyr (1999) and Hofmeyr & Lucas (2001) investigate in particular the role of unionisation in South African labour markets. During apartheid years, Black South Africans were restricted in their job prospects; in addition collective units such as labour unions were banned under the dispensation. Already during the run-up to the regime change, active moves were implemented to protect workers. As a consequence, the labour market (which has an undersupply of labour relative to demand) is segmented and polarised into unionised high earners and non-unionised low-earning workers, despite the same productivity potential in both cohorts (Hofmeyr, 1999: 1). The increasing influence of unionisation is witnessed by the escalating wage premium of unionised workers over non-unionised workers (for urban African males) from 8% in 1985 to 23.5% in 1993 (Hofmeyr & Lucas, 2001: 708). The effect of the union dummy therefore quantifies "non-investment" (in Mincerian terms) action which influences earnings positively.

5.6 Urban

Bhorat & Leibbrandt (2001: 124) use the urban-rural dummy to distinguish between geographic divisions. Provincial dummies may present results contrary to common knowledge, as each province has internally heterogeneous characteristics which attract different workers. While these may possess interpretative value, parsimony is emphasised and the focus remains on the urban dummy. The urban rural divide has a clearer differentiating power, as the type of work (with associated earnings) is more accurately divided between the groups: industrial workers are likely to agglomerate in urban areas, while agricultural workers will remain in rural areas. In South Africa a positive return for urbanisation is witnessed, as in any modern economy.

5.7 Selection Equation

The selection equation's specification includes household and demographic variables which may hinder or lead people to seek employment. Age is expected to have a positive influence, as the yet unskilled youth is less probable to be employed (this may also be indicative of the fact that recent jobless growth means low absorption rates of younger workers, and retention of older employees). Provincial dummies reflect the unemployment situation within each region. The number of children younger than 6 in a household particularly influences a female's choice to participate in the labour market negatively, as care is afforded to her offspring, while males might seek employment more fervently to provide for the young. The number of working and pension age household inhabitants also negatively influences the probability of employment via the decision to participate, as a result of a household safety net. Per capita household income should negatively influence the probability of employment, as other household members support each other in the case of unemployment. Care was taken to heed the warning of Hill et al (2003: 18) to keep variables in the earnings equation separate from the selection specification to avoid adverse effects on standard errors.

6 Results

6.1 Simulation Evidence

A known dependent variable was generated according to the following structure:

$$y = x + e \quad e \sim N(0,1)$$

y was subsequently converted into both narrow and wider categories[2], and simulations proceeded with 1000 repetitions, using midpoint OLS and interval regression estimates. Results are shown in Table 2. Narrow intervals appear to be insensitive to either method, with mean coefficients very close to the true value of 1. P-values do not deviate substantially from the expected 0.05. Wider intervals' coefficients do not appear visually different from 1, however, the p-value for the midpoint imputations (0.143) is significantly larger than 0.05. Interval regressions remain relatively unscathed. This highlights that if intervals are "too wide", midpoint imputation distorts inference, compared to the sustained reliability of interval regressions. It is therefore imperative to establish whether survey brackets in South Africa are suitably narrow with parametric comparisons. Should they not be, it is evident that interval regressions are the most suitable econometric tools to prevent misleading judgments.

Table 2 Monte Carlo Simulation - Midpoints and Interval Regressions

α=1, 1000 repetitions	**Narrower Intervals**		**Wider Intervals**	
	Midpoints (OLS)	**Interval Regression**	**Midpoints (OLS)**	**Interval Regression**
Coefficient	0.9998	0.9996	0.9850	0.9997
(p-value)[3]	(0.042)	(0.050)	(0.143)	(0.056)

6.2 A brief word on some of the coefficients

While this study is focussed on parameter comparisons, a short exposition of their magnitudes is called for. Discussion is limited to the male equation with bootstrapped confidence intervals (Table 12). The inclusion of females in the sample distorts economic interpretation via the experience variable, as discussed above.

First, the sample selection correction term is significant in all cases. This underlines that earnings and employment processes are intertwined. In South Africa, the high unemployment scenario should be kept in mind when wage determination is considered.

Experience enters positively and experience squared negatively, which underlines Mincerian theory. The convexity of returns to education is confirmed in this context, with both the linear and quadratic forms exhibiting a positive relationship with earnings. While education is the only variable to enter insignificantly (at a 5% level) for all methods in the linear form, it is significant in the quadratic form and joint interpretation should be exercised. White males reap larger returns than Indian counterparts, who earn more than Coloureds, who in turn earn more than Black

[2] Narrow intervals divided the income range into 7 equally sized ranges, with open ends on either side. Wider intervals were twice as large. Open categories exceeded their nearest bounds by 10% for the midpoint imputation.
[3] The proportion of the 1000 replications for which the *T* statistic which tests H_0: $a=1$ exceeds 1.96, the 95th percentile of the applicable *t* distribution.

males. This scenario depicts the racial segmentation still prominent from the Apartheid labour market. Urbanisation and unionisation exhibit positive returns, as expected.

The most interesting feature of the coefficients is that the influence on earnings of the Mincerian variables is only small in the context of the rest of the model: the coefficients on human capital investment are overshadowed by "non-investment" features such as race, location and union membership. This underlines the fact that South African earnings are still largely determined by "by-products" of political marks on society and interference in the labour market. Traditional routes to improve earnings (and the equality thereof) therefore prove to have little effect. Can one really rely on education to make disparities obsolete? Even union membership provides greater returns than an additional year of education or experience. Ineffective labour markets therefore allocate more reward to non-productive activity than to skills development.

6.3 Method Comparison by Confidence Intervals

This section presents intuitive evidence of parameter equality: do the compared coefficients' 95% confidence intervals overlap? Tables 5, 9 and 13 provide a good overview of the results obtained via bootstrap methods. The cross-tabulations consider whether the coefficients of the interval regression and the imputations fall within each others' confidence intervals. Similar computations were done for the robust intervals, which produced near identical conclusions. The overwhelming result is that all coefficients (independent of method or imputation) fall within each of the others' 95% confidence interval. This is true for each cohort investigated. The results obtained here therefore confirm the preliminary analysis performed on the data. It should be emphasised again, that these conclusions apply to the specific data structure concerned and cannot be generalised to all household surveys. While a quick scan of the coefficients would convince the analyst that they approach equality, the naked eye fails to detect some underlying statistical differences.

6.4 Multivariate Testing Framework

Appendix 6 introduces a more rigorous multivariate testing procedure to test the intuitive results obtained above. It is possible to model a multivariate regression, with each of the imputations (interval regressions are not compatible with this framework) constituting the dependent variable vector with a common matrix of explanatory variables. It is simple to perform joint Wald tests to compare coefficients across the constituent equations. This procedure takes into account not

only the variances of the coefficients, but also their covariances. A Bonferonni adjustment is implemented to account for the dependencies of hypotheses.

It is evident that no estimated equation is in its entirety equivalent to another in each of the samples. Which variables drive the differences? In each case, the coefficient of the Inverse Mills Ratio in one equation significantly differs from that in the others at a 5% level of significance. At this point, sceptics might question the inclusion of a selection correction term within the interval regression. The exclusion, however, delivers substantially different results compared to any of the traditional imputation methods, which *do* require the correction (the coefficient of the Inverse Mills' Ratio is statistically significant at 1% in all cases, even in the interval regression). The stochastic nature of the Inverse Mills Ratio therefore provides a more satisfactory explanation for the discrepancy. The equality of magnitudes, is however not the emphasis in this case, but the fact that it corrects for biases.

The next striking feature of the analysis is the large number of differences between the coefficients of the Lognormal and Pareto-Midpoint imputation equations. The source of this discrepancy can be traced to the fact that the Pareto-Midpoint method was applied in a gender-specific manner, while the lognormal imputation considered the sample jointly. The Pareto-Midpoint variable is not determined by a single imputation and does not result in a satisfactory representation of the DGP. Researchers should take care to generate separate imputations, specific to the sub-grouped or entire samples to be used.

The male estimates are least affected by this difficulty (with only the coefficients of *Selection, White* and *Indian* rejecting the hypothesis that $\beta_{pareto} = \beta_{\log normal}$ at a 5% level of significance). This suggests that the specific male imputation is best at capturing information which is also relevant to an imputation which considers the entire sample. Female estimates degenerate further, with *Selection, Experience, Education, Education*^{2} *, Coloured and Union* rejecting the hypothesis at 5%. This list is dominated by the Mincerian "investment" variables, which highlights that these are not particularly stably determined for females, as noted above. The female-specific imputation is less representative of a joint imputation. This highlights that researchers are able to model the male DGP with greater ease, but that the underlying process in the female sample is less well-known and differs more substantially from the entire population's DGP than does the male DGP. The single equation estimates compare very poorly; this discredits a gender-specific imputation strategy.

The fact that these differences are fewer when Midpoint estimates are compared to the Lognormal imputation, (in particular that only *Selection, Coloured and Urban* differ at a 5% level of significance for males), reveals that a sample-specific imputation does matter. In both cases imputations were conducted on the sample as a whole.

Further differences are few. It should be noted, however, that the male equation appears to have the most stable coefficients spanning the different methods employed. For the Midpoint/Midpoint-Pareto comparison, only the Inverse Mills ratios are statistically different (this can be ascribed to the fact that the lower tail is generated identically, but still asserts that the upper tails are close to each other), which is joined by *Coloured* and *Urban* in the Midpoint-Lognormal comparison: it is encouraging that these are not Mincerian variables, which form the basis all earnings equations. This study therefore confirms specifically for the male sample, that the different imputations exhibit some statistical differences from each other, however many of these can be ascribed to methodological strategy. Overall, for males, traditional Mincerian models can be modelled with confidence by any method : some other coefficients might fail rigorous statistical tests, though the intuitive results show that they are economically similar.

6.5 Robust or Bootstrapped Confidence Intervals?

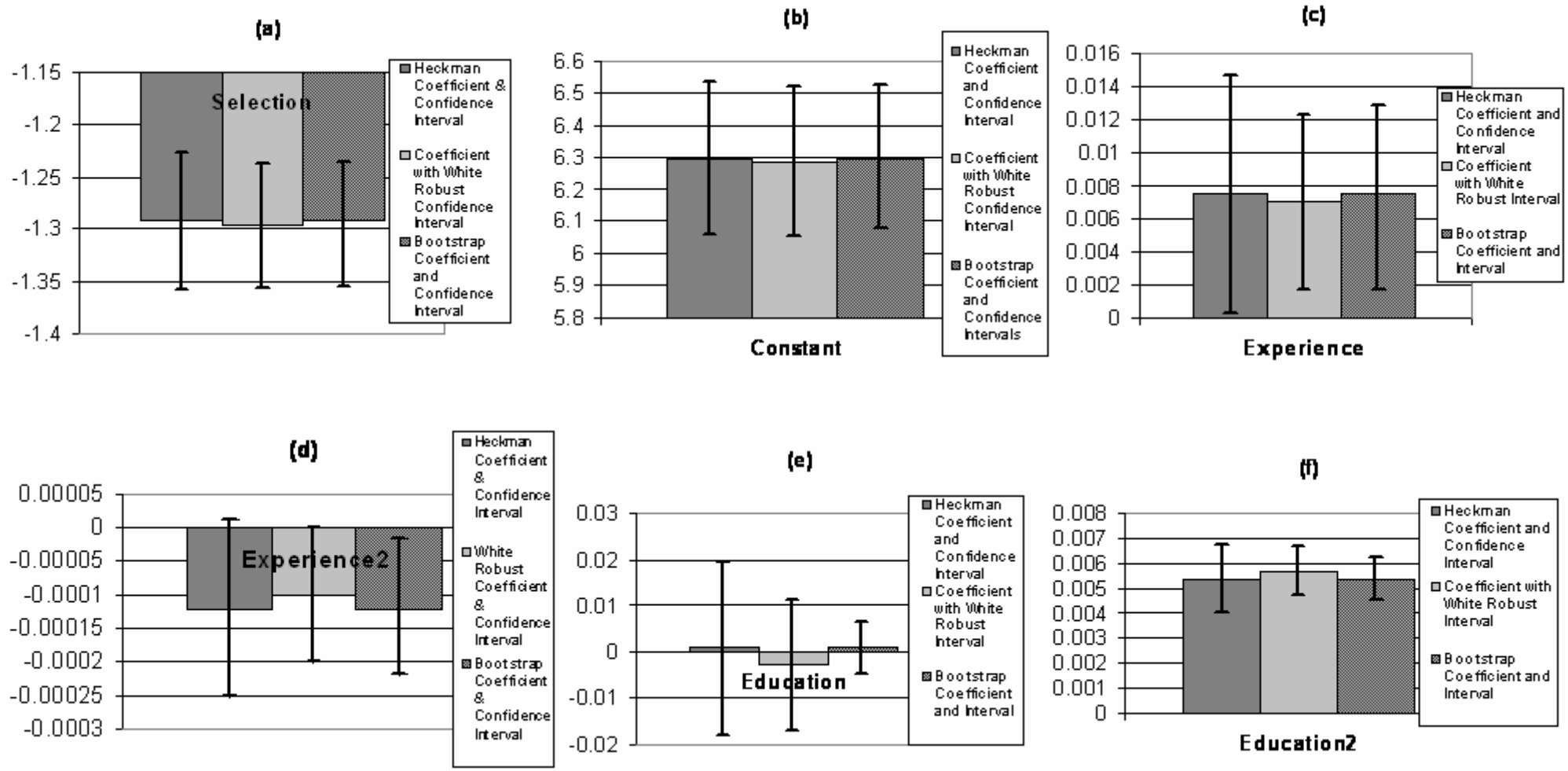
Figure 3 investigates confidence intervals, given the applied corrections: Male midpoint estimates are chosen to illustrate conclusions, as the interval regression cannot be readily implemented with a Heckman covariance structure. The most apparent feature is that the Heckman intervals are the broadest. The bootstrap and robust intervals are fairly close to each other in length. While some bootstrapped intervals exhibit an improvement in efficiency (compared to the robust intervals), this is only very conclusive in the case of *Education*; many other cases deliver no efficiency gains, and in some instances the robust intervals are the most efficient.

