

# DEALING WITH EARNINGS BRACKET RESPONSES IN HOUSEHOLD SURVEYS – HOW SHARP ARE MIDPOINT IMPUTATIONS?

DIETER VON FINTEL<sup>1</sup>

## *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 built into surveys intend to maintain response rates. Econometric tools to incorporate brackets vary from "simplistic" imputation to interval regressions. Coefficient differences are investigated here to establish reliable remedies. Monte-Carlo simulations suggest that "simple" methods fail only under extreme skewness and when a substantial number of right-censored observations appear in the sample. Testing procedures applied to LFS data reveal that in practice coefficients are virtually invariant to the proposed methods.

## 1) INTRODUCTION

The important characteristics of individuals' earnings have been investigated extensively by estimating the classical Mincerian earnings function using household survey data. Education and labour market experience contribute to human capital accumulation, which in turn has a non-linear impact on what individuals earn. Additional factors, such as racial and gender discrimination, union membership, geographic location and other demographic features, are often found to have an important influence on earnings. To monitor changes in the labour market, the effects of new policies, migration and other socioeconomic determinants over time, it is important to establish consistent techniques to estimate coefficients reliably. Are the returns to education increasing or decreasing over time, and which types of education are rewarded more favourably in different periods? Is labour market discrimination decreasing following the inception of affirmative action policies in South Africa? These questions can only be answered if data from successive surveys are used effectively and in a way which allows comparison, given the manner in which data was collected.

The quality of household survey data is constantly subject to the scrutiny of survey designers and analysts alike. Statistical agencies face many trade-offs in data collection: they must establish feasible, cost-effective collection methods, take into account how co-operative respondents might be and also consider the needs of end users. Labour market research would be a smooth task if respondents provided accurate and true information. Neither of these prerequisites is entirely satisfied by any survey. Earnings measurement is a case in point. Earnings questions in surveys cover sensitive subject matter and are met with much suspicion in the field, particularly in wealthy communities. Furthermore, respondents are not always in the position to offer precise information for fellow household members. These factors result in the distortion or refusal of earnings data. Given that a considerable section of labour market analysis rests on earnings functions, it is necessary to evaluate the extent to which imperfect data result in unrealistic conclusions about the economically active population.

To solve this problem, survey designers incorporate income bracket options into questionnaires. Hesitant respondents become more likely to offer a rough indication of earnings, and information on these individuals is no longer "lost" to the analyst. This is true when exact income cannot be ascertained, but an approximate category is known; it is also true for wealthier individuals who respond favourably to the higher degree of anonymity. 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.

---

<sup>1</sup> Department of Economics, University of Stellenbosch. E-mail: dieter2@sun.ac.za

Traditionally economists have been satisfied with applying category midpoints to earnings brackets, and using this imputed variable in further analysis. Is such a simple solution too rudimentary? Is this method justified, given that no distributional or statistical considerations are made? This study compares results using midpoint imputations to other methods which do have a distributional basis: conditional mean imputations from both the lognormal and Pareto distributions offer a parametric solution. It is furthermore possible to estimate interval regressions, which incorporates bracket earnings into a likelihood function: this method serves as a basis from which to establish credibility.

Which tools are the sharpest to convert “bad” data into reliable estimates? A Monte-Carlo simulation study tests in which scenarios the various methods compare favourably. This is followed by an analysis with Labour Force Survey Data. Coefficients of earnings functions are compared, first by various estimates’ confidence intervals and then by multivariate tests. To ensure that these tests have a sound base, it is necessary to account for the impure variance-covariance structures introduced when following the Heckman (1979) two-step procedure to correct for sample selection bias. Robust corrections are evaluated against bootstrapping. If resulting confidence intervals are too broad, too many values are regarded to be admissible; over-precision may similarly lead to *non*-rejection of invalid hypotheses (Brownstone & Valetta, 2001: 129).

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

## **2) WHY AND WHERE TO WITH EARNINGS BRACKETS?**

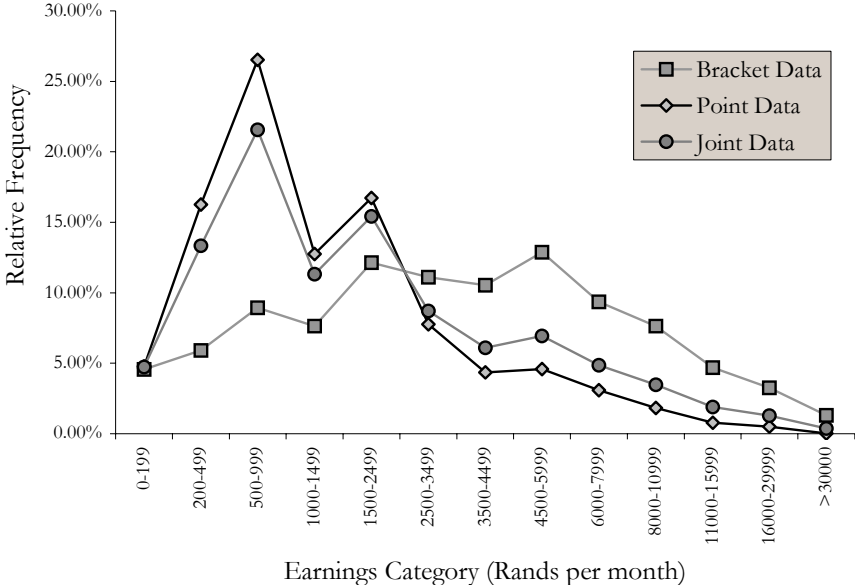
Unsatisfactory earnings response rates in household surveys stem from respondents’ unwillingness to disclose private financial details to survey enumerators (who are considered to be public sector employees) and from uncertainty regarding the positions of fellow household members (Posel & Casale, 2005: 1). Respondents with sporadic income inflows can also not necessarily attach an exactly representative amount to the earnings question. This point is relevant to the self-employed, informal sector employees and seasonal workers. Earnings brackets not only allow breadwinners some anonymity, but respondents may also provide a coarse estimate of fellow inhabitants’ earnings position, where precise knowledge is absent. The former benefit is particularly relevant if individuals earn excessively more than the lower bound of the open upper category: if respondents are offered the option to record monthly earnings of “above R30000” (which is the case in a typical South African household or labour force survey), those who earn R500000 and R50000 (say), would both be categorised in the same fashion. While it is clear that very different earnings positions are captured as “the same”, it is nonetheless true that the richer of the two hypothetical individuals would in all likelihood not have responded at all if prompted for exact details. Given this additional yet different form of information, an approximate position in the income distribution can be ascertained. But how are rough indicators used in precise estimates? Econometricians no longer have a continuous variable at their disposal, and can therefore not apply well-grounded techniques such as OLS.

