Estimating an Earnings Function from Coarsened Data by an Interval Censored Regression Procedure

Reza C. Daniels

School of Economics
University of Cape Town
rdaniels@commerce.uct.ac.za

Sandrine Rospabé

Faculté des Sciences Economique
Université de Rennes I (France)
sandrine.rospabe@univ-rennes1.fr

Development Policy Research Unit
Working Paper 05/91

February 2005
ISBN 1-920055-05-3

Abstract

This paper estimates an earnings function where the dependent variable is a mix of point and interval data using an interval regression model based on a pseudo-maximum likelihood estimation procedure. The analysis uses the 1999 OHS, and takes into account point and interval income observations, as well as design features of the survey including stratification, clustering and weights. In developing and applying the methodology, it is shown that researchers interested in analysing the determinants of income in a meaningful way need not be hampered by the presence of both point and interval observations, and can in fact account for these simultaneously using a generalised Tobit model. By incorporating survey design features into the analysis of the variance, some changes were needed to the estimation procedure and this is where the pseudo-likelihood becomes useful. However, this then affects how the coefficients of the model are interpreted, and researchers are encouraged to focus attention on the confluence of these factors.

JEL Classification: **C42, C51**

Key words: **Generalised Tobit Model, Pseudo Maximum Likelihood Estimation, Complex Survey Data**

Acknowledgements

The authors would like to thank participants at the 6th Annual Conference of the African Econometric Society at the University of Pretoria for their useful comments. We would also like to thank the DPRU for their financial assistance in the publication of this paper

Development Policy
Research Unit
Tel: +27 21 650 5705
Fax: +27 21 650 5711

Table of Contents

1. Introduction

In this paper we discuss an approach to estimating earnings functions from complex survey data using both point and interval observations simultaneously. Typically, survey questions that ask respondents to provide information on income, expenditure, assets and liabilities are subject to both high levels of item missing data as well as to potential measurement error if point observations are required for these variables. As a consequence, Statistics South Africa provide respondents to their household surveys with two options for the income question, namely actual (point) income and interval income categories (e.g. R10,000-R15,000). The resulting distribution of the income variable contains a mixture of actual value responses, interval censored responses, and missing data. Heitjan and Rubin (1990, 1991) call this mixture of data types "coarsened data" – and the phrase has become more widely used within the survey statistics literature (see also Heeringa et al, 2002).

The consequence of having both point and interval income observations makes estimating earnings functions more complex. Our innovation in this paper is to use a generalised *Tobit* model for this procedure. A further dimension of complexity is added to this task when survey sampling design features are considered, including stratification, clustering and weights (see Kish, 1965; Lehtonen & Pahkinen, 1995), where conventional maximum likelihood estimation is no longer possible and pseudo-likelihood estimation must be used. Thus, two analytical questions are addressed here. The first is how to estimate an earnings function using both point and interval observations. The second is how to account for survey design in the estimation method. As will become evident, both of these questions must be addressed in order to obtain accurate coefficients and correct estimates of their precision. The analysis is conducted on the 1999 October Household Survey (OHS) (Statistics South Africa, 1999).

The analysis below proceeds as follows. Firstly, the methodology is discussed. In this section, the model is presented as well as the estimation procedure given the features of complex surveys. Thereafter, the data and variables are described. Section 5 displays the empirical outcomes, where descriptive statistics for the regressed covariates are firstly provided before the results of the earnings function are discussed. Lastly, the conclusion summarises.