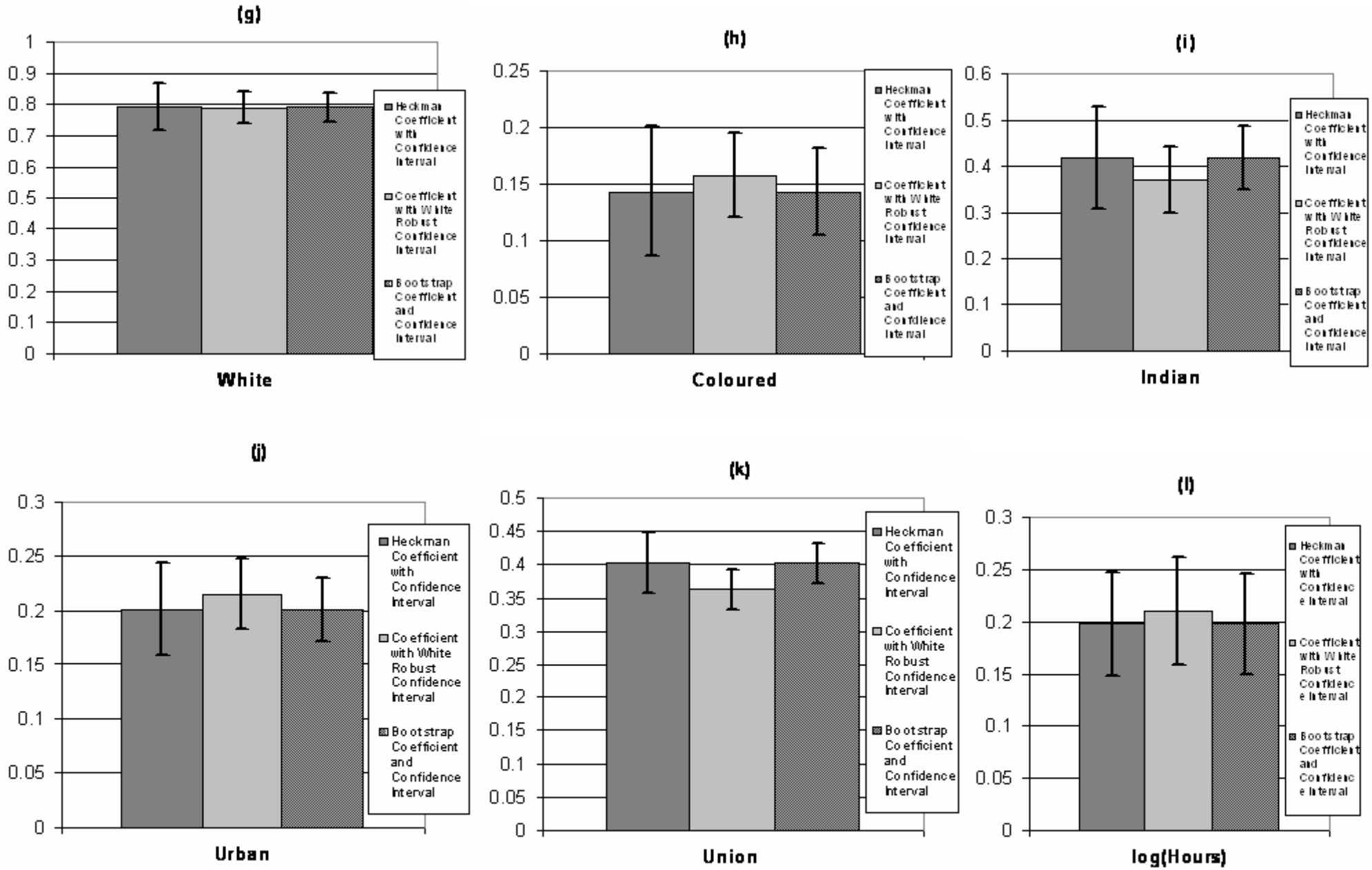
As Hill et al (2003: 19) conclude, work on large samples with little censoring produces satisfactory conclusions when the usual Heckit covariance matrix is combined with bootstrap estimation. Smaller samples with extensive censoring require heteroskedasticity corrected covariance forms in conjunction with bootstrap techniques. This sample can therefore be seen as relatively unscathed by censoring and finite size, and may even perform well without bootstrap estimates (but still require a form of correction). The efficiency gains from corrections are notable, however the additional gains from bootstrapping are limited. Judged in the light that this

dataset appears to be largely unaffected by censoring, it would be logical to conclude that the influence of imputed values play a less critical role. It is therefore important to undergo pre-testing to establish whether censoring is a significant problem: in this case it is necessary to establish whether imputations will accord different information to the sample, and steps to achieve greater efficiency are required (to avoid non-rejection of an invalid hypothesis). Bootstrapping is advocated with a correction incorporated in the procedure, as emphasised by Hill et al (2003: 19).

The advances in software are indicative of the general consensus as to the suitability of various confidence intervals and standard errors. STATA9.0's built-in Heckit 2-step procedure is no longer directly compatible with robust corrections: should these be desired, the Inverse Mills Ratio needs to be obtained manually and included in an earnings regression. However, the Heckman structure is still available (presumably to stay in line with the original theory) for computational ease and because robust modifications have not improved the overall outlook to the extent to which bootstrapping has. Statistics by the latter method are now directly computable within the procedure: this is deemed the most preferable, as it doesn't only provide "half" a correction in seriously censored samples. The trade-off is, however, computational time.

Figure 3 Comparison of Coefficients Magnitudes and 95% Confidence Intervals (Male Midpoint Estimates)




(g)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
White
Heckman Coefficient with Confidence Interval
Coefficient with White Robust Confidence Interval
Bootstrap Coefficient and Confidence Interval
(h)
0.25
0.2
0.15
0.1
0.05
0
Coloured
Heckman Coefficient with Confidence Interval
Coefficient with White Robust Confidence Interval
Bootstrap Coefficient and Confidence Interval
(i)
0.6
0.5
0.4
0.3
0.2
0.1
0
Indian
Heckman Coefficient with Confidence Interval
Coefficient with White Robust Confidence Interval
Bootstrap Coefficient and Confidence Interval
(j)
0.3
0.25
0.2
0.15
0.1
0.05
0
Urban
Heckman Coefficient with Confidence Interval
Coefficient with White Robust Confidence Interval
Bootstrap Coefficient and Confidence Interval
(k)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Union
Heckman Coefficient with Confidence Interval
Coefficient with White Robust Confidence Interval
Bootstrap Coefficient and Confidence Interval
(l)
0.3
0.25
0.2
0.15
0.1
0.05
0
log(Hours)
Heckman Coefficient with Confidence Interval
Coefficient with White Robust Confidence Interval
Bootstrap Coefficient and Confidence Interval

7 Conclusion

While the specific requirements of bracket data and sample selection pose obstacles before one can commence with economic interpretation, one need not lean on these as an excuse to call data "bad". Adler et al (1998: ix) also define the perceptions of "good data" as those which researchers are readily able to use: note that it is labelled in terms of *perceptions*. Have the methods of this study altered views of datasets? It has been shown that particularly for LFS September 2003, parameters exhibit reasonable stability, regardless of which techniques are applied. While certain statistical differences are apparent (and potential reasons can be pinpointed), the fact that magnitudes' confidence interval estimates overlap in all cases, reveals that for the purposes of economic interpretation, a satisfactory compromise has been reached. Ziliak and McCloskey (2004) in fact warn econometricians not to attach the entire emphasis of conclusions to statistical results, when economic magnitude is of importance. In this case, the economic quantities are for all intents and purposes the same, regardless of whether new econometric techniques (interval regressions) or traditional imputation methods are applied. It should be emphasised that these properties are specific to this dataset, but that kernel density estimation can quickly reveal the validity of any proposed imputation for any dataset.

Developments in thinking on impure variance-covariance structures have also improved accuracy in the earnings function framework. Whilst simple asymptotic corrections do not provide a rosier outlook, the sacrifice of computational time by bootstrapping is certainly a price worth paying. The lack of distributional assumptions removes any undue restrictions in inference, and underlines the cause for simulations in the evaluation of models. The small changes in efficiency, however, suggest that this dataset is not unduly affected by censored values – a possible reason why different imputations do not swing the results.

The tools in this shed therefore prove themselves to be sharp for the purposes of economic evaluation. The simplest methods are interval regressions, midpoint imputations and lognormal mean imputations (in that order). The estimation of the α parameter for the Pareto tail is somewhat restrictive. The suitability of traditional methods (should they be employed) should be confirmed before potentially biased results are held to depict true magnitudes.

The validation of these methods certainly does translate perceptions of "bad data" to "good data", and researchers should feel confident to apply them until further advances are made. These tools capacitate researchers to analyse the South African labour market; they enable a depiction

of reality, and add value to the Mincerian-Heckman framework. Economists can therefore make more accurate recommendations, despite the fact that they are supplied with information which is not conventionally easy to process. In effect, interval-coding should not deter labour market analysis, but add significant information and lead to improved practice in econometrics.

8 Bibliography

ADLER, R.J., FELDMAN, R.E. and TAQQU, M.S., 1998. *A Practical Guide to Heavy Tails – Statistical Techniques and Applications*. Boston: Birkhäuser.

BHORAT, H. and LEIBBRANDT, M., 2001. Modelling Vulnerability and Low Earnings in the South African Labour Market. In Bhorat, H., Leibbrandt, M., Maziya, M., van der Berg, S. and Woolard, I. *Fighting Poverty – Labour Markets and Inequality in South Africa*. Landsdowne: UCT Press.

BIRKIN, M. AND CLARKE, G., 1995. Using microsimulation methods to synthesize census data. In Openshaw, S. (ed.), Census Users' Handbook. Cambridge: Pearson Professional Limited.: 363-387.

BROWN, B.W., 2000. Simulation Variance Reduction for Bootstrapping. In Mariano, R., Schuermann, T. and Weeks, M.J. (eds), Simulation-Based Inference in Econometrics – Methods and Applications. Cambridge: Cambridge University Press. pp437-457

BROWNSTONE, D. and VALETTA, R., 2001. The bootstrap and multiple imputations: Harnessing Increased computing power. *The Journal of Economic Perspectives*. Fall 2001, Vol 15: 4. 129-141.

CHAMBERLAIN, D. and VAN DER BERG, S., 2002. *Earnings Functions, Labour Market Discrimination and Quality of Education in South Africa*. Stellenbosch Economic Working Papers: 2/2002.

CRATO, N., 2000. Estimation Of The Maximal Moment Exponent With Censored Data. *Communications in Statistics – Simulation and Computation*, Vol29 No4: 1239-1254.

DACUYCUY, L., 2005. Is the earnings-schooling relationship linear? A semiparametric analysis. *Economics Bulletin,* Vol. 3, No. 37 pp. 1−8.

DAGSVIK, J.K. and VATNE, B.H., 1999. *Is the Distribution of Income Compatible with a Stable Distribution?* Discussion Paper No. 246, Statistics Norway, Research Department. Kongsvinger.

DANIELS, R. and ROSPABÉ, S. 2005. *Estimating an Earnings Function from Coarsened Data by an Interval Censored Regression Procedure*. Development Policy Research Unit Working Paper 05/91.

GUSTAVSSON, M., 2004. *Trends in the Transitory Variance of Earnings: Evidence from Sweden 1960-1990 and a Comparison with the United States*. Uppsala University, Economics Working Paper 2004:11. Available [Online]: http://www.sofi.su.se/sem/GustavssonMagnus.pdf

HAJIVASSILOU, V.A., 2000. Some practical issues in maximum simulated likelihood. In Marioan, R., Schuermann, T., Weeks, M.J., (eds.), *Simulation-Based Inference in Econometrics – Methods and Applications*. Cambridge: Cambridge University Press. pp 71-99

HAJIVASSILOU, V.A., McFADDEN and D., RUUD, P., 1996. Simulation of multivariate normal rectangle probabilities and their derivatives - Theoretical and computational results. *Journal of Econometrics*. 72 (1996) 85- 134.

HAYASHI, F., 2000. *Econometrics*. Princeton: Princeton University Press.

HECKMAN, J.J., 1979. Sample Selection Bias as a Specification Error. *Econometrica*. Vol 47, No. 1. : 153-161.

HENDERSON, A.R., 2005. The Bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clinica Chimica Acta 35*: 1-26.

HILL, R.C., ADKINS, L.C. and BENDER, K.A., 2003. *Test Statistics and Critical Values in Selectivity Models*. Available [Online]: http://www.bus.lsu.edu/academics/economics/faculty/chill/personal/heckit.pdf Later Published in: Fomby, T. and Carter Hill, R. (eds), 2003, *Maximum Likelihood Estimation Of Misspecified Models: Twenty Years Later*. Elsevier

HOFMEYR, J.F. and LUCAS, R.E.B., 2001. The Rise in Union Wage Premiums in South Africa. *Labour* 15 (4): 685-719.

HOFMEYR, J.F., 1999. *Segmentation in the South African Labour Market in 1999*. Working Paper No. 15. South African Network of Economic Research: Potchefstroom.

KESWELL, M. and POSWELL, L., 2004. Returns to Education in South Africa: A Retrospective Sensitivity Analysis of the Available Evidence. *The South African Journal of Economics*. Vol 72:4. September 2004.

MADDALA, G.S., 1983. *Limited Dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

MINCER, J.,1974. *Schooling, Experience and Earnings*. New York: National Bureau of Economic Research

MWABU, G. and SCHULTZ, T.P., 2000. Wage Premiums for Education and Location of South African workers, by Gender and Race. *Economic Development and Cultural Change*. Vol 48 No.2: 307-334.

PARKER, R.N. and FENWICK, R., 1983. The Pareto Curve and Its Utility for Open-Ended Income Distributions in Survey Research. *Social Forces*. Vol 61:3: 872-885.

RICE, J.A., 1988. *Mathematical Statistics and Data Analysis*. Pacific Grove, CA: Wadsworth and Brooks.

ROSPABÉ, S., 2002. How did Labour Market Racial Discrimination Evolve After the End of Apartheid?. *The South African Journal of Economics*. Vol 70 No.1: 185-217.

SEIVER, D.A., 1979. A Note on the Measurement of Income Inequality with Income Data. *The Review of Income and Wealth*. 1979, Series 25: 229-234.

STATACORP, 2003. *Stata Base Reference Manual Volume 4. S-Z, Release 8*. College Station TX: Stata Press.

STATACORP, 2005. *Stata/SE 9.0 for Windows*, College Station TX: StataCorp LP.

STATISTICS SOUTH AFRICA (StatsSA), 2003a. *METADATA. Labour Force Survey2003:2*. Pretoria: Statistics South Africa.

SULTAN, A.M., 1997. New Approximation For Parameters Of Normal Distribution Using Type II-Censored Sampling. *Microelectronics Reliability,* Vol. 37, No. 8, pp. 1169-1171

VAN DER BERG, S. and BURGER, R., 2003. Education and Socioeconomic Differentials: A Study of School Performance in the Western Cape. *The South African Journal of Economics*. Vol 71: 3.

WEST, S.A., 1986. *Estimation of the Mean from Censored Income Data.* Proceedings of the Survey Research Methods Section, American Statistical Association: 665-670 Available [Online]: http://www.amstat.org/sections/srms/Proceedings/

WHITEFORD, A. and McGRATH, M. 1994, *The Distribution of Income in South Africa.* Pretoria: Human Sciences Research Council.

WIK, M., KEBEDE, T.A., BERGLAND, O. and HOLDEN, S.T., 2004. On the measurement of risk aversion from experimental data. *Applied Economics*. 36: 2443-2451.