The simplest remedy is to impute midpoints to each categorical response, and to choose a suitable value for the open category. While other authors (Posel & Casale, 2005: 21; Malherbe, 2007: 80) conclude that there is no indication that this method distorts summary poverty and inequality measures, this paper tests whether the same can be said for coefficients of earnings functions. Wooldridge (2002: 71-72) would have us believe that measurement error in the dependent variable does not accord bias to coefficients if it is uncorrelated with any of the regressors. In this sense, any imputation can be regarded as true earnings with noise. If this noise is uncorrelated with regressors, it is absorbed into the error term, and has no influence on the model. Should this be the case, “rudimentary” methods such as midpoint imputation should not be dismissed.

What does appear to be influential, however, is whether only point data or additional bracket information is incorporated into estimates. Characteristics of point and bracket reporting earners are found to be markedly different (Posel & Casale, 2005: 23): ignoring the one or the other results in information loss. Keswell and Poswell (2004: 855) show that the data generating process (DGP) of point respondents appears to be different to a simulated lognormal theoretical benchmark. Midpoint imputations also introduce substantial differences from the implied DGP, bar for the 1997 October Household Survey (OHS97). These results imply biased coefficient estimates based on the point reporting cohort, but also questions whether midpoint imputation is a satisfactory approximation. Point reporters do not adequately represent the population. The systematic, non-random decision which underlies bracket reporting can therefore not be ignored, and needs to be accommodated.

Which methods have previous studies used? 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. Hofmeyr (1999: 8) implements the midpoint method, without imputing a value to the open category, citing narrow brackets in defence. Rospabé (2002) and Daniels & Rospabé (2005) capitalise on the innovative “interval regression”, which implicitly imputes lognormally within a likelihood function. It is clear that the varying methods, as well as an apparent disregard for the potential bracket pitfalls could discredit past estimates and render them incomparable with other work. Should estimates, however, prove to be insensitive to methodology, such concerns will be mitigated.

Figure 1 Distribution of Bracket, Point and Joint Earnings (Source: LFS 2003b)



a) The Data Divide – LFS2003b

The importance of including earnings range questions in the Labour Force Survey (September 2003) is evident in Figure 1. While 63.9% of the employed sample provided point earnings, a further 28.1% of the sample responded within an income band. 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 offering only 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 reliability of regression 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 bracket-coded data.

Figure 1 illustrates how the inclusion of earnings brackets alters the relative frequency of income: the lower tail of the distribution of joint data is afforded less weight compared to point data, while the upper tail undergoes an upward adjustment to account for a reluctance to report exact amounts in well-off communities.

### 3) DEPENDENT VARIABLE VARIANTS

To satisfactorily include the representative sample of both point and bracket earners, it is imperative that a number of methods be implemented and exposed to testing procedures. Each is compared to the others to establish the most credible way forward.

#### *a) Generalised Tobit - Interval regression*

As a basis case, an interval regression is implemented. This is a generalised Tobit model and is estimated via pseudo-maximum likelihood procedures when sampling weights are brought into account. Daniels and Rospabé (2005: 31) present a log-likelihood function which provides for point, left-censored, right-censored (top income category with only a lower bound) and bracket-censored data. Point data is treated in the typical fashion and the procedure resembles OLS. Bracket information is incorporated by means of a cumulative normal distribution function. As a result, this procedure rests on the assumption of normality of logged earnings, and consequently lognormality of earnings. Supplementary bracket information is soundly incorporated into estimates, which broadens the base of research from only point observations. Further work is judged in light of this specification.

#### *b) 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. The focus of this study is to establish whether the former is a reliable point of departure.

This first method is conceptually simple and widely implemented by researchers. It is assumed that each person who offers an 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%. Imputed variables are then used with standard econometric techniques (such as OLS or Heckman estimates), as if they are continuous data. The pitfall of this method is its lack of theoretical backing (Whiteford & McGrath, 1994: 28). At the same time it may be attractive due to the limited knowledge of statistics required. The most fundamental question is whether this method is too rudimentary to apply in accurate earnings models.

Survey design and the size of brackets introduce sensitivity to estimation. In particular, the broad lowest category in the 1995 October Household Survey afforded too much weight to the upper portions of that bracket when traditional methods were followed (Keswell & Poswell, 2004: 855); other surveys broke bands down into smaller intervals, and midpoint estimates fare better in comparison. Furthermore, Seiver (1979: 230, 232) maintains that the true mean of any interval is always below its midpoint, and that distributional computations are influenced by the number of intervals chosen to span the range – fewer, wider brackets distort the picture.

Given that lower income categories are often 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 appears. Crato (2000:1239) emphasises the need to “model situations in which extreme values are observed with a relatively high probability” by use of heavy-tailed distributions such as the Pareto. Hence, this approach entails estimating conditional “Pareto Means” for categorical earnings in the upper tail, with midpoints maintained below that. This is the Pareto-Midpoint method.

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  $\alpha$  a shape parameter, which must be estimated. This can also be expressed in the log-linear form (Whiteford & McGrath, 1994: 81):

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

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

The above equation can be estimated by OLS on the point data in the sample to obtain an estimate of  $\alpha$ . 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 “agreed” 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 containing 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 to find the shape parameter is estimated with all categories initially included. Successively the lowest income band is excluded from the estimation. The equation with the highest  $R^2$  is deemed to contain the most reliable estimate of  $\alpha$ , 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 considered to be a suitable starting point to impute Pareto Means.

Crato (2000: 1251-1252) concludes that the regression estimator of  $\alpha$  has a smaller bias than the proposed Hill-Hall estimator in the presence of censoring. A modified version of the latter, however, has a smaller variance than the regression estimator: results are nonetheless 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 brackets.