2. Methodology

2.1 Estimation of a Generalised Tobit Model for Interval Regression

Since income is a censored distribution in this case, the appropriate foundation from which to develop the estimation procedure is to use a censored regression or Tobit model, where the latent variable y_i^* is modelled by $y_i^* = \theta' x_i + \varepsilon_i$. Here, $y_i = 0$ if $y_i^* \leq 0$; $y_i = y_i^*$ if $y_i^* > 0$; and $\varepsilon \sim N(0, \sigma^2 I)$. Greene (2000: 911) provides the standard log-likelihood for the censored regression model, where:

$$\log L = \sum_{y_i>0} -\frac{1}{2}\left[\log(2\pi) + \log\sigma^2 + \frac{(y_i - \boldsymbol{\theta}'\mathbf{x}_i)^2}{\sigma^2}\right] + \sum_{y_i=0} \log\left[1 - \Phi\left(\frac{\boldsymbol{\theta}'\mathbf{x}_i}{\sigma^2}\right)\right]$$

To generalise the model and adapt the estimation procedure in order to account for the mixture of point, interval and missing observations needed for an accurate treatment of the income variable, we follow the procedure given below.

As before, let $y_i^* = \theta' x_i + \varepsilon_i$ be the model. We denote y_i, the observed dependent variable, as:

$$y_i = y_{Li} \text{ if } y_i^* \leq y_{Li}; y_i = y_{Ri} \text{ if } y_i^* \geq y_{Ri};$$
$$y_i = y_i^* \text{ if } y_{Li} \leq y_i^* \leq y_{Ri}; y_i = y_i^* \text{ if } y_{1i} \leq y_i^* \leq y_{2i}$$

Given this, the weighted log likelihood for the interval regression procedure is therefore given by the following (adapted from StataCorp, 2003a: 262):

$$\log L = -\frac{1}{2}\sum_{i\in C} w_i\left[\left(\frac{y_i - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right)^2 + \log 2\pi\sigma^2\right] + \sum_{i\in L} w_i \log\Phi\left(\frac{y_{Li} - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right)$$
$$+\sum_{i\in R} w_i \log\left[1 - \Phi\left(\frac{y_{Ri} - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right)\right] + \sum_{i\in I} w_i \log\left[\Phi\left(\frac{y_{2i} - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right) - \Phi\left(\frac{y_{1i} - \boldsymbol{\theta}'\mathbf{x}_i}{\sigma}\right)\right] \quad (1)$$

Here, observations $i \in C$ are point data, $i \in L$ are left censored, $i \in R$ are right censored and observations $i \in I$ are intervals. $\Phi(.)$ is the standard cumulative normal. Thus, regardless of the types of observations, the estimation method is able to account for them simultaneously.

However, the estimation of this model is complicated when survey design features are incorporated into the calculation of θ. Essentially, this implies that it no longer becomes possible to use a standard likelihood function, and a pseudo-likelihood has to be developed instead.

2.2 Pseudo-Likelihood Estimation (PML) of Parameters

The estimation technique for interval regression using the generalised Tobit uses a weighted maximum likelihood estimator. For complex survey data, however, this weighted likelihood is not the distribution function for the sample, since (i) when there is clustering, individual observations are no longer independent and the likelihood does not reflect this, and (ii) when there are sampling weights, the likelihood does not fully account for the randomness of the weighted sample. As it is not a true likelihood, it is termed a pseudo-likelihood. One of the consequences of the pseudo-likelihood is that standard likelihood-ratio (LR) tests are no longer valid, and Wald tests need to be used instead (see Eliason, 1993, 34-35 for a good discussion of other convenient features of Wald tests over LR tests in ML estimation).

Binder (1983) provided a rigorous treatment of how the variance of asymptotically normal estimators should be estimated from complex surveys, and it was this theoretical framework that subsequently became synonymous with PML estimation. It should be noted that the estimation of variance for a complex survey statistic is complicated not only by the nature of the survey's design, but also by the form of the statistic. In the event of a Tobit regression coefficient estimated by PML and incorporating survey design components, the variance formulae take on an added dimension of complexity. Therefore, while equation (1) is an efficient model to use with an interval-censored dependent variable, it will not yield either the correct coefficients or precise standard errors if it were estimated from complex survey data without taking into account the relevant survey design features.