WINTER, C., 1999. *Women Workers in South Africa: Participation, Pay and Prejudice in the Formal Labour Market*. South Africa: Poverty and Inequality – Informal Discussion Paper Series 19752, World Bank Country Department I, Africa Region. Washington: World Bank.

WOOLDRIDGE, J.M., 2002. *Econometric Analysis of cross section and panel data.* Cambridge, MA: MIT Press.

ZILIAK, S.T., and McCLOSKEY, D.N., 2004. Size Matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*. 33 (2004): 527-546.

APPENDIX 1 – ESTIMATION RESULTS

Single Equation

Table 3 Automated Heckit Estimates with Heckman Covariance Matrix – Single Equation

Single Equation	**Heckit Estimates with Heckman Covariance Matrix**		
	Midpoint	**Lognormal**	**Midpoint-Pareto**
Selection (λ)	**-1.39674**	**-1.382957**	**-1.390056**
	(-1.449898 -1.343582)**	(-1.43559 -1.330324)**	(-1.442959 -1.337152)**
Experience	**-0.00635**	**-0.00593**	**-0.00632**
	(-0.01197 - -0.00072)*	(-0.01151 - -0.00036)*	(-0.01192 - -0.00072)*
Experience²	**-0.00005**	**-0.00005**	**-0.00004**
	(-0.00015 - 0.00006)	(-0.00015 - 0.00005)	(-0.00015 - 0.00006)
Education	**0.00299**	**0.0028**	**0.00352**
	(-0.01143 - 0.01740)	(-0.01147 - 0.01707)	(-0.01082 - 0.01787)
Education²	**0.00483**	**0.00484**	**0.00479**
	(0.00382 - 0.00585)**	(0.00384 - 0.00585)**	(0.00378 - 0.00580)**
White	**0.6585**	**0.65693**	**0.65598**
	(0.59961 - 0.71739)**	(0.59862 - 0.71524)**	(0.59737 - 0.71459)**
Coloured	**0.10808**	**0.10575**	**0.1084**
	(0.06244 - 0.15371)**	(0.06056 - 0.15093)**	(0.06298 - 0.15382)**
Indian	**0.43079**	**0.42977**	**0.42989**
	(0.34135 - 0.52023)**	(0.34121 - 0.51832)**	(0.34088 - 0.51890)**
Urban	**0.16921**	**0.16761**	**0.16874**
	(0.13652 - 0.20190)**	(0.13524 - 0.19998)**	(0.13620 - 0.20127)**
Union	**0.52641**	**0.52493**	**0.52609**
	(0.48953 - 0.56330)**	(0.48841 - 0.56145)**	(0.48938 - 0.56280)**
log(Hours)	**0.33222**	**0.32978**	**0.33309**
	(0.29958 - 0.36487)**	(0.29746 - 0.36211)**	(0.30060 - 0.36558)**
Constant	**6.10502**	**6.09994**	**6.09461**
	(5.93144 - 6.27861)**	(5.92807 - 6.27182)**	(5.92186 - 6.26737)**
Observations	38469	38469	38469

95% confidence intervals in parentheses

* significant at 5%; ** significant at 1%

Table 4 Manual Weighted Heckman 2-step with Robust Confidence Intervals - Single Equation

Single Equation	**Manual Weighted Heckman 2-step with Robust Confidence Intervals**			
	Midpoint	**Lognormal**	**Midpoint-Pareto**	**Interval Regression**
Selection (λ)	**-1.40195**	**-1.38915**	**-1.39409**	**-1.39007**
	(-1.44711 - -1.35678)**	(-1.43438 - -1.34392)**	(-1.43970 - -1.34847)**	(-1.43485 - -1.34530)**
Experience	**-0.00536**	**-0.00497**	**-0.0053**	**-0.00493**
	(-0.00939 - -0.00134)**	(-0.00899 - -0.00095)*	(-0.00934 - -0.00125)*	(-0.00892 - -0.00094)*
Experience²	**-0.00005**	**-0.00006**	**-0.00005**	**-0.00006**
	(-0.00013 - 0.00002)	(-0.00013 - 0.00001)	(-0.00013 - 0.00002)	(-0.00013 - 0.00001)
Education	**0.00041**	**0.00005**	**0.00091**	**0.0001**
	(-0.01083 - 0.01166)	(-0.01114 - 0.01125)	(-0.01038 - 0.01220)	(-0.01108 - 0.01129)
Education²	**0.00513**	**0.00515**	**0.00509**	**0.00517**
	(0.00437 - 0.00589)**	(0.00439 - 0.00591)**	(0.00433 - 0.00586)**	(0.00441 - 0.00593)**
White	**0.65156**	**0.65098**	**0.64968**	**0.6516**
	(0.61163 - 0.69149)**	(0.61068 - 0.69128)**	(0.60941 - 0.68995)**	(0.61180 - 0.69139)**
Coloured	**0.10754**	**0.10501**	**0.10775**	**0.10615**
	(0.07904 - 0.13604)**	(0.07672 - 0.13330)**	(0.07938 - 0.13612)**	(0.07801 - 0.13428)**
Indian	**0.38217**	**0.38115**	**0.38069**	**0.38241**
	(0.32938 - 0.43495)**	(0.32882 - 0.43347)**	(0.32826 - 0.43312)**	(0.33039 - 0.43444)**
Urban	**0.19482**	**0.19256**	**0.19477**	**0.19331**
	(0.16870 - 0.22095)**	(0.16665 - 0.21846)**	(0.16860 - 0.22093)**	(0.16749 - 0.21913)**
Union	**0.4755**	**0.47436**	**0.47541**	**0.4751**
	(0.45142 - 0.49958)**	(0.45035 - 0.49837)**	(0.45138 - 0.49945)**	(0.45119 - 0.49902)**
log(Hours)	**0.33348**	**0.33117**	**0.33442**	**0.33296**
	(0.30154 - 0.36542)**	(0.29964 - 0.36271)**	(0.30244 - 0.36641)**	(0.30123 - 0.36469)**
Constant	**6.10884**	**6.10509**	**6.09662**	**6.09585**
	(5.95212 - 6.26555)**	(5.94913 - 6.26106)**	(5.93879 - 6.25445)**	(5.94005 - 6.25165)**
Observations	21389	21389	21389	21389
R-Squared	0.64348	0.64418	0.64192	

Robust 95% confidence intervals in parentheses
* significant at 5%; ** significant at 1%
Heckit Interval Regression is Estimated Manually

Table 5 Bootstrapped Coefficients and Bias-Corrected Confidence Intervals - Single Equation

Single Equation	**Heckit with Bias-Corrected Bootstrapped Confidence Intervals**							
	Midpoint		**Lognormal**		**Midpoint-Pareto**		**Interval Regression**	
Selection (λ)	**-1.39674**	***0.00184860***	**-1.38296**	***0.00197050***	**-1.39006**	***0.00220980***	**-1.38312**	***-0.00002210***
	-1.44543	-1.35281500*	-1.43047	-1.33908100*	-1.43989	-1.34681000*	-1.42282	-1.34401800*
Experience	**-0.00635**	***0.00009280***	**-0.00593**	***0.00009110***	**-0.00632**	***0.00011080***	**-0.00581**	***-0.00000184***
	-0.01082	-0.00202280*	-0.01039	-0.00174700*	-0.01084	-0.00228590*	-0.00923	-0.00249910*
Experience²	**-0.00005**	***-0.00000100***	**-0.00005**	***-0.00000099***	**-0.00004**	***-0.00000113***	**-0.00005**	***0.00000008***
	-0.00012	0.00003520	-0.00013	0.00003100	-0.00012	0.00003630	-0.00011	0.00001150
Education	**0.00299**	***-0.00000391***	**0.00280**	***-0.00008210***	**0.00352**	***-0.00009470***	**0.00278**	***-0.00004400***
	-0.00669	0.01226570	-0.00658	0.01195850	-0.00580	0.01333450	-0.00644	0.01188870
Education²	**0.00483**	***0.00000656***	**0.00484**	***0.00001090***	**0.00479**	***0.00001390***	**0.00487**	***0.00000221***
	0.00419	0.00548490*	0.00421	0.00547930*	0.00412	0.00543380*	0.00424	0.00549840*
White	**0.65850**	***0.00064230***	**0.65693**	***0.00096840***	**0.65598**	***0.00103880***	**0.65701**	***0.00012420***
	0.62054	0.69395800*	0.61902	0.69318780*	0.61760	0.69296080*	0.62161	0.69181400*
Coloured	**0.10808**	***0.00074160***	**0.10575**	***0.00057580***	**0.10840**	***0.00055990***	**0.10621**	***0.00001800***
	0.07589	0.13803500*	0.07396	0.13611050*	0.07709	0.13892610*	0.08216	0.13072280*
Indian	**0.43079**	***0.00015840***	**0.42977**	***0.00024750***	**0.42989**	***0.00037530***	**0.42952**	***0.00040990***
	0.38095	0.48324900*	0.37798	0.48207390*	0.37622	0.48033590*	0.38012	0.47962090*
Urban	**0.16921**	***0.00013380***	**0.16761**	***0.00025610***	**0.16874**	***-0.00013580***	**0.16915**	***0.00001730***
	0.14593	0.19191510*	0.14504	0.19009250*	0.14587	0.19220200*	0.14739	0.19043960*
Union	**0.52641**	***0.00032140***	**0.52493**	***0.00028260***	**0.52609**	***0.00054150***	**0.52553**	***0.00008000***
	0.50179	0.55040040*	0.50164	0.54855200*	0.50156	0.54934850*	0.50438	0.54657040*
log(Hours)	**0.33222**	***-0.00006370***	**0.32978**	***-0.00000853***	**0.33309**	***-0.00006620***	**0.33204**	***-0.00001680***
	0.30221	0.36187270*	0.30123	0.35885100*	0.30408	0.36198440*	0.30331	0.36184630*
Constant	**6.10502**	***-0.00312870***	**6.09994**	***-0.00336870***	**6.09461**	***-0.00361110***	**6.08765**	***0.00020800***
	5.95690	6.25628700*	5.95559	6.25185100*	5.95267	6.24959100*	5.94481	6.22661000*
Observations	38469		38469		38469		21389	
Replications	10000		10000		10000		10000	

95% Bias-Corrected Confidence Intervals: *significant at 5%

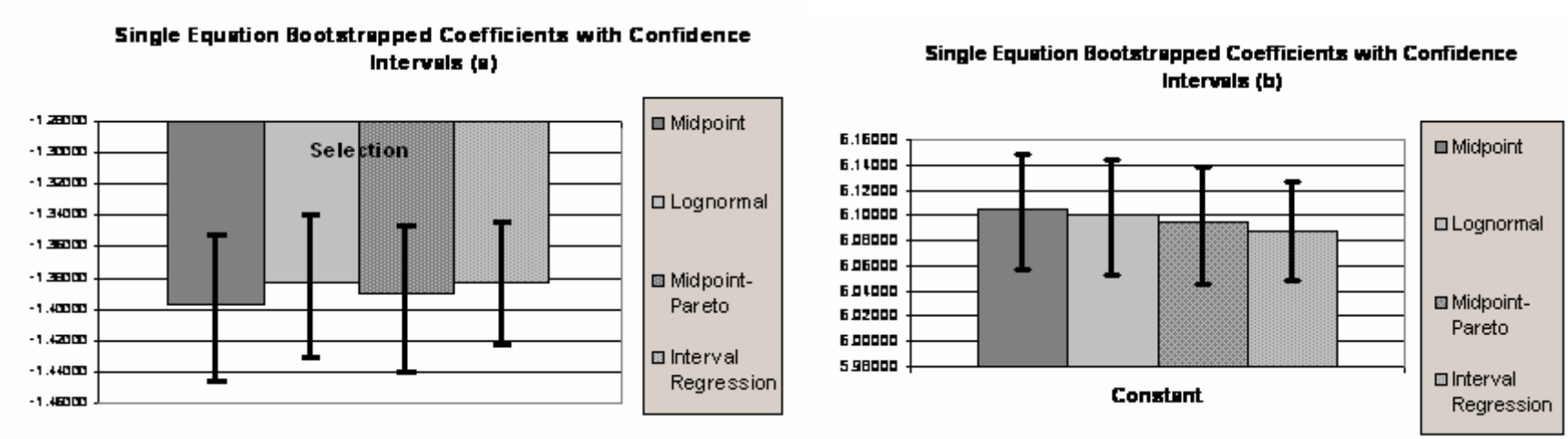
Coefficients: Observed with bias in italics

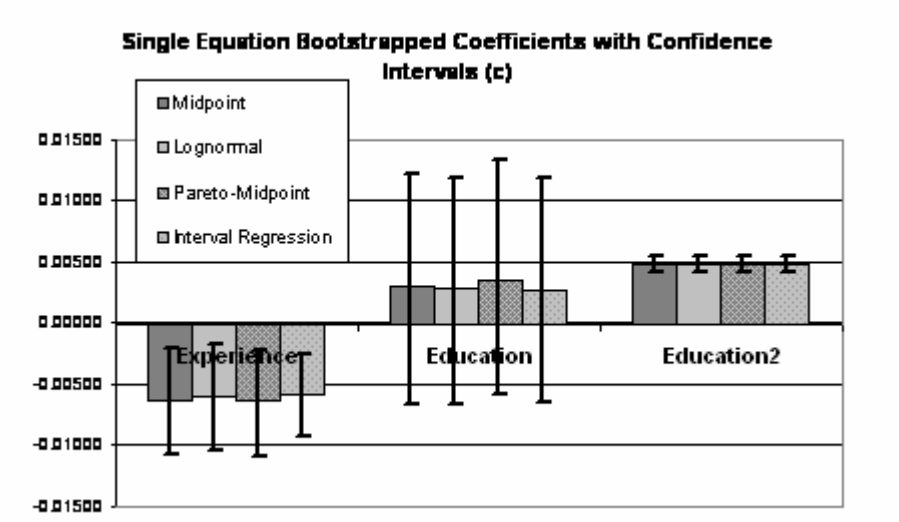
Heckit Interval Regression is Estimated Manually

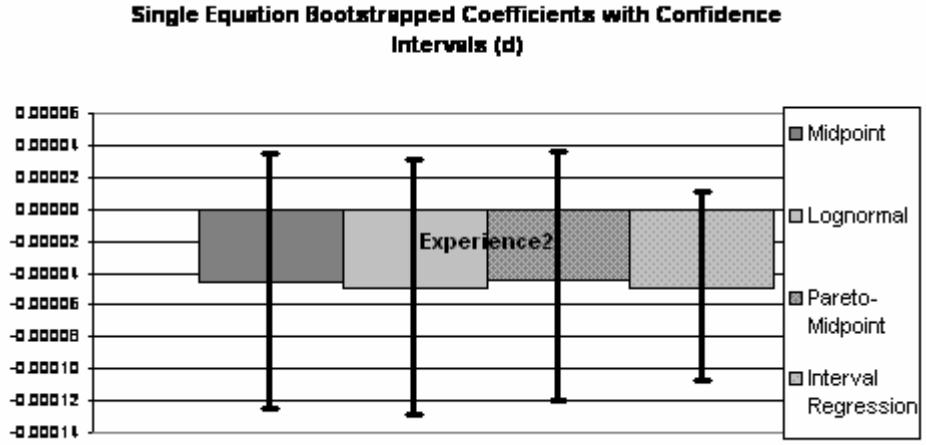
Table 6 Does Bootstrapped Confidence Interval Contain other methods' bootstrapped coefficients? - Single Equation