Consequently, the Pareto means (conditional on the range of each applicable category) can be calculated as follows, where  $a$  and  $b$  are the lower and upper bound of the *bounded* category concerned and  $\hat{\alpha}$  is the regression estimate of the Pareto shape parameter (von Fintel, 2006: 59):

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

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

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

Gustavsson (2004: 20-21) implements a lognormal distribution over earnings data. This distribution also has a heavy tail, and justifies the assumption in its use to model earnings. When data is expressed in log form, a normal distribution is fitted, and as a result the untransformed data will be lognormally distributed. Maximum likelihood estimation on the log of earnings is standardly used to find the mean and variance of the distribution; these are used as normal distribution parameters to simulate the rest of the data. The introduction of censored and interval-coded data complicates 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).

An interval regression including only a constant successfully estimates the mean ( $\hat{\mu}$  – the constant) and standard error ( $\hat{\sigma}$  – the standard error of regression) of the log-transformed variable. Von Fintel (2006:

62) elaborates the imputation of *normal*<sup>2</sup> means to the intervals in *log* format by the following formula, where *a* and *b* are the *logged* lower and upper bounds of the earnings categories respectively:

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

#### 4) METHODOLOGICAL CONSIDERATIONS

Throughout, augmented Mincerian earnings functions are estimated. A parsimonious model is chosen, in accordance with knowledge from previous work. While the expected signs and the relative magnitude of coefficients are well-known for South Africa, 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. Sampling design is accounted for according to the weights of Statistics South Africa (2003a: 2-3).

##### a) *Sample Selection Bias*

Given the extent of both voluntary (where wages offered are below reservation wages) and apparent structural unemployment in the South African labour market, it is necessary to test for sample selection bias. Heckman (1979: 153-154), in his seminal article, outlines that population estimates based on non-randomly selected samples are mis-specified. An Inverse Mills Ratio ( $\lambda$ ) is calculated for each individual from preliminary estimates of an employment probit. This is a function of the probability that each observation is included in the sample (Heckman, 1979: 156).

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. Wooldridge (2002: 564) indicates that the test for significant selection bias is  $H_0: \beta_\lambda=0$  within the earnings specification. Under this hypothesis, the standard assumption of homoskedasticity holds, while the presence of significant sample selection bias causes it to be violated. This leads to the next point of concern.

##### b) *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 is implemented, though this is marred by heteroskedasticity once sample selection is a statistically valid concern. The first correction applies robust standard errors, according to the Huber-White robust covariance matrix (Hill et al., 2003: 5).

Wooldridge (2002: 564), however, sounds the warning that robust standard errors may nonetheless be misleading, as  $\beta_\lambda$  is itself the coefficient of an estimated stochastic quantity. Hill et al. (2003: 4-12, 18) evaluate the adequacy of implementing various asymptotic variance-covariance matrices and bootstrapping with Monte-Carlo simulation studies. For small samples and considerable censoring, none of the variants perform well. For large samples, bootstrap estimates offered the best solution. Since LFS survey data constitute large samples, this course of action is considered 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. Hence, the parent population is approximated as the number

---

<sup>2</sup> As a result, the untransformed variable is lognormally imputed

of repetitions is increased. Sprent, in Henderson (2005: 3), claims: “The more vague the supposition about population distributions, the more useful the bootstrap becomes.” Brown (2000: 437) advocates the use of bootstrap for cases where asymptotic variance is impossible or difficult to calculate. Distributions of parameters are considered closer to the true population approximation than limiting distributions: consequently confidence intervals (which are used below to present intuitive results) are unaffected by impure standard errors<sup>3</sup>.

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, confidence intervals cannot be constructed with pivotal statistics, standard errors and tabled percentiles. It is therefore necessary to increase the bootstrap iterations 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.

## 5) VARIABLES AND SPECIFICATION

The standard Mincerian earnings function (Mincer, 1974: 130) 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.

### a) *Earnings*

Monthly earnings are used in this analysis. 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. The sample is restricted to those typically assumed to be in the labour force, namely workers between the ages of 16 and 64.

### b) *Returns to experience and education – human capital investment*

The positive influence of labour market experience on earnings has been acknowledged in early estimates. Potential experience<sup>4</sup> is used here, yet with caution. Mincer (1974: 129-130) himself warns against the difficulties of approximating the variable in this fashion: in particular, females and the chronically unemployed do not necessarily accrue labour market experience during all their time out of education. Indeed, once sample selection is accounted for, the inclusion of females in the sample causes experience coefficients to assume theoretically inconsistent signs for the survey under consideration (von Fintel, 2006: 27). Since this study does not attempt to directly establish the determinants of earnings in South Africa, these difficulties are noted. Coefficients are tested for stability across methods, and not used to confirm or refute their theoretical bases. Results presented here therefore focus on a male-only sample. Similar conclusions with respect to the central hypothesis apply to joint samples and female estimates, despite theoretically inconsistent signs in some cases (von Fintel, 2006: 28-29).

Van der Berg and Burger (2003: 496) commence a study on educational inequalities by posing the question whether the intended rectification of socioeconomic disparities in South Africa is indeed being achieved via the education system. The returns to education in South Africa have been found to be non-linear, which may have substantial implications for individuals pursuing different types of education. Keswell and Poswell (2004: 844), who show that when controlling for potential experience, the returns to education are positive for the first 12 years. The additional positive quadratic term causes predicted income to rise more sharply following this attainment.

---

<sup>3</sup> For the mechanics of bootstrap regression and the construction of bootstrapped confidence intervals, see Brownstone & Valetta (2001:131)

<sup>4</sup> *Potential Experience = Age – Years of Education - 6*

### c) Other Variables

A number of additional variables are included to capture “non-investment” (in terms of human capital) features of South African earnings. South Africa’s historical context dictates that race-specific discrimination still persists in South Africa (Rospabé, 2002; Burger & Jafta, 2006). Here, the black cohort is chosen as a reference group, with relative estimates obtained for whites, Indians and coloureds. A dummy is included to investigate the escalating wage premium associated with unionisation found by Hofmeyr (1999) and Hofmeyr & Lucas (2001). Bhorat & Leibbrandt (2001: 124) use the urban-rural dummy to distinguish between geographic differences, rather than establishing provincial effects.

The selection equation includes household and demographic variables. Specification follows Chamberlain & van der Berg (2002). Variables included are age, age squared, a set of provincial dummies, the number of infants in the individual’s household, the number of working age males and females in the household, the number of adults above age 60 in the household and household income. 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

### a) Simulation Evidence

A Monte-Carlo simulation study precedes the analysis of real data to delineate various factors which influence the robustness of the various methods. The investigation probes the stability of these results by considering a number of models subject to different limitations. The dimensions considered are sample size, the number of regressors, the underlying skewness of the earnings distribution and the size of the brackets applied to the variables. To design this experiment, the spectrum of these characteristics had to be considered.