In order to obtain accurate coefficients, appropriate survey weights must be used. In order to obtain precise standard errors, the effects of stratification, multi-stage sampling and weighting should be incorporated into the coefficient and variance estimates. Since *n* is large in the OHS99, finite population corrections need not be included. These features of complex surveys are standard in all Statistics South Africa's household surveys, and their omission constitutes an important, though frequently unrecognised source of error. Below we show how the coefficients and their variance are estimated using PML (adapted from StataCorp, 2003b: 39-40).

Let (h, α, β) index the elements in the population, where $h = 1, \ldots, H$ are the strata, $\alpha = 1, \ldots, A_h$ are the clusters (or primary sampling units – PSUs) in stratum h, and $\beta = 1, \ldots, B_{h\alpha}$ are the elements in PSU (h, α). Suppose that we observed $(Y_{h\alpha\beta}, X_{h\alpha\beta})$ for the entire population, and that $(Y_{h\alpha\beta}, X_{h\alpha\beta})$ arose from a suitable likelihood model as in equation (1). Let $l\left(\theta; Y_{h\alpha\beta}, X_{h\alpha\beta}\right)$ be the associated log-likelihood under this model. Then, for a finite population, we define the parameter θ by the vector estimating equation:

$$G(\theta) = \sum_{h=1}^{H} \sum_{\alpha=1}^{A_h} \sum_{\beta=1}^{B_{h\alpha}} S(\theta; Y_{h\alpha\beta}, X_{h\alpha\beta}) = 0$$

Where $S = \partial l / \partial \theta$ is the score vector, i.e. the first derivative with respect to θ of $l\left(\theta; Y_{h\alpha\beta,}, X_{h\alpha\beta}\right)$. Then, the PML estimator $\hat{\theta}$ is the solution to the weighted sample estimating equation:

$$\widehat{G}(\theta) = \sum_{h=1}^{H} \sum_{\alpha=1}^{A_h} \sum_{\beta=1}^{B_{h\alpha}} w_{h\alpha\beta} S(\theta; y_{h\alpha\beta}, x_{h\alpha\beta}) = 0 \qquad (2)$$

For the estimated coefficient $\hat{\theta}$ in (2) above, it is then possible to use a first-order matrix Taylor series expansion to produce the variance estimate.

$$V(\hat{\theta}) = \left\{ \left[\frac{\partial G(\theta)}{\partial \theta} \right]^{-1} \hat{V}(\hat{G}(\theta)) \left[\frac{\partial \hat{G}(\theta)}{\partial \theta} \right]^{-1} \right\} \Big|_{\theta=\theta} = H^{-1} \hat{V}(\hat{G}(\theta)) \Big|_{\theta=\theta} H^{-1} \qquad (3)$$

Where H is the Hessian for the weighted sample log-likelihood.

The use of the Taylor series expansion in equation (3) is one (tractable) example of how we can calculate the variance for the regression coefficients, and follows Kish's (1965) general identification of this method for complex surveys and Binder's (1983) specific adaptation to the PML framework. However, it is by no means the only one. Sul Lee et al (1989) provide a simple analysis of replication methods for estimating variance, including Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR). These replication methods are generally more useful than the Bootstrap when the underlying distributional assumptions of key variables are known. However, replication methods are generally more computationally intensive to derive than Taylor series approximations, which have become the standard approach in most current software programs (e.g. *Stata*, *SAS*). In the analysis below, the standard errors are estimated using the Taylor series.

Once we estimate the variance, it is then also possible to evaluate the precision of the coefficients estimated from the OHS99 given its complex survey features, relative to a simple random sample (SRS) of the same size. This is known as the design effect *(deff)* (Kish, 1965), and provides us with additional insight into the effect of survey design on the precision of the estimates. It is computed as:

$$deff\left(\hat{\theta}\right) = \frac{Var_{complex}\left(\hat{\theta}\right)}{Var_{srs}(\theta)} \qquad (4)$$

Where θ is the parameter of interest.