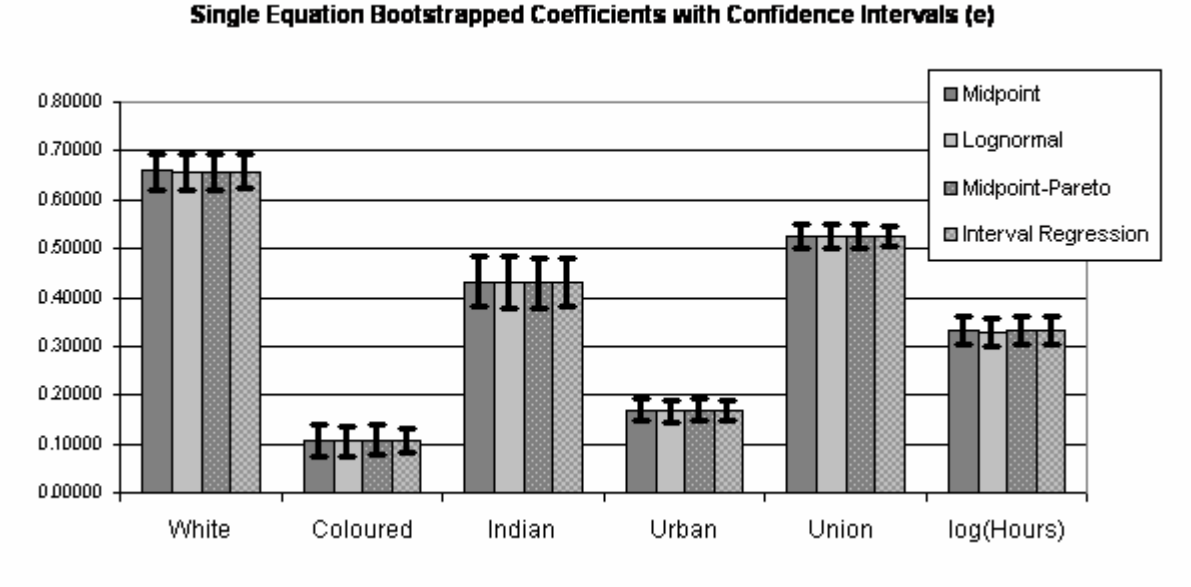
Conf. Interval	Interval Regression			Midpoint			Lognormal			Midpoint-Pareto		
Coefficient	Midpoint	Lognormal	Midpoint-Pareto	Interval Regression	Lognormal	Midpoint-Pareto	Interval Regression	Midpoint	Midpoint-Pareto	Interval Regression	Midpoint	Lognormal
Selection (λ)	√	√	√	√	√	√	√	√	√	√	√	√
Experience	√	√	√	√	√	√	√	√	√	√	√	√
Experience^2	√	√	√	√	√	√	√	√	√	√	√	√
Education	√	√	√	√	√	√	√	√	√	√	√	√
Education^2	√	√	√	√	√	√	√	√	√	√	√	√
White	√	√	√	√	√	√	√	√	√	√	√	√
Coloured	√	√	√	√	√	√	√	√	√	√	√	√
Indian	√	√	√	√	√	√	√	√	√	√	√	√
Urban	√	√	√	√	√	√	√	√	√	√	√	√
Union	√	√	√	√	√	√	√	√	√	√	√	√
log(Hours)	√	√	√	√	√	√	√	√	√	√	√	√
Constant	√	√	√	√	√	√	√	√	√	√	√	√

Figure 4 Comparison of Bootstrapped Coefficients and Confidence Intervals - Single Equation




Single Equation Bootstrapped Coefficients with Confidence Intervals (c)
Midpoint
Lognormal
Pareto-Midpoint
Interval Regression
0.01500
0.01000
0.00500
0.00000
-0.00500
-0.01000
-0.01500
Experience
Education
Education2


Single Equation Bootstrapped Coefficients with Confidence Intervals (d)
Experience2
Midpoint
Lognormal
Pareto-Midpoint
Interval Regression


Single Equation Bootstrapped Coefficients with Confidence Intervals (e)
0.80000
0.70000
0.60000
0.50000
0.40000
0.30000
0.20000
0.10000
0.00000
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression
White
Coloured
Indian
Urban
Union
log(Hours)

Female Equation

Table 7 Automated Heckit Estimates with Heckman Covariance Matrix – Female Equation

Female Equation	**Heckit Estimates with Heckman Covariance Matrix**		
	Midpoint	**Lognormal**	**Midpoint-Pareto**
Selection (λ)	**-1.382899**	**-1.365195**	**-1.375433**
	(-1.461246 - -1.304553)**	(-1.442538 - -1.287851)**	(-1.453356 - -1.29751)**
Experience	**-0.01533**	**-0.01473**	**-0.01525**
	(-0.02357 - -0.00710)**	(-0.02286 - -0.00661)**	(-0.02344 - -0.00707)**
Experience²	**-0.00005**	**-0.00005**	**-0.00005**
	(-0.00020 - 0.00010)	(-0.00020 - 0.00009)	(-0.00020 – 0.00010)
Education	**0.00225**	**0.0018**	**0.00297**
	(-0.01835 - 0.02285)	(-0.01853 - 0.02213)	(-0.01752 – 0.02345)
Education²	**0.00469**	**0.00473**	**0.00464**
	(0.00325 - 0.00614)**	(0.00330 - 0.00616)**	(0.00320 - 0.00608)**
White	**0.52893**	**0.52733**	**0.52702**
	(0.44256 - 0.61530)**	(0.44207 - 0.61260)**	(0.44112 - 0.61293)**
Coloured	**0.07995**	**0.07768**	**0.08139**
	(0.01336 - 0.14654)*	(0.01194 - 0.14341)*	(0.01517 - 0.14762)*
Indian	**0.39571**	**0.39495**	**0.39576**
	(0.25712 - 0.53430)**	(0.25813 - 0.53176)**	(0.25792 - 0.53361)**
Urban	**0.18138**	**0.17985**	**0.18164**
	(0.13477 - 0.22798)**	(0.13384 - 0.22586)**	(0.13528 - 0.22799)**
Union	**0.65421**	**0.65222**	**0.65403**
	(0.59697 - 0.71144)**	(0.59572 - 0.70872)**	(0.59711 - 0.71096)**
log(Hours)	**0.30446**	**0.30179**	**0.30478**
	(0.26272 - 0.34619)**	(0.26059 - 0.34299)**	(0.26327 - 0.34629)**
Constant	**6.33881**	**6.32929**	**6.3269**
	(6.09580 - 6.58182)**	(6.08939 - 6.56919)**	(6.08520 - 6.56860)**
Observations	19212	19212	19212

95% confidence intervals in parentheses

* significant at 5%; ** significant at 1%

Table 8 Manual Weighted Heckman 2-step with Robust Confidence Intervals - Female Equation

Female Equation	**Manual Weighted Heckman 2-step with Robust Confidence Intervals**			
	Midpoint	**Lognormal**	**Midpoint-Pareto**	**Interval Regression**
Selection (λ)	**-1.40195**	**-1.38915**	**-1.39409**	**-1.39007**
	(-1.44711 - -1.35678)**	(-1.43438 - -1.34392)**	(-1.43970 - -1.34847)**	(-1.43485 - -1.34530)**
Experience	**-0.00536**	**-0.00497**	**-0.0053**	**-0.00493**
	(-0.00939 - -0.00134)**	(-0.00899 - -0.00095)*	(-0.00934 - -0.00125)*	(-0.00892 - -0.00094)*
Experience²	**-0.00005**	**-0.00006**	**-0.00005**	**-0.00006**
	(-0.00013 - 0.00002)	(-0.00013 - 0.00001)	(-0.00013 - 0.00002)	(-0.00013 - 0.00001)
Education	**0.00041**	**0.00005**	**0.00091**	**0.0001**
	(-0.01083 - 0.01166)	(-0.01114 - 0.01125)	(-0.01038 - 0.01220)	(-0.01108 - 0.01129)
Education²	**0.00513**	**0.00515**	**0.00509**	**0.00517**
	(0.00437 - 0.00589)**	(0.00439 - 0.00591)**	(0.00433 – 0.00586)**	(0.00441 - 0.00593)**
White	**0.65156**	**0.65098**	**0.64968**	**0.6516**
	(0.61163 - 0.69149)**	(0.61068 - 0.69128)**	(0.60941 – 0.68995)**	(0.61180 - 0.69139)**
Coloured	**0.10754**	**0.10501**	**0.10775**	**0.10615**
	(0.07904 - 0.13604)**	(0.07672 - 0.13330)**	(0.07938 – 0.13612)**	(0.07801 - 0.13428)**
Indian	**0.38217**	**0.38115**	**0.38069**	**0.38241**
	(0.32938 - 0.43495)**	(0.32882 - 0.43347)**	(0.32826 – 0.43312)**	(0.33039 - 0.43444)**
Urban	**0.19482**	**0.19256**	**0.19477**	**0.19331**
	(0.16870 - 0.22095)**	(0.16665 - 0.21846)**	(0.16860 – 0.22093)**	(0.16749 - 0.21913)**
Union	**0.4755**	**0.47436**	**0.47541**	**0.4751**
	(0.45142 - 0.49958)**	(0.45035 - 0.49837)**	(0.45138 – 0.49945)**	(0.45119 - 0.49902)**
log(Hours)	**0.33348**	**0.33117**	**0.33442**	**0.33296**
	(0.30154 - 0.36542)**	(0.29964 - 0.36271)**	(0.30244 – 0.36641)**	(0.30123 - 0.36469)**
Constant	**6.10884**	**6.10509**	**6.09662**	**6.09585**
	(5.95212 - 6.26555)**	(5.94913 - 6.26106)**	(5.93879 – 6.25445)**	(5.94005 - 6.25165)**
Observations	21389	21389	21389	21389
R-Squared	0.64348	0.64418	0.64192	

Robust 95% confidence intervals in parentheses
* significant at 5%; ** significant at 1%

Heckit Interval Regression
Estimated Manually

Table 9 Bootstrapped Coefficients and Bias-Corrected Confidence Intervals - Female Equation

Female Equation	**Heckit with Bias-Corrected Bootstrapped Confidence Intervals**							
	Midpoint		**Lognormal**		**Midpoint-Pareto**		**Interval Regression**	
Selection (λ)	**-1.38290**	***0.00497420***	**-1.37543**	***0.00463520***	**-1.37543**	***0.00463520***	**-1.36658**	***0.00009750***
	-1.45562	-1.31971700*	-1.44881	-1.31243300*	-1.44881	-1.31243300*	-1.42543	-1.30664700*
Experience	**-0.01533**	***0.00023180***	**-0.01525**	***0.00017770***	**-0.01525**	***0.00017770***	**-0.01448**	***-0.00000490***
	-0.02232	-0.00910560*	-0.02221	-0.00895280*	-0.02221	-0.00895280*	-0.01954	-0.00932620*
Experience²	**-0.00005**	***-0.00000226***	**-0.00005**	***-0.00000126***	**-0.00005**	***-0.00000126***	**-0.00006**	***-0.00000008***
	-0.00016	0.00007860	-0.00017	0.00007380	-0.00017	0.00007380	-0.00015	0.00003240
Education	**0.00225**	***-0.00017450***	**0.00297**	***-0.00002160***	**0.00297**	***-0.00002160***	**0.00212**	***-0.00013810***
	-0.01142	0.01659620	-0.01157	0.01714570	-0.01157	0.01714570	-0.01166	0.01548420
Education²	**0.00469**	***0.00002370***	**0.00464**	***0.00001500***	**0.00464**	***0.00001500***	**0.00474**	***0.00000976***
	0.00370	0.00562200*	0.00368	0.00561290*	0.00368	0.00561290*	0.00379	0.00571150*
White	**0.52893**	***0.00199390***	**0.52702**	***0.00210210***	**0.52702**	***0.00210210***	**0.52820**	***0.00017110***
	0.47155	0.58382830*	0.46906	0.57941260*	0.46906	0.57941260*	0.47541	0.58163550*
Coloured	**0.07995**	***0.00124020***	**0.08139**	***0.00111680***	**0.08139**	***0.00111680***	**0.07834**	***-0.00027780***
	0.02969	0.12753430*	0.03195	0.12873040*	0.03195	0.12873040*	0.04078	0.11707000*
Indian	**0.39571**	***0.00111150***	**0.39576**	***0.00155250***	**0.39576**	***0.00155250***	**0.39385**	***0.00056500***
	0.31238	0.47745590*	0.31321	0.47697100*	0.31321	0.47697100*	0.31534	0.47182650*
Urban	**0.18138**	***0.00057890***	**0.18164**	***0.00041330***	**0.18164**	***0.00041330***	**0.18089**	***-0.00001040***
	0.14435	0.21571460*	0.14588	0.21634720*	0.14588	0.21634720*	0.14782	0.21423220*
Union	**0.65421**	***0.00101070***	**0.65403**	***0.00100190***	**0.65403**	***0.00100190***	**0.65298**	***0.00003890***
	0.61503	0.69212850*	0.61377	0.69218700*	0.61377	0.69218700*	0.61689	0.68803880*
log(Hours)	**0.30446**	***-0.00011930***	**0.30478**	***-0.00018850***	**0.30478**	***-0.00018850***	**0.30627**	***-0.00016530***
	0.26653	0.34091150*	0.26774	0.34171270*	0.26774	0.34171270*	0.26979	0.34337560*
Constant	**6.33881**	***-0.00848380***	**6.32690**	***-0.00806020***	**6.32690**	***-0.00806020***	**6.30680**	***0.00099480***
	6.14141	6.55294400*	6.12855	6.54574100*	6.12855	6.54574100*	6.10783	6.49363600*
Observations	19212		19212		19212		9454	
Replications	10000		10000		10000		10000	