A number of South African studies that employed earnings functions<sup>6</sup> were surveyed to gather information regarding sample sizes ( $n$ ) and the number of variables ( $p$ ) typically implemented. These ranges lead to the choice of 100, 500, 1000, 2000, 5000, 15000 and 30000 for  $n$  combined with the choice of 3 (a standard Mincerian earnings function), 5, 7, 10, 20 and 40 for  $p$ . Nationally representative household surveys were used to establish bounds for the moments of typical South African earnings (see Table 3). These surveys’ bracket structures were also used to categorise the dependent variable in the simulations.

A benchmark simulation with 1000 repetitions was run with each of the combinations of  $n$  and  $p$  according to the following data-generating process:

$$x_i \sim N(0;1) \quad i=1 \cdots p$$

$$\beta_i = \sqrt{0.5/p} \quad i=1 \cdots p$$

$$e \sim N(0;1)$$

$$\Rightarrow \log(y^*) = \left( 7.2 + \sum_i x_i' \beta_i + e \right) \sim N(7.2; 1.5)$$

This delivers the desired distributional characteristics of  $\log(\text{Earnings})$ , with a mean of 7.2 which is typical for the LFS’s and an implied skewness of 12.094. The continuous  $\log(y^*)$  was artificially categorised according to the bounds in LFS questionnaires. This new categorical variable was subjected to interval regressions, as well as OLS with a midpoint imputation<sup>8</sup>, the results of which were compared to the known population values.

---

<sup>6</sup> Magruder & Natrass, 2006; Natrass & Walker, 2005; Daniels & Rospabé, 2005; Keswell & Poswell, 2002; Rospabé, 2002; Posel & Casale, 2005.

<sup>8</sup> The open category was imputed with a value of R33000, 10% above the lowerbound of R30000.

Table 1 Means and Standard Deviations of  $\log(\text{Earnings})$  and implied Skewness of Earnings<sup>9</sup>: 1995-2005

	$\mu$	Mean Earnings $=\exp(\mu)$	$\sigma$	Implied Lognormal Skewness
OHS95	7.183	1316.25	1.080	7.757
OHS96	7.147	1269.97	1.163	9.941
OHS97	7.241	1395.85	1.122	8.780
OHS98	7.223	1371.25	1.171	10.173
OHS99	7.142	1264.39	1.296	15.367
LFS2000a	7.037	1137.66	1.244	12.901
LFS2000b	7.219	1364.48	1.228	12.219
LFS2001a	7.095	1206.27	1.237	12.601
LFS2001b	7.241	1396.16	1.206	11.371
LFS2002a	7.236	1388.92	1.214	11.712
LFS2002b	7.297	1475.57	1.230	12.323
LFS2003a	7.314	1501.45	1.222	11.981
LFS2003b	7.398	1632.12	1.202	11.224
LFS2004a	7.430	1685.70	1.205	11.329
LFS2004b	7.463	1742.43	1.161	9.868
LFS2005a	7.476	1766.00	1.184	10.597

**Error! Not a valid bookmark self-reference.** shows the simulation results for each of the sample sizes, including only 3 regressors. The first apparent result is that the midpoint imputation leads to a general overestimation of the mean, compared to a slight downward bias for the other regressors. Interval regressions perform better. Benchmarking the size of each test against 0.05, it is evident that the interval regression performs remarkably well. The midpoint imputation results are satisfactory for small sample sizes (though the tests for the constants generally have a poor size), but then show a marked deterioration for sample sizes of 5000 and above. A possible reason for this is that a substantial number of observations fall in the open category when the sample size is increased. The arbitrary value of R33000 assigned to this category appears to

capture this quantity poorly. This exercise was repeated by imputing a conditional lognormal mean to this category (not shown), without delivering any improvements. A first consideration in applying midpoints (or any imputation for that matter) is therefore the number of right-censored values present in the earnings variable. This problem may be less of a concern if most individuals in the unbounded category provide point data there or if most individuals fall within bounded brackets, where reflective midpoints are known. These conclusions concerning the midpoint imputation hold true for each of the other simulations run with different number of parameters and are omitted here to conserve space.

Table 3 repeats the simulation with  $p=10$ . It is first evident that the same pattern appears for the midpoint imputation as before. What is important, however, is the pattern which develops for interval regressions, and continues as more regressors are added (not shown here). For  $n=100$ , the p-values start deviating substantially from 0.05, while this is generally not true for  $n>100$ . This feature also holds true for the midpoint imputation, but only once 40 regressors are included. It is therefore evident that when the sample size is restrictively small, degrees of freedom sensitivity is problematic with either solution.

The entire exercise was repeated, but the dimension of broader brackets was explored. To achieve this, successive LFS brackets were merged to form new wider brackets. The simulations were re-run. Results (not shown) reveal that interval regressions showed similar patterns as before, with reliable estimates throughout (bar for small sample sizes combined with many regressors). The midpoint estimates, however, deteriorated substantially, with most of the test sizes deviating from 0.05. This reveals that interval regressions appear to be the more robust method if sample sizes are significantly large. Sufficiently narrow brackets are essential for reliable coefficient estimates when using midpoint imputation, given the moments of earnings prevalent in the specific survey.

The aspect of extreme skewness was also investigated. The implied DGP was adjusted to reflect a skewness of 15.809, which reflects the scenario of OHS99 more realistically (see Table 3). The results (not shown) highlight that no method is robust to extreme skewness, with the size of each test substantially different from the expected 0.05. The reason for this may lie yet again with the larger numbers of individuals in the open category. Should surveys start containing many right-censored earnings values, it becomes advisable to append another category to the questionnaire.