2.3 A Note on the Weights

There are two different weights that are applicable to this analysis. The first is the computation of $w_{h\alpha\beta}$ in equation (2), which is part of the weighted log likelihood in the pseudo-ML procedure. It accounts for heteroskedasticity and the number of replicates in an iterative likelihood procedure. The second weight is distinct from the first, and is developed in order to account for the design features of a complex survey. This weight is, in turn, comprised of three components: (i) compensation for unequal probability of selection (denoted w_1), (ii) adjustment for non-response (denoted w_2), and (iii) post-stratification adjustments (denoted w_3). The three weights are calculated as follows:

$$w_1 = 1/p(\beta) \quad ; \qquad w_2 = 1/r_\beta \qquad \text{and} \qquad w_3 = c.N/m_\beta$$

Here, $p(\beta)$ is the probability that unit β is sampled; r_β is the response rate; c is a constant chosen so that the weights sum to the number of respondents; and N is the population total (e.g. obtained from the Census) of a given number of respondents m_β.

The final weight (w) is then the product of the three individual weights, given by:

$$w = \prod_{HA_hB_{h\alpha}} w_1 w_2 w_3$$

Therefore, it is w that must be used as the weight of choice in the survey design adjusted parameter estimates.

In the analysis below, we use Statistics South Africa's (SSA) weight in the OHS99 since it is computed in this manner. It is important to be aware of the fact that adjustments to the OHS99 weight – i.e. the weight provided by SSA in the publicly released version of the dataset – in order to compensate for population growth and other demographic changes, constitutes an adjustment to w_3 only (i.e. the post-stratification weight). If this adjustment is made without factoring out w_1 and w_2 (thereby isolating the post-stratification factor of the product), then the resulting weight would be incorrect.[1]

1 Since the weight is a product function, it would be useful for Statistics South Africa to include all three weights plus the combined weight in the survey released to the public. This would allow researchers to make their own post-stratification adjustments, or, indeed, to create alternative weights based on some other procedure (e.g. imputation).

3. Data and Variables

The data for this exercise is taken from the 1999 October Household Survey (OHS99), conducted by Statistics South Africa. A two-stage sampling procedure was applied in the OHS, and the sample was stratified, clustered and selected to meet the requirements of probability sampling. The sampling procedure involved primary stage stratification by province and area type (urban/rural). Independent samples of Enumerated Areas (EAs) were systematically selected with probability proportional to size in each stratum; these are the clusters. The measure of size was the estimated number of households in each Enumerated Areas. A systematic sample of 10 households was then drawn from each EA, amounting to 30 000 households in 3 000 EAs.

The sub-sample of individuals evaluated in this study is limited to:

- Workers whose ages range from 15 to 65 (i.e. all economically active individuals);
- Those who are employed by someone else (we thus exclude self-employed people who only report their gross turnover); and
- Those for which information is available for wages and all other relevant attributes.

These restrictions reduce the original sample size to 17 945 individuals.

3.1 Dependent Variable

Information on earnings relates to total salary/pay, including overtime, allowances and bonuses before tax. The worker is asked to give either the precise amount of their salary or the income interval in which it fits, on a weekly, monthly or annual basis. Thus, the observations for the dependent variables consist of a mixture of point and interval data. Despite the fact that we omit both item and unit missing data from the regression (rather than imputing as per Heeringa et al, 2002), the data is still termed “coarsened” in the Heitjan and Rubin (1990, 1991) sense. Indeed, their definition of this phrase is flexible enough to be applied even to data that have only been grouped to ensure confidentiality.

All the observations were then converted into monthly data, though **it is common to use hourly earnings to abstract from the effect of variations in hours worked. However, even if workers report the number of hours they usually work per week, the presence of interval income data prevents us from working with the hourly wage rate. In order to account for this, working hours are introduced as an independent variable.** Lastly, the model we use assumes normality, and since the distribution of wages is skewed and non-normal, we more closely approximate normality if we model the log of wages.