95% Bias-Corrected Confidence Intervals: *significant at 5%

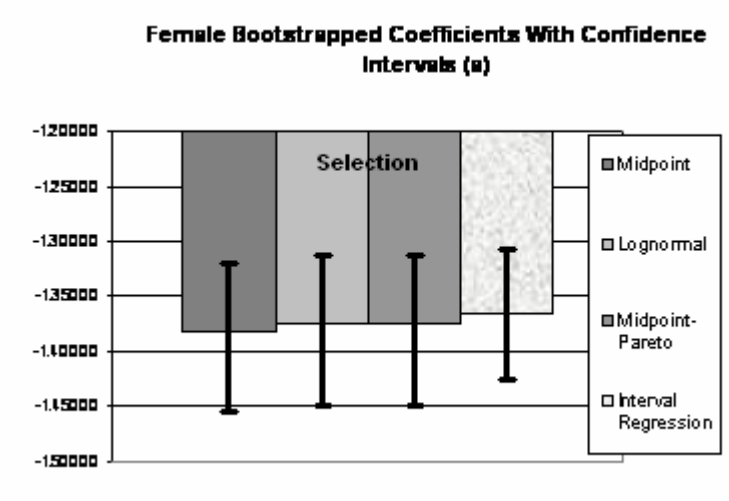
Coefficients: Observed with bias in italics

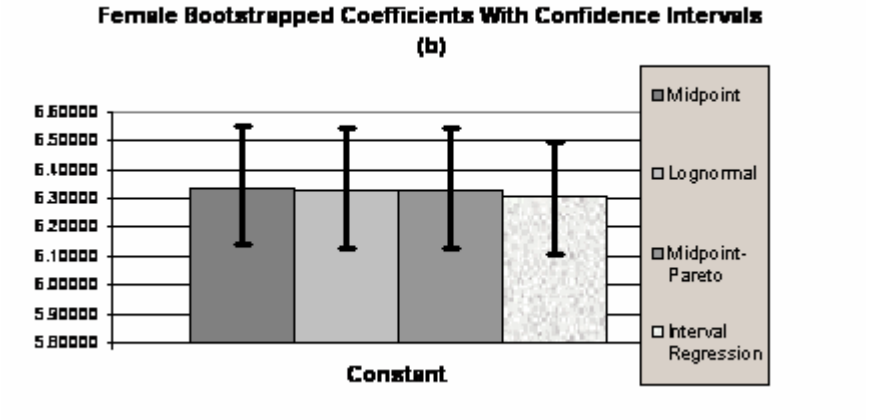
Heckit Interval Regression Estimated Manually

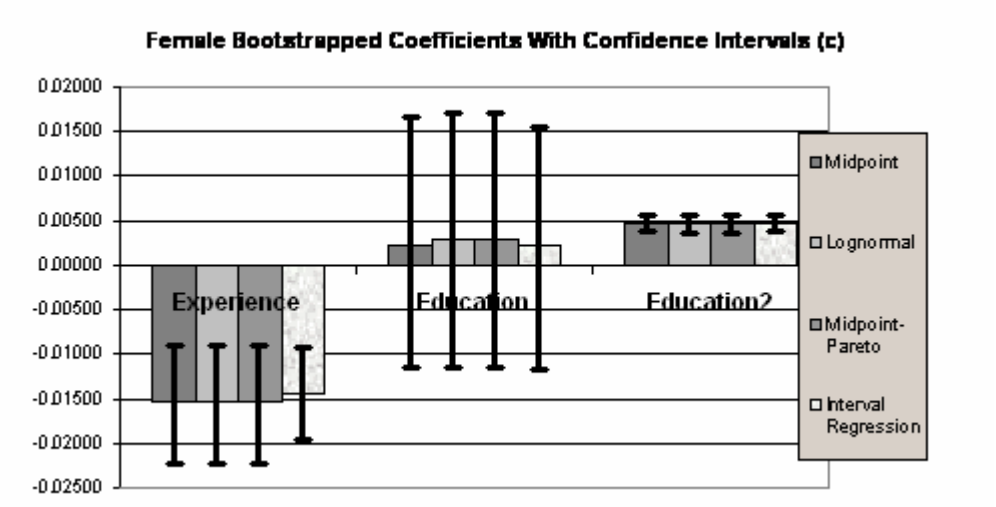
Table 10 Does Bootstrapped Confidence Interval Contain other methods' bootstrapped coefficients? Female Equation

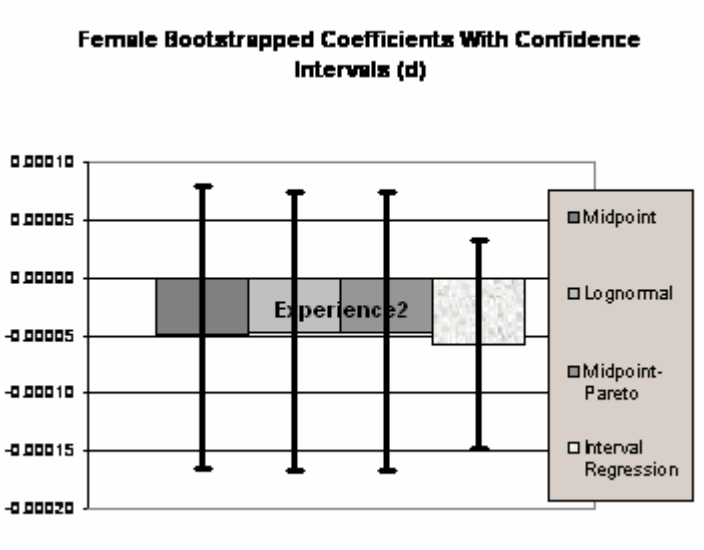
Conf Interval	Interval Regression			Midpoint			Lognormal			Midpoint-Pareto		
Coefficient	Midpoint	Lognormal	Midpoint-Pareto	Interval Regression	Lognormal	Midpoint-Pareto	Interval Regression	Midpoint	Midpoint-Pareto	Interval Regression	Midpoint	Lognormal
Selection (λ)	√	√	√	√	√	√	√	√	√	√	√	√
Experience	√	√	√	√	√	√	√	√	√	√	√	√
Experience²	√	√	√	√	√	√	√	√	√	√	√	√
Education	√	√	√	√	√	√	√	√	√	√	√	√
Education²	√	√	√	√	√	√	√	√	√	√	√	√
White	√	√	√	√	√	√	√	√	√	√	√	√
Coloured	√	√	√	√	√	√	√	√	√	√	√	√
Indian	√	√	√	√	√	√	√	√	√	√	√	√
Urban	√	√	√	√	√	√	√	√	√	√	√	√
Union	√	√	√	√	√	√	√	√	√	√	√	√
log(Hours)	√	√	√	√	√	√	√	√	√	√	√	√
Constant	√	√	√	√	√	√	√	√	√	√	√	√

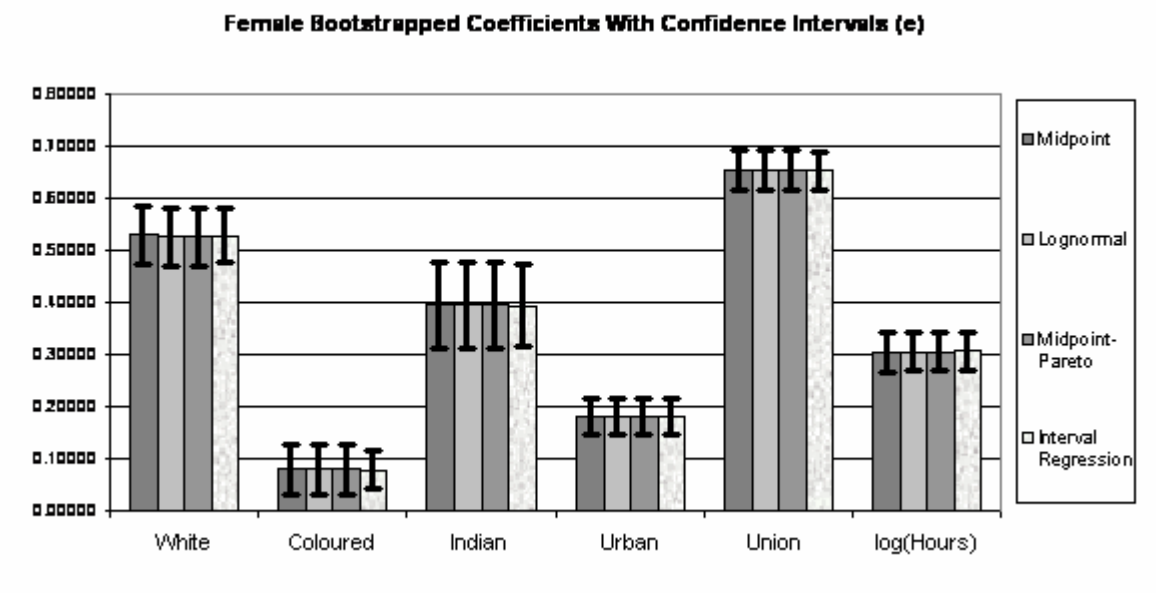
Figure 5 Comparison of Bootstrapped Coefficients and Confidence Intervals – Female Equation






Female Bootstrapped Coefficients With Confidence Intervals (c)
0.02000
0.01500
0.01000
0.00500
0.00000
-0.00500
-0.01000
-0.01500
-0.02000
-0.02500
Experience
Education
Education2
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression


Female Bootstrapped Coefficients With Confidence Intervals (d)
0.00010
0.00005
0.00000
-0.00005
-0.00010
-0.00015
-0.00020
Experience2
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression


Female Bootstrapped Coefficients With Confidence Intervals (e)
0.80000
0.70000
0.60000
0.50000
0.40000
0.30000
0.20000
0.10000
0.00000
White
Coloured
Indian
Urban
Union
log(Hours)
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression

Male Equation

Table 11 Automated Heckit Estimates with Heckman Covariance Matrix – Male Equation

Male Equation	**Heckit Estimates with Heckman Covariance Matrix**		
	Midpoint	**Lognormal**	**Midpoint-Pareto**
Selection (λ)	**-1.29163**	**-1.283029**	**-1.286555**
	(-1.357492 - -1.225767)**	(-1.348453 -1.217606)**	(-1.352159 -1.220951)**
Experience	**0.00749**	**0.00764**	**0.00744**
	(0.00033 - 0.01464)*	(0.00053 - 0.01474)*	(0.00032 - 0.01457)*
Experience²	**-0.00012**	**-0.00012**	**-0.00012**
	(-0.00025 - 0.00001)	(-0.00025 - 0.00001)	(-0.00025 - 0.00001)
Education	**0.00093**	**0.00099**	**0.00133**
	(-0.01782 - 0.01967)	(-0.01763 - 0.01961)	(-0.01734 - 0.02000)
Education²	**0.0054**	**0.00539**	**0.00537**
	(0.00409 - 0.00672)**	(0.00408 - 0.00670)**	(0.00406 - 0.00668)**
White	**0.79247**	**0.79064**	**0.78908**
	(0.71850 - 0.86645)**	(0.71716 - 0.86412)**	(0.71540 - 0.86276)**
Coloured	**0.1437**	**0.14137**	**0.14297**
	(0.08562 - 0.20177)**	(0.08368 - 0.19905)**	(0.08512 - 0.20081)**
Indian	**0.41882**	**0.41781**	**0.41723**
	(0.30965 - 0.52799)**	(0.30937 - 0.52625)**	(0.30849 - 0.52597)**
Urban	**0.20168**	**0.19989**	**0.20061**
	(0.15891 - 0.24444)**	(0.15741 - 0.24237)**	(0.15801 - 0.24321)**
Union	**0.40223**	**0.40108**	**0.40157**
	(0.35657 - 0.44789)**	(0.35572 - 0.44643)**	(0.35609 - 0.44706)**
log(Hours)	**0.19821**	**0.19682**	**0.19977**
	(0.14922 - 0.24720)**	(0.14815 - 0.24548)**	(0.15097 - 0.24857)**
Constant	**6.29691**	**6.29434**	**6.2878**
	(6.05840 - 6.53541)**	(6.05743 - 6.53126)**	(6.05024 - 6.52537)**
Observations	19257	19257	19257

95% confidence intervals in parentheses

* significant at 5%; ** significant at 1%

Table 12 Manual Weighted Heckman 2-step with Robust Confidence Intervals - Male Equation

Male Equation	**Manual Weighted Heckman 2-step with Robust Confidence Intervals**			
	Midpoint	**Lognormal**	**Midpoint-Pareto**	**Interval Regression**
Selection (λ)	**-1.29626**	**-1.28726**	**-1.29003**	**-1.28487**
	(-1.35607 - -1.23645)**	(-1.34698 - -1.22753)**	(-1.35025 - -1.22980)**	(-1.34396 - -1.22578)**
Experience	**0.00697**	**0.00717**	**0.00698**	**0.00716**
	(0.00167 - 0.01227)**	(0.00187 - 0.01247)**	(0.00165 - 0.01230)*	(0.00192 - 0.01241)**
Experience²	**-0.0001**	**-0.0001**	**-0.0001**	**-0.0001**
	(-0.00020 - -0.00000)*	(-0.00020 - -0.00001)*	(-0.00020 - -0.00000)*	(-0.00019 - -0.00001)*
Education	**-0.00276**	**-0.00292**	**-0.00262**	**-0.0029**
	(-0.01692 - 0.01141)	(-0.01710 - 0.01126)	(-0.01691 - 0.01167)	(-0.01703 - 0.01122)
Education²	**0.0057**	**0.00571**	**0.00569**	**0.00573**
	(0.00474 - 0.00667)**	(0.00474 - 0.00668)**	(0.00471 - 0.00667)**	(0.00477 - 0.00670)**
White	**0.78915**	**0.7884**	**0.78669**	**0.78861**
	(0.73852 - 0.83977)**	(0.73695 - 0.83985)**	(0.73524 - 0.83813)**	(0.73785 - 0.83938)**
Coloured	**0.15826**	**0.15511**	**0.15702**	**0.1556**
	(0.12167 - 0.19484)**	(0.11878 - 0.19145)**	(0.12068 - 0.19336)**	(0.11954 - 0.19167)**
Indian	**0.37119**	**0.36956**	**0.36808**	**0.37145**
	(0.30073 - 0.44166)**	(0.29987 - 0.43926)**	(0.29813 - 0.43803)**	(0.30246 - 0.44045)**
Urban	**0.21537**	**0.21374**	**0.21497**	**0.21502**
	(0.18310 - 0.24764)**	(0.18170 - 0.24579)**	(0.18264 - 0.24730)**	(0.18320 - 0.24683)**
Union	**0.36268**	**0.36181**	**0.36222**	**0.36244**
	(0.33269 - 0.39267)**	(0.33183 - 0.39179)**	(0.33220 - 0.39225)**	(0.33264 - 0.39225)**
log(Hours)	**0.21053**	**0.209**	**0.21212**	**0.20755**
	(0.15885 - 0.26221)**	(0.15799 - 0.26001)**	(0.16033 - 0.26391)**	(0.15667 - 0.25843)**
Constant	**6.28658**	**6.28405**	**6.27631**	**6.28496**
	(6.05311 - 6.52005)**	(6.05234 - 6.51577)**	(6.04135 - 6.51127)**	(6.05494 - 6.51499)**
Observations	11935	11935	11935	11935
R-Squared	0.63723	0.63716	0.63499	
	Robust 95% confidence intervals in parentheses * significant at 5%; ** significant at 1%			