<sup>9</sup> An interval regression with only a constant was run for each survey. The constant signifies the mean of  $\log(\text{Earnings})$ , while the standard error of regression is taken to be the standard deviation of  $\log(\text{Earnings})$ . The skewness parameter of *Earnings* is calculated assuming a lognormal distribution, and is based on the estimate of  $\sigma$  rather than a population value (Pouloukas, 2004: 158):

$$Skewness = \left( e^{\sigma^2} + 2 \right) \sqrt{e^{\sigma^2} - 1}$$

In summary, a number of concerns emerge: first, should a large number of datapoints be right-censored, any method should be considered with caution; secondly, if brackets are “too wide”, midpoint imputation distorts inference, compared to the sustained reliability of interval regressions; thirdly, extreme underlying skewness cannot be overcome by any known method. If these concerns are suspected to hinder inference for a specific survey, midpoints should be used with caution, even if the bracket structure is known to typically provide satisfactory results. It is, however, important to establish whether these peculiarities result from observed social phenomena or whether data collection in that specific survey can be trusted (see Burger & Yu, 2006 for an exposition of the potential problems with earnings variables in South African household surveys). Table 3 shows that particularly for the LFS era, skewness of *Earnings* has remained relatively stable. The validation of one survey therefore contributes largely to the credibility of using midpoints and other imputations in these surveys. Some concerns with the characteristics of the OHS surveys (in particular OHS99 – see Table 3), however, remind researchers to not apply midpoints blindly.

Table 2 Monte-Carlo Simulations:  $p=3$ ; mean of  $\log(\text{Earnings}) = 7.2$ ; Skewness of Earnings =12.094; replications = 1000

Population Values	$c = 7.2$ $\beta = 0.408$	Interval Regression		Midpoints	
		Coefficient	$p$ -value <sup>10</sup>	Coefficient	$p$ -value
100	constant	7.197	0.057	7.215	0.052
	x1	0.414	0.063	0.407	0.059
	x2	0.410	0.056	0.404	0.055
	x3	0.410	0.059	0.404	0.062
500	constant	7.200	0.052	7.218	0.067
	x1	0.407	0.065	0.401	0.069
	x2	0.409	0.040	0.404	0.047
	x3	0.406	0.055	0.401	0.054
1000	constant	7.200	0.054	7.218	0.086
	x1	0.407	0.063	0.402	0.057
	x2	0.408	0.053	0.403	0.058
	x3	0.408	0.043	0.403	0.053
2000	constant	7.200	0.060	7.217	0.122
	x1	0.409	0.035	0.404	0.052
	x2	0.408	0.046	0.403	0.054
	x3	0.408	0.053	0.403	0.060
5000	constant	7.200	0.039	7.217	0.226
	x1	0.408	0.057	0.403	0.077
	x2	0.408	0.043	0.403	0.076
	x3	0.408	0.054	0.403	0.063
15000	constant	7.200	0.051	7.217	0.553
	x1	0.409	0.048	0.403	0.096
	x2	0.408	0.057	0.403	0.108
	x3	0.408	0.047	0.403	0.101
30000	constant	7.200	0.067	7.218	0.848
	x1	0.408	0.039	0.403	0.159
	x2	0.408	0.049	0.403	0.165
	x3	0.408	0.048	0.403	0.153

<sup>10</sup> The proportion of the 1000 replications for which the  $T$  statistic exceeds the 97.5<sup>th</sup> percentile of the applicable  $t$  distribution. This value should therefore be benchmarked against the theoretical size of 0.05 for the two-sided test of  $H_0 : \hat{\beta} = \beta$ .

**Table 3a Monte-Carlo Simulations:  $p=10$ ; mean of  $\log(\text{Earnings}) = 7.2$ ; Skewness of Earnings =12.094; replications = 1000**

Population Values	$c=7.2$ $\beta=0.224$	Interval Regression		Midpoints		$n$		Interval Regression		Midpoints	
		Coefficient	p-value	Coefficient	p-value			Coefficient	p-value	Coefficient	p-value
100	constant	7.196	0.059	7.215	0.047	1000	constant	7.200	0.055	7.217	0.084
	x1	0.228	0.072	0.224	0.058		x1	0.224	0.040	0.221	0.038
	x2	0.226	0.060	0.223	0.050		x2	0.224	0.048	0.221	0.049
	x3	0.227	0.059	0.223	0.050		x3	0.222	0.051	0.219	0.055
	x4	0.228	0.069	0.225	0.056		x4	0.224	0.044	0.221	0.037
	x5	0.226	0.061	0.223	0.056		x5	0.224	0.054	0.222	0.054
	x6	0.228	0.069	0.224	0.057		x6	0.224	0.059	0.221	0.056
	x7	0.224	0.064	0.221	0.055		x7	0.224	0.045	0.221	0.038
	x8	0.225	0.072	0.222	0.058		x8	0.225	0.048	0.222	0.052
	x9	0.224	0.068	0.221	0.050		x9	0.224	0.049	0.221	0.043
	x10	0.222	0.055	0.219	0.045		x10	0.224	0.046	0.221	0.041
500	constant	7.200	0.047	7.218	0.065	2000	constant	7.201	0.051	7.218	0.126
	x1	0.222	0.051	0.219	0.051		x1	0.224	0.062	0.221	0.058
	x2	0.221	0.055	0.218	0.054		x2	0.224	0.041	0.221	0.046
	x3	0.222	0.058	0.219	0.055		x3	0.226	0.045	0.223	0.046
	x4	0.223	0.047	0.220	0.046		x4	0.224	0.042	0.221	0.039
	x5	0.225	0.053	0.222	0.052		x5	0.223	0.037	0.220	0.040
	x6	0.224	0.052	0.221	0.048		x6	0.224	0.051	0.221	0.052
	x7	0.226	0.048	0.223	0.045		x7	0.224	0.059	0.221	0.060
	x8	0.223	0.061	0.221	0.058		x8	0.222	0.071	0.220	0.078
	x9	0.225	0.051	0.222	0.049		x9	0.224	0.043	0.222	0.042
	x10	0.224	0.043	0.222	0.044		x10	0.224	0.048	0.221	0.051

**Table 3b Monte-Carlo Simulations:  $p=10$ ; mean of  $\log(\text{Earnings}) = 7.2$ ; Skewness of Earnings =12.094; replications = 1000**