3.2 Independent Variables

Independent variables include the following:

- A set of educational dummies[2], a variable for age and one for tenure – which proxies on-the-job-learning – are introduced to test the human capital theory.
- Quadratic terms for age and tenure are included to allow for increasing and then decreasing returns to age and experience over the life cycle.
- Racial dummies are introduced to assess whether, other things being equal, race plays a role in the determination of earnings. This is a simple, but not comprehensive, way of detecting racial discrimination in the labour market.
- Following the same method for race, a dummy for male is included to test for gender discrimination.
- Variables for marriage and headship status are traditionally set as determinants of earnings as proxies for factors such as stability, motivation and discipline.
- We also add a dummy variable for location to test the hypothesis that workers in urban areas earn more than in rural areas.
- Dummies for the provinces are also included, to take into account the differences in the cost of living.
- A dummy for union membership is introduced to investigate the union power over wage setting. We thus also test whether unionised workers earn higher wages than non-unionised.
- A dummy for the nature of the activity – formal / informal – is also included to test if working in a registered activity is more lucrative than in a non-registered activity.
- Finally, we introduce a set of 10 sectoral dummies and 10 occupational categories, since earnings are expected to vary substantially among industries and occupations.

4. Results and Discussion

4.1 Descriptive Statistics

In order to discuss the effect of sampling design on the analysis of simple descriptive statistics, Table 1 presents the mean, proportion and standard errors of the set of variables described above. They are successively calculated first under simple random sampling (columns 2), then integrating weights into the computations (column 3) and then including stratification, clusters and weights (column 4). The last column shows the individual design effect (*deff)* values for each variable; see equation (4) above.

2 No education, primary (grade1-grade7), secondary (grade8-grade12), further education (National Technical Certificate), higher education (diploma with grade12, degree, postgraduate degree or diploma).

Table 1 : Descriptive statistics of earning with and without survey design features: 1999 OHS

Variable	Without weights,clusters and strata		With weights only		With weights,clusters and strata		Design Effect
	Mean proportion	Std. Error	Mean or proportion	Std. Error	Mean or proportion	Std. Error.	
Dependant variable							
Monthly income[1]	2782.0290	269.3999	2963.6650	304.7622	2963.6650	317.7013	1.34
< 200[2]	0.0498	0.0026	0.0438	0.0025	0.0438	0.0032	1.83
[201-500]	0.1137	0.0037	0.0979	0.0038	0.0979	0.0047	1.80
[501-1000]	0.1344	0.0040	0.1173	0.0041	0.1173	0.0054	2.02
[1001-1500]	0.1388	0.0041	0.1349	0.0045	0.1349	0.0055	1.90
[1501-2500]	0.1780	0.0045	0.1794	0.0051	0.1794	0.0060	1.77
[2501-3500]	0.1180	0.0038	0.1195	0.0043	0.1195	0.0048	1.56
[3501-4500]	0.0856	0.0033	0.0925	0.0040	0.0925	0.0048	1.96
[4501-6000]	0.0783	0.0032	0.0869	0.0040	0.0869	0.0046	1.95
[6001-8000]	0.0470	0.0025	0.0561	0.0033	0.0561	0.0037	1.83
[8001-11000]	0.0243	0.0018	0.0301	0.0026	0.0301	0.0029	2.05
[11001-16000]	0.0194	0.0016	0.0249	0.0024	0.0249	0.0029	2.46
[16001-30000]	0.0095	0.0011	0.0129	0.0017	0.0129	0.0018	1.90
> 30000	0.0030	0.0006	0.0039	0.0009	0.0039	0.0010	2.03
Independent variables[3]							
Schooling							
No education	0.0990	0.0022	0.0783	0.0021	0.0783	0.0028	1.96
Primary	0.2889	0.0034	0.2553	0.0036	0.2553	0.0051	2.47
Secondary	0.4746	0.0037	0.5047	0.0043	0.5047	0.0059	2.46
Further education	0.0206	0.0011	0.0243	0.0014	0.0243	0.0016	1.95
Higher education	0.1170	0.0024	0.1374	0.0032	0.1374	0.0051	3.89
Age	36.9282	0.0786	36.0724	0.0882	36.0724	0.1004	1.69
Age square	1474.6920	6.1715	1408.1640	6.8404	1408.1640	7.7319	1.65
Tenure	6.9706	0.0587	6.5930	0.0633	6.5930	0.0790	1.95
Tenure square	110.3507	1.8955	100.7855	2.2087	100.7855	2.4671	1.71
Race							
White	0.1201	0.0024	0.1686	0.0037	0.1686	0.0082	8.63
African	0.6834	0.0035	0.6637	0.0042	0.6637	0.0092	6.87
Coloured	0.1710	0.0028	0.1354	0.0027	0.1354	0.0060	5.55
Indian	0.0245	0.0012	0.0311	0.0016	0.0311	0.0037	8.20