Table 13 Bootstrapped Coefficients and Bias-Corrected Confidence Intervals - Male Equation

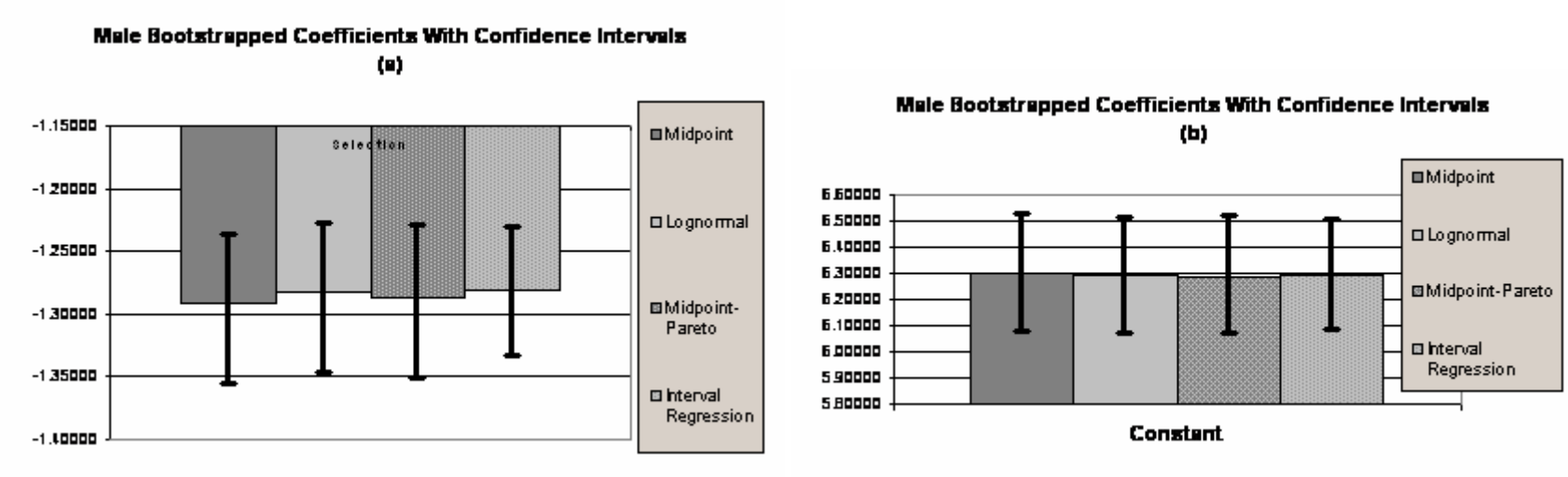
	Heckit with Bias-Corrected Bootstrapped Confidence Intervals							
Male Equation	Midpoint		Lognormal		Midpoint-Pareto		Interval Regression	
Selection (λ)	**-1.29163**	***0.00356110***	**-1.28303**	***0.00406230***	**-1.28656**	***0.00373480***	**-1.28025**	***0.00050860***
	-1.35583	-1.23578800*	-1.34647	-1.22736500*	-1.35052	-1.22941200*	-1.33263	-1.23107200*
Experience	**0.00749**	***0.00022050***	**0.00764**	***0.00018230***	**0.00744**	***0.00013360***	**0.00769**	***0.00003460***
	0.00169	0.01287170*	0.00181	0.01284620*	0.00163	0.01298550*	0.00324	0.01217600*
Experience²	**-0.00012**	***-0.00000328***	**-0.00012**	***-0.00000194***	**-0.00012**	***-0.00000136***	**-0.00012**	***-0.00000036***
	-0.00022	-0.00001550*	-0.00022	-0.00001570*	-0.00022	-0.00001360*	-0.00020	-0.00004120*
Education	**0.00093**	***-0.00006410***	**0.00099**	***-0.00009440***	**0.00133**	***0.00001850***	**0.00089**	***0.00003370***
	-0.01092	0.01300720	-0.01068	0.01333040	-0.01093	0.01314370	-0.01080	0.0123598
Education²	**0.00540**	***0.00001040***	**0.00539**	***0.00001570***	**0.00537**	***0.00000998***	**0.00543**	***-0.00000109***
	0.00455	0.00623050*	0.00453	0.00620620*	0.00451	0.00622140*	0.00462	0.00625170*
White	**0.79247**	***0.00150590***	**0.79064**	***0.00139130***	**0.78908**	***0.00133520***	**0.79044**	***0.00037000***
	0.74255	0.83738190*	0.74094	0.83604320*	0.74151	0.83614800*	0.74596	0.83618370*
Coloured	**0.14370**	***0.00060970***	**0.14137**	***0.00048790***	**0.14297**	***0.00073990***	**0.14144**	***0.00003190***
	0.10440	0.18147460*	0.10380	0.17825120*	0.10390	0.18134980*	0.10990	0.17276220*
Indian	**0.41882**	***0.00125180***	**0.41781**	***0.00009870***	**0.41723**	***0.00109250***	**0.41781**	***0.00081330***
	0.34981	0.48702490*	0.35016	0.48487370*	0.34938	0.48291720*	0.35245	0.48091700*
Urban	**0.20168**	***0.00023240***	**0.19989**	***0.00023170***	**0.20061**	***0.00017900***	**0.20212**	***0.00006240***
	0.17164	0.22995910*	0.17149	0.22873640*	0.17223	0.22901120*	0.17546	0.22801790*
Union	**0.40223**	***0.00078830***	**0.40108**	***0.00095750***	**0.40157**	***0.00072190***	**0.40186**	***0.00010990***
	0.37281	0.43085780*	0.37051	0.42987110*	0.37229	0.42954570*	0.37545	0.42865650*
log(Hours)	**0.19821**	***-0.00014240***	**0.19682**	***-0.00031680***	**0.19977**	***-0.00076130***	**0.19574**	***-0.00024290***
	0.15035	0.24649110*	0.14957	0.24430740*	0.15081	0.24688660*	0.14823	0.24264990*
Constant	**6.29691**	***-0.00544380***	**6.29434**	***-0.00502690***	**6.28780**	***-0.00277320***	**6.29316**	***-0.00024820***
	6.07589	6.52523200*	6.07372	6.51712200*	6.07014	6.52045400*	6.08686	6.50458500*
Observations	19257		19257		19257		11935	
Replications	10000		10000		10000		10000	

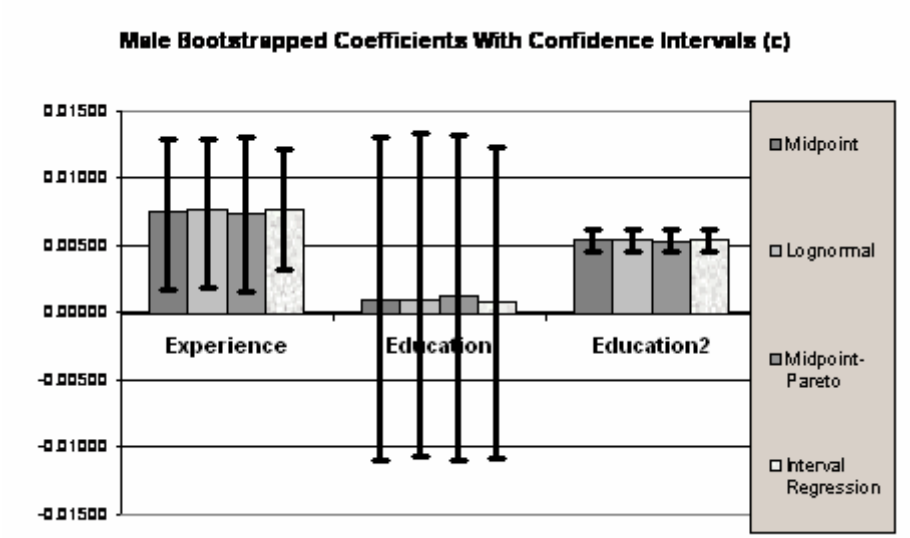
95% Bias-Corrected Confidence Intervals: *significant at 5%
Coefficients: Observed with bias in italics
Heckit Interval Regression Estimated Manually

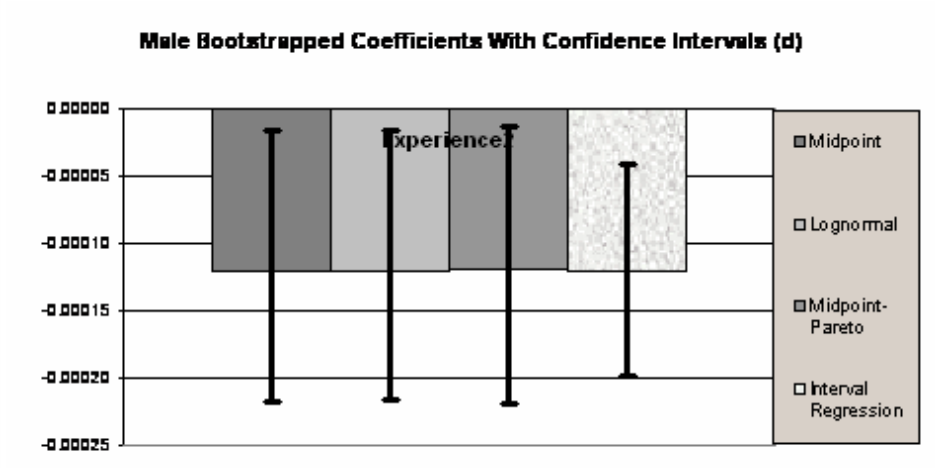
Table 14 Does Bootstrapped Confidence Interval Contain other methods' bootstrapped coefficients? Male Equation

Conf Interval	Interval Regression			Midpoint			Lognormal			Midpoint-Pareto		
Coefficient	Midpoint	Lognormal	Midpoint-Pareto	Interval Regression	Lognormal	Midpoint-Pareto	Interval Regression	Midpoint	Midpoint-Pareto	Interval Regression	Midpoint	Lognormal
Selection (λ)	√	√	√	√	√	√	√	√	√	√	√	√
Experience	√	√	√	√	√	√	√	√	√	√	√	√
Experience²	√	√	√	√	√	√	√	√	√	√	√	√
Education	√	√	√	√	√	√	√	√	√	√	√	√
Education²	√	√	√	√	√	√	√	√	√	√	√	√
White	√	√	√	√	√	√	√	√	√	√	√	√
Coloured	√	√	√	√	√	√	√	√	√	√	√	√
Indian	√	√	√	√	√	√	√	√	√	√	√	√
Urban	√	√	√	√	√	√	√	√	√	√	√	√
Union	√	√	√	√	√	√	√	√	√	√	√	√
log(Hours)	√	√	√	√	√	√	√	√	√	√	√	√
Constant	√	√	√	√	√	√	√	√	√	√	√	√

Figure 6 Comparison of Bootstrapped Coefficients and Confidence Intervals – Male Equation




Male Bootstrapped Coefficients With Confidence Intervals (c)
0.01500
0.01000
0.00500
0.00000
-0.00500
-0.01000
-0.01500
Experience
Education
Education2
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression


Male Bootstrapped Coefficients With Confidence Intervals (d)
0.00000
-0.00005
-0.00010
-0.00015
-0.00020
-0.00025
Experience2
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression

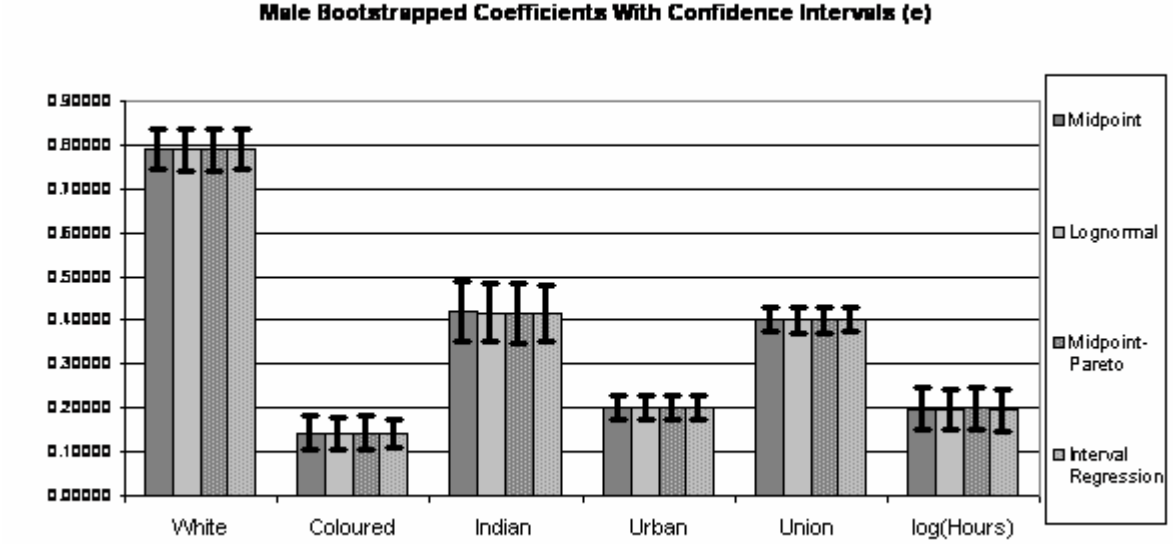

Male Bootstrapped Coefficients With Confidence Intervals (e)
0.90000
0.80000
0.70000
0.60000
0.50000
0.40000
0.30000
0.20000
0.10000
0.00000
White
Coloured
Indian
Urban
Union
log(Hours)
Midpoint
Lognormal
Midpoint-Pareto
Interval Regression