$n$		Interval Regression		Midpoints		$n$		Interval Regression		Midpoints	
		Coefficient	p-value	Coefficient	p-value			Coefficient	p-value	Coefficient	p-value
5000	constant	7.200	0.047	7.218	0.232	30000	constant	7.200	0.051	7.218	0.860
	x1	0.224	0.046	0.221	0.050		x1	0.224	0.050	0.221	0.083
	x2	0.224	0.046	0.221	0.052		x2	0.224	0.052	0.221	0.091
	x3	0.224	0.053	0.221	0.058		x3	0.224	0.051	0.221	0.068
	x4	0.223	0.040	0.220	0.049		x4	0.224	0.047	0.221	0.062
	x5	0.223	0.052	0.220	0.058		x5	0.224	0.034	0.221	0.071
	x6	0.224	0.048	0.221	0.054		x6	0.224	0.050	0.221	0.073
	x7	0.224	0.042	0.221	0.052		x7	0.223	0.043	0.220	0.078
	x8	0.224	0.041	0.221	0.045		x8	0.224	0.059	0.221	0.078
	x9	0.224	0.048	0.221	0.046		x9	0.224	0.055	0.221	0.083
	x10	0.224	0.044	0.221	0.056		x10	0.223	0.051	0.221	0.091
15000	constant	7.200	0.038	7.218	0.553						
	x1	0.224	0.056	0.221	0.073						
	x2	0.224	0.041	0.221	0.071						
	x3	0.224	0.046	0.221	0.065						
	x4	0.223	0.054	0.221	0.062						
	x5	0.224	0.053	0.221	0.058						
	x6	0.224	0.052	0.221	0.066						
	x7	0.224	0.036	0.221	0.055						
	x8	0.224	0.045	0.221	0.062						
	x9	0.224	0.045	0.221	0.058						

	x10	0.224	0.055	0.221	0.073
--	-----	-------	-------	-------	-------

*b) Analysis of Labour Force Survey (September 2003) Data*

It is imperative to test whether survey brackets in South Africa are suitably narrow with parametric comparisons of real data. Bracket structures have remained unchanged since OHS1997, which entails that conclusions presented here apply for all of these surveys. This assumes that skewness and right-censored datapoints are of no concern. Should methods be dissimilar, it is evident that an interval regression remains the most suitable econometric tool to prevent misleading judgments.

How does this survey compare to the hypothetical scenarios presented above? Single equation estimates use 21389 earnings observations, while this can be split into male and female models, with 11935 and 9454 observations respectively. This tally borders on the range where midpoint imputation could be hindered by skewness and right-censored values. Skewness, however, is moderate (see Table 3). 80 earnings values are right-censored, of which 68 are male values and 12 female. This is moderate compared to the tally in excess of 100 present in a typical round of the simulation with this sample size.

*i) A brief word on the coefficients*

While this study is focussed on parameter comparisons, a short exposition of their magnitudes is called for. Discussion is limited to male estimates with bootstrapped confidence intervals (see Table 4). The inclusion of females in the sample distorts economic interpretation due to the potential experience variable, as discussed above. A full set of results for single and female equations are presented in von Fintel (2006).

First, the sample selection correction term is significant in all cases. This underlines that earnings and employment processes are intertwined, and that coefficients would be otherwise biased.

The convexity of returns to education is confirmed in this context, with both the linear and quadratic terms 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.

The most interesting feature of the coefficients is that the returns to Mincerian variables are 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. Ineffective labour markets allocate more reward to non-productive activity than to skills development: union membership (for instance) has higher returns than an additional year of education.

*ii) Method Comparison by Confidence Intervals*

It is useful to first consider intuitive evidence of parameter equality: do respective coefficients fall within the 95% bootstrapped confidence intervals of the other methods? The overwhelming result is that this is true for each coefficient in male, female and single samples, independent of imputation or methodology. This in itself provides a strong indication that each tool is as sharp as the other. 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. This necessitates formal testing.

*Table 4 Heckman 2-Step with Bootstrapped Confidence Intervals (Male Sample)*

<i>Dependent Variable: log(Earnings) following imputation</i>	Midpoint	Lognormal	Midpoint-Pareto	Interval Regression
---	----------	-----------	-----------------	---------------------

<b>Inverse Mills Ratio</b> ( $\hat{\theta}$ )	<b>-1.292</b> (-1.356; -1.236) *	<b>-1.283</b> (-1.346; -1.227) *	<b>-1.287</b> (-1.351; -1.229) *	<b>-1.280</b> (-1.333; -1.231) *
<b>Experience</b>	<b>0.008</b> (0.002; 0.013) *	<b>0.008</b> (0.002; 0.013) *	<b>0.007</b> (0.002; 0.013) *	<b>0.008</b> (0.003; 0.012) *
<b>Experience<sup>2</sup></b>	<b>-0.0001</b> (-0.0002; 0.0000) *	<b>-0.0001</b> (-0.0002; 0.0000) *	<b>-0.0001</b> (-0.0002; 0.0000) *	<b>-0.0001</b> (-0.0002; 0.0000) *
<b>Education</b>	<b>0.001</b> (-0.011; 0.013)	<b>0.001</b> (-0.011; 0.013)	<b>0.001</b> (-0.011; 0.013)	<b>0.001</b> (-0.011; 0.012)
<b>Education<sup>2</sup></b>	<b>0.005</b> (0.005; 0.006) *	<b>0.005</b> (0.005; 0.006) *	<b>0.005</b> (0.005; 0.006) *	<b>0.005</b> (0.005; 0.006) *
<b>White</b>	<b>0.793</b> (0.743; 0.837) *	<b>0.791</b> (0.741; 0.836) *	<b>0.789</b> (0.742; 0.836) *	<b>0.790</b> (0.746; 0.836) *
<b>Coloured</b>	<b>0.144</b> (0.104; 0.181) *	<b>0.141</b> (0.104; 0.178) *	<b>0.143</b> (0.104; 0.181) *	<b>0.141</b> (0.110; 0.173) *
<b>Indian</b>	<b>0.419</b> (0.350; 0.487) *	<b>0.418</b> (0.350; 0.485) *	<b>0.417</b> (0.349; 0.483) *	<b>0.418</b> (0.352; 0.481) *
<b>Urban</b>	<b>0.202</b> (0.172; 0.230) *	<b>0.200</b> (0.171; 0.229) *	<b>0.201</b> (0.172; 0.229) *	<b>0.202</b> (0.175; 0.228) *
<b>Union</b>	<b>0.402</b> (0.373; 0.431) *	<b>0.401</b> (0.371; 0.430) *	<b>0.402</b> (0.372; 0.430) *	<b>0.402</b> (0.375; 0.429) *
<b>log(Hours)</b>	<b>0.198</b> (0.150; 0.246) *	<b>0.197</b> (0.150; 0.244) *	<b>0.200</b> (0.151; 0.247) *	<b>0.196</b> (0.148; 0.243) *
<b>Constant</b>	<b>6.297</b> (6.076; 6.525) *	<b>6.294</b> (6.074; 6.517) *	<b>6.288</b> (6.070; 6.520) *	<b>6.293</b> (6.087; 6.505) *
Observations	19257	19257	19257	11935
Bootstrap Replications	10000	10000	10000	10000

95% Bias-Corrected Confidence Intervals in parentheses: \*significant at 5% level

Note that the number of observations differ for the interval regression – this equation was estimated by a manual two-step, and the final number of observations includes only employed individuals. Other equations are estimated by pre-programmed Heckman estimators, which consider both the employed and the unemployed in its observation tally.