Other race	0.0010	0.0002	0.0012	0.0003	0.0012	0.0004	2.82
Male	0.5637	0.0037	0.5729	0.0043	0.5729	0.0046	1.56
Monthly hours	203.7437	0.4688	201.6373	0.5167	201.6373	0.6904	2.24
Urban	0.6433	0.0036	0.7007	0.0038	0.7007	0.0060	3.10
Union	0.3725	0.0036	0.3693	0.0041	0.3693	0.0064	3.17
Marital status	0.5012	0.0037	0.4984	0.0043	0.4984	0.0060	2.60
Headship status	0.5807	0.0037	0.5817	0.0043	0.5817	0.0050	1.85
Formal	0.8068	0.0029	0.8146	0.0033	0.8146	0.0046	2.50
Industries							
Manufacturing	0.1278	0.0025	0.1423	0.0031	0.1423	0.0042	2.55
Agriculture	0.1555	0.0027	0.1139	0.0025	0.1139	0.0054	5.25
Mining	0.0691	0.0019	0.0604	0.0019	0.0604	0.0046	6.83
Utilities	0.0080	0.0007	0.0085	0.0008	0.0085	0.0010	1.94
Construction	0.0461	0.0016	0.0485	0.0019	0.0485	0.0022	1.85
Trade	0.1405	0.0026	0.1531	0.0032	0.1531	0.0039	2.15
Transport	0.0416	0.0015	0.0486	0.0020	0.0486	0.0023	2.01
Finance	0.0702	0.0019	0.0875	0.0027	0.0875	0.0033	2.40
Services	0.2035	0.0030	0.2119	0.0035	0.2119	0.0053	3.00
Domestic services	0.1330	0.0025	0.1201	0.0027	0.1201	0.0035	2.03
Occupations							
Managers	0.1171	0.0024	0.1194	0.0028	0.1194	0.0034	2.02
Professionals	0.0381	0.0014	0.0466	0.0020	0.0466	0.0024	2.37
Technicians	0.0461	0.0016	0.0545	0.0021	0.0545	0.0028	2.73
Clerks	0.0974	0.0022	0.1074	0.0028	0.1074	0.0035	2.27
Salesperson	0.0991	0.0022	0.1145	0.0029	0.1145	0.0035	2.22
Artisans	0.1029	0.0023	0.1109	0.0028	0.1109	0.0034	2.06
Skill agricultural workers	0.0414	0.0015	0.0367	0.0016	0.0367	0.0021	2.19
Operators	0.1289	0.0025	0.1223	0.0027	0.1223	0.0036	2.13
Elementary workers	0.2173	0.0031	0.1877	0.0032	0.1877	0.0050	2.94
Domestic workers	0.1118	0.0024	0.1001	0.0024	0.1001	0.0030	1.84

Provinces							
Western Cape	0.1752	0.0028	0.1609	0.0031	0.1609	0.0041	2.19
Eastern Cape	0.0826	0.0021	0.0876	0.0024	0.0876	0.0040	3.60
Northern Cape	0.0578	0.0017	0.0263	0.0009	0.0263	0.0017	2.09
Free State	0.0991	0.0022	0.0848	0.0021	0.0848	0.0031	2.22
Kwazulu-Natal	0.1243	0.0025	0.1635	0.0035	0.1635	0.0055	3.91
North West	0.0968	0.0022	0.0785	0.0020	0.0785	0.0027	1.81
Gauteng	0.1853	0.0029	0.2605	0.0041	0.2605	0.0059	3.28
Mpumalanga	0.0978	0.0022	0.0707	0.0019	0.0707	0.0030	2.47
Northern Province	0.0810	0.0020	0.0671	0.0020	0.0671	0.0029	2.44