Table 15 Selection Probit Equations

Selection Equation	**Single**	**Female**	**Male**
	Broad Employment	*Broad Employment*	*Broad Employment*
Age	**0.08614**	**0.06987**	**0.1048**
	(0.07722 - 0.09506)**	(0.05676 - 0.08297)**	(0.09221 - 0.11739)**
Age²	**-0.0006**	**-0.00029**	**-0.00095**
	(-0.00072 - -0.00049)**	(-0.00047 - -0.00012)**	(-0.00111 - -0.00078)**
Eastern Cape	**-0.53382**	**-0.50822**	**-0.54569**
	(-0.59603 - -0.47161)**	(-0.59411 - -0.42232)**	(-0.63755 - -0.45382)**
Northern Cape	**-0.22348**	**-0.36838**	**-0.08031**
	(-0.30333 - -0.14362)**	(-0.48197 - -0.25480)**	(-0.19611 - 0.03549)
Free State	**-0.35492**	**-0.52041**	**-0.17465**
	(-0.42200 - -0.28784)**	(-0.61449 - -0.42633)**	(-0.27395 - -0.07535)**
KwaZulu-Natal	**-0.39592**	**-0.39192**	**-0.4163**
	(-0.45344 - -0.33839)**	(-0.47190 - -0.31193)**	(-0.50092 - -0.33168)**
North West	**-0.5278**	**-0.62029**	**-0.46557**
	(-0.59269 - -0.46290)**	(-0.71259 - -0.52800)**	(-0.55951 - -0.37163)**
Gauteng	**-0.58143**	**-0.60503**	**-0.58459**
	(-0.64123 - -0.52162)**	(-0.68902 - -0.52103)**	(-0.67199 - -0.49719)**
Mpumalanga	**-0.34949**	**-0.4007**	**-0.2754**
	(-0.41498 - -0.28401)**	(-0.49131 - -0.31010)**	(-0.37256 - -0.17825)**
Limpopo	**-0.61569**	**-0.64564**	**-0.56301**
	(-0.67982 - -0.55156)**	(-0.73220 - -0.55908)**	(-0.66122 - -0.46481)**
#Children <6	**-0.02623**	**-0.09231**	**0.0917**
	(-0.04612 - -0.00635)**	(-0.11824 - -0.06638)**	(0.05957 - 0.12382)**
#Males 16-59	**-0.09597**	**-0.12937**	**-0.17047**
	(-0.10973 - -0.08221)**	(-0.15037 - -0.10838)**	(-0.19213 - -0.14881)**
#Females 16-59	**-0.13195**	**-0.06721**	**-0.11139**
	(-0.14604 - -0.11786)**	(-0.08744 - -0.04699)**	(-0.13403 - -0.08875)**
#Adults >60	**-0.20539**	**-0.16163**	**-0.24322**
	(-0.23411 - -0.17667)**	(-0.20170 - -0.12156)**	(-0.28516 - -0.20129)**
pc Household Income	**0.00068**	**0.00055**	**0.00091**
	(0.00066 - 0.00070)**	(0.00053 - 0.00058)**	(0.00088 - 0.00095)**
(pc Household Income)²	**0**	**0**	**0**
	(-0.00000 - -0.00000)**	(-0.00000 - -0.00000)**	(-0.00000 - -0.00000)**
Constant	**-1.72501**	**-1.63259**	**-1.88493**
	(-1.89237 - -1.55765)**	(-1.87684 - -1.38833)**	(-2.12378 - -1.64608)**
Observations	**38469**	**19212**	**19257**

Heckman Corrected 95% confidence intervals in parentheses

* significant at 5%; ** significant at 1%

APPENDIX 2 – PARETO MEAN IMPUTATION

Following West (1986: 665):
First the conditional Pareto mean for the unbounded category is found:

$Y \equiv Earnings$

$Y \sim pareto(\alpha, k)$

$F_Y(y)$

$= P(Y \le y)$

$$= \begin{cases} 1-\left(\frac{k}{y}\right)^{\alpha} & \text{for y} \ge k \ge 0 \text{ and } \alpha > 0 \\ 0 & \text{for y<k} \end{cases}$$

where k is the lowest point for which Pareto tail is applicable and α is the shape parameter.

Let a be the lowerbound of the open category. Now:

$1-F_Y(y|Y \ge a)$ for y $\ge k \ge 0$ and $\alpha > 0$

$$= 1-\frac{P(Y \le y, Y \ge a)}{P(Y \ge a)}$$

$$= 1-\frac{P(Y \le y)}{P(Y \ge a)} \quad \text{because } a \ge k$$

$$= 1-\frac{P(Y \le y)}{1-P(Y \le a)}$$

$$= 1-\frac{F_Y(y)}{1-F_Y(a)}$$

$$= 1-\frac{1-\left(\frac{k}{y}\right)^{\alpha}}{\left(\frac{k}{a}\right)^{\alpha}} \quad \text{for y} \ge k \ge 0 \text{ and } \alpha > 0$$

$$\Rightarrow F_Y\left(y\middle|Y\geq a\right)$$

$$=\frac{1-\left(\frac{k}{y}\right)^{\alpha}}{\left(\frac{k}{a}\right)^{\alpha}}$$

$$=\frac{1}{\left(\frac{k}{a}\right)^{\alpha}}-y^{-\alpha}a^{\alpha}$$

$$\Rightarrow f_Y\left(y\middle|Y\geq a\right)$$

$$=\frac{\partial}{\partial y}\left[F_Y\left(y\middle|Y\geq a\right)\right]$$

$$=\alpha a^{\alpha}y^{-\alpha-1}\quad \text{for y}\geq k\geq 0 \text{ and } \alpha>0$$

Use the above and substitute the regression estimator of α, namely $\hat{\alpha}$, to obtain the conditional Pareto mean for the open category:

$$\overline{y}_{pareto|Y\geq a}$$

$$=\int_a^{\infty} y f_Y\left(y\middle|Y\geq a\right)dy$$

$$=\alpha a^{\hat{\alpha}}\lim_{t\to\infty}\int_a^t y^{-\hat{\alpha}}dy$$

$$=\hat{\alpha}a^{\hat{\alpha}}\lim_{t\to\infty}\frac{y^{-\hat{\alpha}+1}}{-\hat{\alpha}+1}\Bigg]_a^t$$

$$=\frac{\hat{\alpha}a^{\hat{\alpha}}}{-\hat{\alpha}+1}\lim_{t\to\infty}\left(t^{1-\hat{\alpha}}-a^{-\hat{\alpha}+1}\right)$$

$$=\frac{\hat{\alpha}a^{\hat{\alpha}}}{-\hat{\alpha}+1}\left(0-a^{-\hat{\alpha}+1}\right)\quad \text{if } \hat{\alpha}>1 \quad \ldots(1)$$

$$=\frac{\hat{\alpha}}{\hat{\alpha}-1}a\quad \text{for y}\geq k\geq 0 \text{ and } \hat{\alpha}>1$$

Now a mean is calculated for the bounded categories, following the same procedure, where *a* and *b* are the upper and lower bounds respectively of the concerned category:

$$
\begin{aligned}
&F_Y(y) \\
&= P(Y \le y) \\
&= \begin{cases} 1-\left(\frac{k}{y}\right)^{\alpha} & \text{for } k \le a \le y \le b \text{ and } \alpha > 1 \\ & \text{for the mean to be finite (see above)} \\ 0 & y < k \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
&F_Y(y \,|\, a \le Y \le b) \\
&= \frac{P(Y \le y,\ a \le Y \le b)}{P(a \le Y \le b)} \\
&= \frac{F_Y(y)}{F_Y(b) - F_Y(a)} \\
&= \frac{1-\left(\frac{k}{y}\right)^{\alpha}}{\left[1-\left(\frac{k}{b}\right)^{\alpha}\right] - \left[1-\left(\frac{k}{a}\right)^{\alpha}\right]} \\
&= \frac{k^{-\alpha} - y^{-\alpha}}{a^{-\alpha} - b^{-\alpha}} \quad \text{for } k \le a \le y \le b \text{ and } \alpha > 1
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow f_Y(y \,|\, a \le Y \le b) \\
&= \frac{\partial}{\partial y}\left[F_Y(y \,|\, a \le Y \le b)\right] \\
&= \frac{\alpha y^{-\alpha-1}}{a^{-\alpha} - b^{-\alpha}}
\end{aligned}
$$

Use the above and substitute the regression estimator of α, namely $\hat{\alpha}$, to obtain the conditional Pareto mean for the bounded categories:

$$\bar{y}_{pareto|a\le Y\le b}$$

$$= \int_a^b y\ f_Y\left(y \middle| a \le Y \le b\right) dy$$

$$= \frac{\hat{\alpha}}{a^{-\hat{\alpha}} - b^{-\hat{\alpha}}} \int_a^b y^{-\hat{\alpha}} dy$$

$$= \frac{\hat{\alpha}}{a^{-\hat{\alpha}} - b^{-\hat{\alpha}}} \frac{y^{-\hat{\alpha}+1}}{-\hat{\alpha}+1} \Bigg]_a^b$$

$$= \frac{\hat{\alpha}}{1-\hat{\alpha}} \frac{b^{-\hat{\alpha}+1} - a^{-\hat{\alpha}+1}}{a^{-\hat{\alpha}} - b^{-\hat{\alpha}}} \text{ for } k \le a \le y \le b \text{ and } \hat{\alpha} > 1$$

This formula is similar to that in Whiteford & McGrath (1994: 83), but the form calculated here produces correct results, with the imputed mean falling within the specified boundaries. It should be noted that $\hat{\alpha} > 1$ to obtain a finite mean for the unbounded category (see …(1) above, which necessitates the condition). It is therefore evident that the coefficient observed on *log(Y)* in the regression below is in fact the *negative* of the Pareto coefficient.

$$\log P = k - \alpha \log Y$$

The literature is not clear on this point, and results obtained differ slightly. However, when $\hat{\alpha}$ is used (without the negative sign), the imputed means which are obtained fall below the midpoint, as suggested by Seiver (1979: 230, 232). This also explains the suggestions by Whiteford & McGrath (1994: 83) and Gustavsson (2004:20) that the Pareto mean for the open category is in fact:

$$\bar{y}_{pareto} = \frac{\hat{\alpha}}{\hat{\alpha}+1} a$$

If we substitute -$\hat{\alpha}$ into this version of the formula:

$$\bar{y}_{pareto}$$

$$= \frac{-\hat{\alpha}}{-\hat{\alpha}+1} a$$

$$= \frac{\hat{\alpha}}{\hat{\alpha}-1} a$$

which yields the formula from the calculation above

APPENDIX 3 – LOGNORMAL MEAN IMPUTATION

A conditional mean of a normally distributed variable is found over a specified interval. Since the earnings variable is normally distributed following a log transformation, the original variable will assume a lognormal distribution.

$Y \equiv Earnings$

$X = \log(Y) \sim N(\mu; \sigma^2)$

a is the lowerbound of the category concerned

b is the upperbound of the category concerned

$$f_X(x \mid a < X < b)$$

$$= \frac{f_X(y, a < x < b)}{P(a < X < b)}$$

$$= \frac{f_X(y, a < x < b)}{\int_a^b f_X(x)\,dx}$$

$$E(x \mid a < X < b)$$

$$= \int_a^b x\, f_X(x \mid a < X < b)\, dx$$

$$= \int_a^b x \left(\frac{f_X(x)}{\int_a^b f_X(x)\, dx} \right) dx$$

$$= \frac{1}{\int_a^b f_X(x)\, dx} \int_a^b x\, f_X(x)\, dx$$

$$= \frac{1}{F_X(b) - F_X(a)} \int_a^b x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] dx \qquad \text{Let z=} \frac{x-\mu}{\sigma} \Rightarrow x = z\sigma + \mu \Rightarrow \frac{\partial x}{\partial z} = \sigma \Rightarrow \partial x = \sigma\, \partial z$$

$$= \frac{1}{\Phi(b^*) - \Phi(a^*)} \int_{a^*}^{b^*} (z\sigma + \mu) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z^2 \right] dz \qquad \textit{where } a^* = \frac{a-\mu}{\sigma} \quad \textit{and} \quad b^* = \frac{b-\mu}{\sigma}$$

$$= \frac{1}{\Phi(b^*) - \Phi(a^*)} \left\{ \mu \int_{a^*}^{b^*} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z^2 \right] dz + \sigma \int_{a^*}^{b^*} z \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z^2 \right] dz \right\} \qquad \textit{Let } w = -\frac{1}{2} z^2 \Rightarrow \frac{\partial w}{\partial z} = -z \Rightarrow \partial w = -z\, \partial z$$

$$= \frac{1}{\Phi(b^*) - \Phi(a^*)} \left\{ \mu \left[\Phi(b^*) - \Phi(a^*) \right] - \sigma \frac{1}{\sqrt{2\pi}} \int_{-\frac{1}{2}(a^*)^2}^{-\frac{1}{2}(b^*)^2} \exp[w]\, dw \right\}$$

$$= \mu - \frac{1}{\Phi(b^*) - \Phi(a^*)} \left\{ \sigma \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{1}{2} (b^*)^2 \right) - \exp\left(-\frac{1}{2} (a^*)^2 \right) \right] \right\}$$

$$= \mu - \frac{1}{\Phi(b^*) - \Phi(a^*)} \left\{ \sigma \left[\phi(b^*) - \phi(a^*) \right] \right\} \quad \text{because this is equivalent to the difference of two standard normal pdf's}$$

$$= \mu - \sigma \frac{\phi(b^*) - \phi(a^*)}{\Phi(b^*) - \Phi(a^*)}$$

$$= \mu - \sigma \frac{\phi\left(\frac{b-\mu}{\sigma} \right) - \phi\left(\frac{a-\mu}{\sigma} \right)}{\Phi\left(\frac{b-\mu}{\sigma} \right) - \Phi\left(\frac{a-\mu}{\sigma} \right)}$$

Once estimators $\hat{\mu}$ of μ and $\hat{\sigma}$ of σ are obtained by means of an interval regression with only a constant, the preceding population conditional mean can be used to impute conditional sample means to the brackets:

$$\overline{x}_{\text{normal}|a\le X\le b} = \hat{\mu} - \hat{\sigma}\frac{\phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - \phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}{\Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}$$

The untransformed Y now assumes lognormal imputation.

Hayashi (2000: 512) displays a specific case of this equation for $X > a$, which can be applied to the open ended top category:
This entails that $b \to \infty$
Therefore $\phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) \to 0$ and $\Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) \to 1$,
leaving us with

$\overline{x}_{\text{normal}|X>a}$

$$= \hat{\mu} - \hat{\sigma}\frac{0 - \phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}{1 - \Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}$$

$$= \hat{\mu} + \hat{\sigma}\frac{\phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}{1 - \Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}$$ which confirms Hayashi's form *where* $\frac{\phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}{1 - \Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right)}$ is the Inverse Mills Ratio

Similarly, for the lowest category:

a $\rightarrow$ -∞ as X results from a log transformation of Y.

Therefore $\phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right) \rightarrow 0$ and $\Phi\left(\frac{a-\hat{\mu}}{\hat{\sigma}}\right) \rightarrow 0$,

leaving us with

$$\overline{x}_{\text{normal}|X \le b}$$

$$= \hat{\mu} - \hat{\sigma}\frac{\phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - 0}{\Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right) - 0}$$

$$= \hat{\mu} - \hat{\sigma}\frac{\phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right)}{\Phi\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right)}$$

Similarly, the normal imputation of X results in a lognormal imputation of Y.