### iii) Multivariate Testing Framework

A more rigorous multivariate testing procedure to assess the intuitive results obtained above is necessary. 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 and 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 executed on p-values to account for the dependence of simultaneously tested hypotheses (see for instance Johnson & Wichern, 2002: 232). Results are summarised in Table 5 for males.

It is evident that no estimated equation is in its entirety equivalent to another. 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 1% level of significance. This can be ascribed to the stochastic nature of the Inverse Mills Ratio. The equality of magnitudes is, however, not the emphasis in this case, as its economic interpretation is limited. The comparison of the Midpoint and Midpoint-Pareto method highlights no further differences.

Elsewhere (von Fintel, 2006: 28-30) it is shown that imputation strategies should be carefully considered. If Pareto shape coefficients are estimated separately for each gender, estimates for the joint sample based on the imputed variable appears to be incomparable to those based on other imputations. Male estimates appear to be stable, in contrast to those of females and the whole sample. This raises the question whether the Pareto coefficient for females is reliably determined in this case. It is particularly problematic when the gender-specific imputations are combined to form a “mixed distribution” for the entire sample, while the midpoint imputation does not distinguish a gender effect.

Table 5 Multivariate Tests of Coefficient Equality

Male Equation	Hypotheses (Equality of Individual Coefficients across equations, and entire equations)
---------------	--

	Midpoint = Midpoint-Pareto			Midpoint = Lognormal			Midpoint-Pareto= Lognormal		
	Chi-Squared	df	p-value	Chi-Squared	df	p-value	Chi-Squared	df	p-value
Inverse Mills Ratio (̂)	26.44	1	0.000 **	46.49	1	0.000 **	13.39	1	0.003 **
Experience	0.19	1	1.000	1.65	1	1.000	6.5	1	0.130
Experience <sup>2</sup>	1.34	1	1.000	0.06	1	1.000	2.49	1	1.000
Education	1.14	1	1.000	0.03	1	1.000	3.06	1	0.965
Education <sup>2</sup>	1.18	1	1.000	0.21	1	1.000	2.73	1	1.000
White	3.3	1	0.829	1.24	1	1.000	8.22	1	0.0498 *
Coloured	2.62	1	1.000	17.8	1	0.000 **	14.11	1	0.002 **
Indian	0.84	1	1.000	0.32	1	1.000	0.3	1	1.000
Urban	4.18	1	0.491	8.23	1	0.049 *	2.49	1	1.000
Union	1.28	1	1.000	3.95	1	0.564	2.14	1	1.000
log(Hours)	6.27	1	0.147	1.28	1	1.000	6.85	1	0.106
Constant	6.45	1	0.133	0.21	1	1.000	1.88	1	1.000
Whole Vector	199.46	12	0.000 **	746.72	12	0.000 **	519.46	12	0.000 **

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

\* reject at a 5% level of significance  $H_0$ : coefficient in first equation = coefficient in second equation

\*\*reject at a 1% level of significance  $H_0$ : coefficient in first equation = coefficient in second equation

The comparisons of regressions with Pareto-Midpoint and Lognormal imputations as dependent variables reveal that only the coefficients of the *Inverse Mills Ratio*, *White* and *Coloured* reject the hypothesis that  $\beta_{pareto} = \beta_{lognormal}$  at a 5% level of significance for males. Fewer differences occur when Midpoint estimates are compared to the Lognormal imputation (in particular only the *Inverse Mills Ratio*, *Coloured* and *Urban* differ at a 5% level of significance for males). Overall, traditional Mincerian variables can be modelled with confidence by any method: some other coefficients might fail rigorous statistical tests, though the intuitive results suggest that they are economically similar.

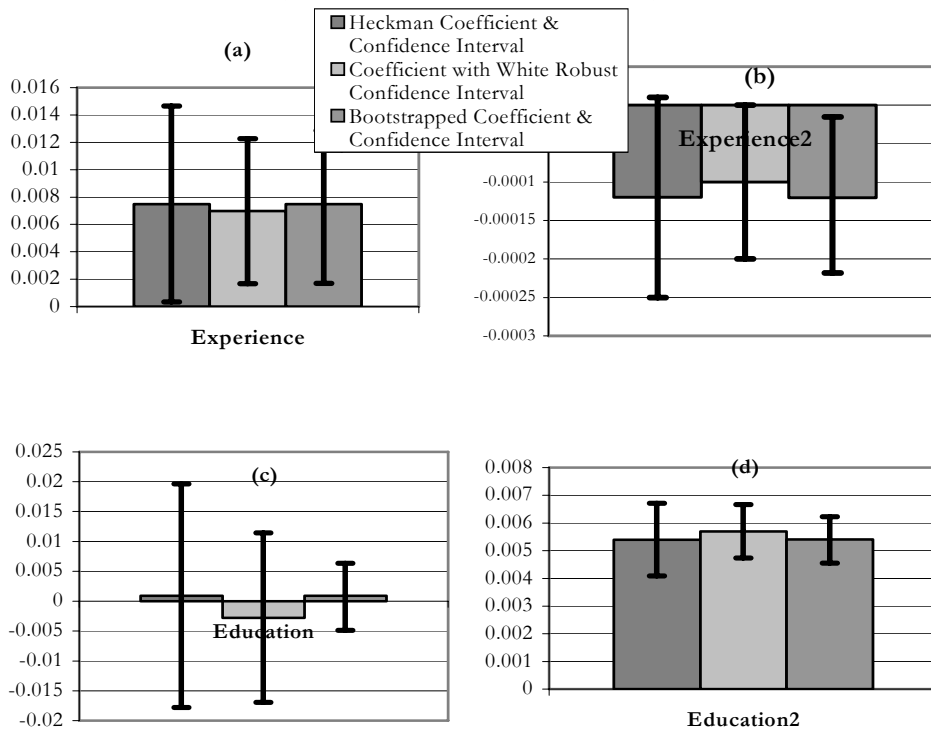
#### iv) Robust or Bootstrapped Confidence Intervals?

Figure 2 compares 95% confidence intervals based on usual Heckman standard errors with Huber-White robust and bootstrapped confidence intervals: Male midpoint estimates of Mincerian variables are chosen for illustrative purposes<sup>13</sup>. The most apparent feature is that the Heckman intervals are consistently the broadest. The bootstrap and robust intervals are fairly similar 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 gains in precision, and in some instances the robust intervals are the most efficient. The relevance of these comparisons is particularly evident for *Experience Squared*: both the Heckman and robust intervals include zero, which implies that the coefficient is insignificant at a 5% level. The bootstrapped confidence interval, in contrast, refutes this evidence, with the quadratic term being negatively significant; this is consistent with theoretical priors. Coefficients appear equal, though associated confidence intervals attach zero or negative value to their influence on earnings, depending solely on the efficiency of the estimator.

As Hill et al. (2003: 19) conclude, work on large samples with relatively 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 with only a robust correction.

<sup>13</sup> Interval regressions are not readily compatible with the Heckman covariance structure, hence midpoint estimates are used to enable comparison.

Figure 2 Comparison of Coefficients Magnitudes and 95% Confidence Intervals (Male Midpoint Estimates)



## 7) CONCLUSION

While bracket data pose obstacles before economic interpretation can credibly commence, 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. Do the available methods alter standard views of datasets? A Monte-Carlo simulation study delineates the circumstances which researchers should heed before applying midpoints blindly. This includes severe skewness, large numbers of right-censored observations and restrictively small sample sizes. It has, however, been shown that given the typical South African household survey bracket structure, parameters of earnings functions exhibit reasonable stability, regardless of the technique applied. While certain underlying statistical differences are apparent, the fact that magnitudes’ confidence interval estimates are in each instance comparable, reveals that (for the purposes of economic interpretation) a satisfactory solution has been reached. Ziliak and McCloskey (2004) in fact warn econometricians not to attach the entire weight of conclusions to strict 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.

The tools in this shed prove themselves to be sharp for the purposes of economic evaluation, should the considerations identified above be of no concern. Interval regressions survive a battery of experimental tests, and in fairly extreme cases of censoring (but not skewness) offer highly satisfactory outcomes. Imputation methods are more sensitive to these conditions: however, it has been shown that imputation in a typical LFS does not induce any notable distortion compared to interval regressions. Given that the bracket structure of South African household surveys remained stable, this bodes well for each of the surveys currently in use for labour market research, even if traditional midpoint imputation is applied. This conclusion is, however, subject to the assumption that skewness and many right-censored values do not distort the picture. This is good news for applied researchers, since midpoint imputation is simpler to execute than any other imputation or interval regressions.

The validation of these methods certainly does translate perceptions of “bad data” to “good data”. In effect, bracket coding should not deter labour market analysis, but add important information and lead to

improved practice in econometrics. “Rudimentary” methods such as midpoint imputation should not be dismissed; they prevent the loss of vital information in a satisfactory manner. This validates much of past work and allows researchers the time to interpret results more carefully instead of attempting to overcome additional statistical obstacles. The fact that respondents are either uncertain about exact earnings or wish to escape the public eye, does not translate to a poorer outlook for estimates.

## 8) REFERENCES

- 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.
- 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: 437-457
- 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.
- BURGER, R.P. and JAFTA, R.C.C., 2006. *Returns to Race: Labour Market Discrimination in Post-Apartheid South Africa*. Stellenbosch Economic Working Papers: 4/2006.
- BURGER, R.P. and YU, D., 2006. Wage trends in post-Apartheid South Africa: Constructing an earnings series from household survey data. *Labour Market Frontiers*. October 2006: No.8: 1-8. Pretoria: South African Reserve Bank.
- 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*, Vol. 29 No4: 1239-1254.
- DACUYCUIY, L., 2005. Is the earnings-schooling relationship linear? A semiparametric analysis. *Economics Bulletin*, Vol. 3 No. 37: 1–8.
- DANIELS, R. and ROSPABÉ, S. 2005. Estimating an Earnings Function from Coarsened Data by an Interval Censored Regression Procedure. *Journal of Studies in Economics and Econometrics*. Vol. 29 No. 1: 29-45.
- 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>
- HAIJIVASSILOU, 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: 71-99
- HAIJIVASSILOU, V.A., McFADDEN and D., RUUD, P., 1996. Simulation of multivariate normal rectangle probabilities and their derivatives - Theoretical and computational results. *Journal of Econometrics*. Vol. 72 :85-134.
- 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* Vol. 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* Vol. 15 No. 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.

- JOHNSON, R.A. and WICHERN, D.W.W., 2002. *Applied Multivariate Statistical Analysis, Fifth Edition*. Upper Saddle River: Prentice Hall.
- 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 No. 4: 834-860.
- MAGRUDER, J. and NATTRASS, N., 2006. Exploring attrition bias: The Case of the Khayelitsha Panel Study (2000-2004). *The South African Journal of Economics*. Vol. 74 No. 4: 769-781.
- MALHERBE, J.E., 2007. *An Analysis of Income and Poverty in South Africa*. Masters Dissertation, Stellenbosch: Department of Statistics and Actuarial Science, University of Stellenbosch.
- 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.
- NATTRASS, N. and WALKER, R., 2005. Unemployment and Reservation Wages in Working-Class Cape Town. *The South African Journal of Economics*. Vol. 73 No. 3: 498-509.
- 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 No. 3: 872-885.
- POSEL, D. and CASALE, D., 2005. *Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa*. Biennial Conference of the Economic Society of South Africa, September 2005. Durban.
- POULOUKAS, S., 2004. Estimation and Comparison of Lognormal Parameters in the Presence of Censored Data. *Journal of Statistical Computation and Simulation*. Vol 74 No.3: 157-169.
- 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*. Series 25: 229-234.
- STATISTICS SOUTH AFRICA (StatsSA), 2003a. *METADATA. Labour Force Survey 2003:2*. 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: 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 No. 3: 496-522.
- VON FINTEL, D.P., 2006. *Earnings Bracket Obstacles in Household Surveys: How Sharp are the tools in the Shed?* Stellenbosch Economic Working Papers 08/06
- WHITEFORD, A. and McGRATH, M. 1994, *The Distribution of Income in South Africa*. Pretoria: Human Sciences Research Council.
- 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*. Vol. 33: 527-546.