APPENDIX 4 – SUMMARY OF IMPUTATIONS USED

Table 16 Summary of Imputations Employed

Summary of Categories and Imputations Used				*Midpoint-Pareto Imputation*		*Lognormal Mean Imputation*
Earningsgroup	*Lowerbound*	*Upperbound*	*Midpoint*	*Male*	*Female*	
2	0	199	100	100	100	123.97
3	200	499	350	350	350	341.57
4	500	999	750	750	750	723.18
5	1,000	1,499	1250	1250	1250	1227.46
6	1,500	2,499	2000	2000	2000	1930.13
7	2,500	3,499	3000	3000	3000	2945.18
8	3500	4499	4000	4000	4000	3954.46
9	4500	5999	5250	5250	5250	5165.29
10	6000	7999	7000	6816.60	6790.290062	6877.39
11	8000	10999	9500	9196.55	9153.352468	9279.90
12	11000	15999	13500	12908.71772	12825.95305	13032.55
13	16000	29999	23000	20323.80838	19988.78293	20625.31
14	30000	∞	33000	46071.99396	42272.83494	43884.41
				α= 2.866601	**α= 3.444423**	**μ= 7.369894**
						σ= 1.177352

APPENDIX 5 - PRELIMINARY TESTING BY KERNEL DENSITY ESTIMATION

These methods follow Keswell and Poswell (2004: 854-855). Density plots of point data, data simulated according to a lognormal distribution (deemed to be the theoretical benchmark) and the imputed variables were drawn for comparative purposes. Two simulations were implemented: one for the point-reporting cohort, and another for the entire reporting population. Should the shapes of these curves differ substantially, the Data Generating Process (DGP) is considered to be different in each case.

First, a uniform random number generator was used to draw random *U(0,1)* samples for the respective cases. For each interval, a constant-only interval regression was run with the available log lower- and log upper bounds, as well as the respective logged point data, to establish the mean and standard deviation within each earnings group. This was executed in turn according to the ranges of both simulated variables mentioned above. A lognormal transformation was performed on the uniform random variables over each earnings bracket, with the respective associated means and standard deviations as follows:

$$X_{Lognormal} = \exp\left\{\hat{\mu} + \hat{\sigma} * \Phi^{-1}\left[U(0,1)\right]\right\}$$

where $\Phi^{-1}(\bullet)$ is the inverse standard normal distribution function.

The two simulated variables are now representative of theoretical DGP's of the point data by itself, and the joint point-interval data, according to the assumption that income is lognormally distributed. To each variable concerned, a Kernel-Density estimate was applied, in order to approximate a distribution for the random variables: this smoothes the histogram of the concerned variable via a non-negative symmetric weighting function (Rice, 1988: 325):

$w(x)$

$$w_h(x) = \frac{1}{h} w\left(\frac{x}{h}\right)$$

where $w_h(x)$ is a scaled variation of the initial $w(x)$. In this case, a Gaussian weight function was opted for: $w(x)$ is the standard normal density, which implies that $w_h(x)$ is a normal density with standard deviation h. This parameter represents the bandwidth of the smoother, and corresponds to the respective bin widths of the chosen histogram to be smoothed. As h is varied from a small magnitude to infinity, the obtained density changes from a rough (closer to the data) approximation, to a smoother, more drawn out density. The resultant estimated density f_h which approximates f, the true density in question, is estimated from the samples as follows:

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - X_i)$$

$X_1, X_2, \ldots, X_n$ is a random sample from f

This provides a graphical method to ascertain whether point data and interval-coded data have different underlying DGP's; proposed imputations can also be evaluated according to visual deviations.

APPENDIX 6 – MULTIVARIATE TESTING FRAMEWORK

Rigorous testing was implemented by multivariate methods. This takes into account the dependency structures between the different equations specified. Within this framework interval regressions cannot be compared to the imputations, as this option has not been developed within multivariate regression.

The model is set up as follows:

$$\begin{bmatrix} \underset{n\,x\,1}{\underline{Y}_{midpoint}} & \underset{n\,x\,1}{\underline{Y}_{pareto}} & \underset{n\,x\,1}{\underline{Y}_{\log normal}} \end{bmatrix}$$

$$= \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,11} & 1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,11} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,1} & \cdots & x_{n,11} & 1 \end{bmatrix} \begin{bmatrix} \text{Midpoint} & \text{Pareto} & \text{Lognormal} \\ \beta_{selection} & \beta_{selection} & \beta_{selection} \\ \beta_{experience} & \beta_{experience} & \beta_{experience} \\ \beta_{experience^2} & \beta_{experience^2} & \beta_{experience^2} \\ \beta_{education} & \beta_{education} & \beta_{education} \\ \beta_{education^2} & \beta_{education^2} & \beta_{education^2} \\ \beta_{white} & \beta_{white} & \beta_{white} \\ \beta_{coloured} & \beta_{coloured} & \beta_{coloured} \\ \beta_{indian} & \beta_{indian} & \beta_{indian} \\ \beta_{urban} & \beta_{urban} & \beta_{urban} \\ \beta_{union} & \beta_{union} & \beta_{union} \\ \beta_{\ln(hours)} & \beta_{\ln(hours)} & \beta_{\ln(hours)} \\ \beta_0 & \beta_0 & \beta_0 \end{bmatrix} + \begin{bmatrix} \underset{n\,x\,1}{\underline{\varepsilon}_{midpoint}} & \underset{n\,x\,1}{\underline{\varepsilon}_{pareto}} & \underset{n\,x\,1}{\underline{\varepsilon}_{\log normal}} \end{bmatrix}$$

$$\underset{n\,x\,3}{Y} = \underset{n\,x\,12}{X}\ \underset{12\,x\,3}{B} + \underset{n\,x\,3}{E}$$

This model is the multivariate regression. This too was estimated with bootstrap methods to obtain accurate covariance matrices, with 10000 repetitions. Consequently Wald Tests are performed (adapted from StataCorp, 2003: 238):

$$W = \left(\underset{qx12}{\underline{a}'} \underset{12x3}{B} \underset{3x1}{\underline{c}} - \underset{qx1}{\underline{0}} \right)' \left\{ Cov \left[\left(\underset{qx12}{\underline{a}'} \underset{12x3}{B} \underset{3x1}{\underline{c}} \right); \left(\underset{qx12}{\underline{a}'} \underset{12x3}{B} \underset{3x1}{\underline{c}} \right)' \right] \right\}^{-1} \left(\underset{qx12}{\underline{a}'} \underset{12x3}{B} \underset{3x1}{\underline{c}} - \underset{qx1}{\underline{0}} \right) \sim \chi^2_q$$

$$= \left(\underset{qx12}{\underline{a}'} \underset{12x3}{B} \underset{3x1}{\underline{c}} \right)' \left\{ \underset{qx12}{\underline{a}'} Cov \left[\left(\underset{12x3}{B} \underset{3x1}{\underline{c}} \right); \left(\underset{12x3}{B} \underset{3x1}{\underline{c}} \right)' \right] \underset{12xq}{\underline{a}} \right\}^{-1} \left(\underset{qx12}{\underline{a}'} \underset{12x3}{B} \underset{3x1}{\underline{c}} \right)$$

Here $\underline{a}'$ is a selection vector of 1's and 0's, which picks out the row (in other word the coefficients) of B to be tested. $\underline{c}'$ is the specific linear combination of equations to be tested: it

picks out the columns. For instance if one wishes to test whether $\beta_{selection\ midpoint} = \beta_{selection\ pareto}$ choose $\underline{a}' = [1,0,0,0,0,0,0,0,0,0,0,0]$ and $\underline{c}' = [1, -1, 0]$. It is now evident that the covariances between the coefficients are also taken into account in the test statistic, to accommodate for the presence of dependent structures. This combination is tested for equality to $\underline{0}$:qx1 and compared this to a $\chi^2(q)$ critical value, where q is the number of linear combinations being tested simultaneously. This framework allows for individual coefficients to be compared, and can be developed to compare entire equations simultaneously. However, it should be noted that the significance level requires adjustment when simultaneous hypotheses are being tested. This is due to the fact that the individual hypotheses themselves are not independent of each other. As such, a Bonferroni adjustment has been executed on the p-values, which takes into account the number of linear hypotheses being tested.

Table 17 Multivariate Tests - Single Equation

	Hypotheses ***(Equality of Individual Coefficients across equations, and entire equations)***								
Single Equation	**Midpoint = Midpoint-Pareto**			**Midpoint = Lognormal**			**Midpoint-Pareto= Lognormal**		
	χ^2	df	Prob>χ^2	χ^2	df	Prob>χ^2	χ^2	df	Prob>χ^2
Selection	65.27	1	0.00000*	151.12	1	0.00000*	67.31	1	0.00000*
Experience	0.19	1	1.00000	21.11	1	0.00010*	29.67	1	0.00000*
Experience^2	1.87	1	1.00000	5.19	1	0.27300	12.68	1	0.00440*
Education	5.61	1	0.21380	0.47	1	1.00000	13.08	1	0.00360*
Education^2	4.25	1	0.47110	0.26	1	1.00000	17.3	1	0.00040*
White	4.43	1	0.42330	2	1	1.00000	3.96	1	0.55970
Coloured	1.13	1	1.00000	25.61	1	0.00000*	44.65	1	0.00000*
Indian	0.63	1	1.00000	0.79	1	1.00000	0.03	1	1.00000
Urban	2.18	1	1.00000	9.77	1	0.02130*	6.63	1	0.12010
Union	0.68	1	1.00000	12.18	1	0.00580*	15.84	1	0.00080*
log(Hours)	8.26	1	0.04860*	7.43	1	0.07710	14.62	1	0.00160*
Constant	19.06	1	0.00020*	1.45	1	1.00000	2.03	1	1.00000
Joint	435.43	12	0.00000*	1443.58	12	0.00000*	834.4	12	0.00000*

Wald Tests, with Bonferroni adjusted p-values for dependent tests

*reject at a 5% level H_0: coefficient in first equation = coefficient in second equation

Table 18 Multivariate Tests - Female Equation

	Hypotheses ***(Equality of Individual Coefficients across equations, and entire equations)***								
Female Equation	**Midpoint = Midpoint-Pareto**			**Midpoint = Lognormal**			**Midpoint-Pareto= Lognormal**		
	χ^2	df	Prob>χ^2	χ^2	df	Prob>χ^2	χ^2	df	Prob>χ^2
Selection	43.82	1	0.00000*	103.73	1	0.00000*	60.1	1	0.00000*
Experience	1.12	1	1.00000	18.3	1	0.00020*	18.35	1	0.00020*
Experience^2	1.92	1	1.00000	3.94	1	0.56470	6.81	1	0.10850
Education	12.36	1	0.00530*	1.15	1	1.00000	9.49	1	0.02480*
Education^2	7.04	1	0.09580	1.25	1	1.00000	13.91	1	0.00230*
White	2.66	1	1.00000	1.42	1	1.00000	0.15	1	1.00000
Coloured	14.52	1	0.00170*	8.44	1	0.04410*	26.44	1	0.00000*
Indian	0	1	1.00000	0.59	1	1.00000	0.92	1	1.00000
Urban	1.72	1	1.00000	3.25	1	0.85720	4.53	1	0.40060
Union	0.19	1	1.00000	11.95	1	0.00660*	12.42	1	0.00510*
log(Hours)	1.95	1	1.00000	4.53	1	0.40040	5.78	1	0.19430
Constant	16.38	1	0.00060*	2.35	1	1.00000	0.19	1	1.00000
Joint	377.16	12	0.00000*	1190.75	12	0.00000*	399.66	12	0.00000*

Wald Tests, with Bonferroni adjusted p-values for dependent tests

*reject at a 5% level H_0: coefficient in first equation = coefficient in second equation

Table 19 Multivariate Tests - Male Equation

	Hypotheses ***(Equality of Individual Coefficients across equations, and entire equations)***								
Male Equation	**Midpoint = Midpoint-Pareto**			**Midpoint = Lognormal**			**Midpoint-Pareto= Lognormal**		
	χ^2	df	Prob>χ^2	χ^2	df	Prob>χ^2	χ^2	df	Prob>χ^2
Selection	26.44	1	0.00000*	46.49	1	0.00000*	13.39	1	0.00300*
Experience	0.19	1	1.00000	1.65	1	1.00000	6.5	1	0.12970
Experience^2	1.34	1	1.00000	0.06	1	1.00000	2.49	1	1.00000
Education	1.14	1	1.00000	0.03	1	1.00000	3.06	1	0.96490
Education^2	1.18	1	1.00000	0.21	1	1.00000	2.73	1	1.00000
White	3.3	1	0.82920	1.24	1	1.00000	8.22	1	0.04980*
Coloured	2.62	1	1.00000	17.8	1	0.00030*	14.11	1	0.00210*
Indian	0.84	1	1.00000	0.32	1	1.00000	0.3	1	1.00000
Urban	4.18	1	0.49090	8.23	1	0.04940*	2.49	1	1.00000
Union	1.28	1	1.00000	3.95	1	0.56400	2.14	1	1.00000
log(Hours)	6.27	1	0.14710	1.28	1	1.00000	6.85	1	0.10610
Constant	6.45	1	0.13280	0.21	1	1.00000	1.88	1	1.00000
Joint	199.46	12	0.00000*	746.72	12	0.00000*	519.46	12	0.00000*

Wald Tests, with Bonferroni adjusted p-values for dependent tests

*reject at a 5% level H_0: coefficient in first equation = coefficient in second equation

APPENDIX 7 – DESCRIPTIVE STATISTICS

Table 20 Descriptive Statistics

Sample for which Earnings Data are available

	ALL			MALE			FEMALE		
Variable	**Obs**	**Mean/Proportion**	**Std. Dev.**	**Obs**	**Mean/Proportion**	**Std. Dev.**	**Obs**	**Mean/Proportion**	**Std. Dev.**
Midpoint Earnings	22426	2804.721	3985.547	12462	3263.767	4532.675	9964	2230.592	3076.337
Lognormal Earnings	22426	2799.378	4198.092	12462	3266.093	4828.395	9964	2215.657	3145.887
Pareto-Midpoint Earnings	22426	2805.434	4234.651	12462	3277.108	4908.638	9964	2215.51	3098.417
Experience	22288	23.7098	12.45993	12378	23.6102	12.51723	9910	23.83421	12.3875
Experience Squared	22288	717.3975	667.2635	12378	714.1098	675.187	9910	721.504	657.2435
Education	22288	8.534458	3.726815	12378	8.48465	3.732897	9910	8.59667	3.718456
Education Squared	22288	86.7255	54.0523	12378	85.92269	54.0765	9910	87.72825	54.00803
Black	22426	0.693748	0.460946	12462	0.692586	0.461441	9964	0.695203	0.460345
White	22426	0.124811	0.330511	12462	0.125662	0.331481	9964	0.123746	0.329308
Coloured	22426	0.148488	0.355591	12462	0.145001	0.352116	9964	0.15285	0.359861
Indian	22426	0.032239	0.176639	12462	0.035949	0.186171	9964	0.027599	0.16383
Urban	22419	0.629645	0.482911	12455	0.611963	0.487323	9964	0.651746	0.476441
Union	22426	0.268082	0.44297	12462	0.305088	0.460463	9964	0.221799	0.415477
Hours	22341	44.88761	15.22701	12419	46.84661	14.3913	9922	42.4356	15.8761