Notes: [1] 10 692 observations for monthly income

[2] 7253 observations for income intervals

[3] 17945 observations for each independent variable

This table first highlights the importance of using sampling weights in order to obtain the correct point estimates. Proportions from the weighted analysis differ by – 54 per cent (for the Northern Cape) up to 40 per cent (for the Whites) from the point estimates – ignoring the survey design parameters. Put differently, taking into account the weights leads to an increase in the proportion of White workers among the total workforce, which indicates that the proportion of Whites surveyed was lower than the true proportion of the population. Results for the Northern Cape show the opposite case, where the share of workers in the sample was too high compared to the true population proportion in South Africa.

Table 1 also shows that the survey design features of the sample generally reduce the precision of the sampling estimates. The reason is that workers living in the same clusters are usually more similar to one another in behaviour and characteristics than workers living in different clusters (Deaton, 1997). The *deff* is a useful concept to assess how the sample design affects precision. For example, we see that their values are particularly high for the race variable, implying that racial groups are highly clustered in South Africa. The *deff* is also important for agriculture and mining, where we find that people employed in these sectors are largely grouped.

On the other hand, age and gender are expected to cut across clusters uniformly, which explains why their *deff* values are low. Surprisingly, *deff* values associated with income are also low. Here we would have expected that income would have been more clearly associated with the racial groups, and as such be highly clustered. A possible explanation could be that almost 20 per cent of the Whites interviewed did not give either their exact income or the income interval in which they earned. As there is a high probability that these 20 per cent are not the least wealthy, it can explain why observations for high-income intervals are not largely grouped.

All these observations show that design effects from complex survey data do indeed influence the precision of the estimates and thus statistical inference. Consequently, if we ignore them, we increase the probability of making erroneous conclusions. The next section evaluates these issues for the earnings regression.

4.2 Regression Results

This section presents the results of the interval regression procedure in Table 2. The results are presented for the regression coefficients computed in equation (2) and their standard errors computed as the square root of equation (3). These outcomes represent the survey-design adjusted results and are the accurate coefficients and variance estimates described in the methodology. For comparative purposes, the unweighted non-design based coefficients and standard errors are also presented, and these amount to estimation under simple random sampling assumptions, labelled accordingly in Table 2. Lastly, we also present the mean design effects *(deff)* for similarly grouped variables, computed in equation (4).

Table 2 : Earnings interval regression with and without survey design features: 1999 OHS

	Simple Random Sampling		With survey-design		
Variables	Coefficient	Std. Error	Coefficient	Std. Error	Deff
Primary[a]	0.1199***	0.0216	0.1210***	0.0245	1.32
Secondary	0.3627***	0.0230	0.3734***	0.0272	
Further education	0.6363***	0.0469	0.6306***	0.0577	
Higher education	0.8153***	0.0325	0.8239***	0.0403	
Age	0.0376***	0.0038	0.0403***	0.0047	1.38
Age square	-0.0004***	0.0000	-0.0005***	0.0001	
Tenure	0.0233***	0.0016	0.0230***	0.0019	1.26
Tenure square	-0.0004***	0.0000	-0.0003***	0.0000	
African[b]	-0.6700***	0.0210	-0.6348***	0.0333	2.10
Coloured	-0.5429***	0.0257	-0.4632***	0.0385	
Indian	-0.3283***	0.0416	-0.2966***	0.0535	
Other race	-0.1562	0.1796	-0.2287	0.1453	
Male	0.1923***	0.0150	0.2032***	0.0178	1.32
Monthly hours	0.0009***	0.0001	0.0009***	0.0001	1.47
Urban	0.1766***	0.0152	0.1751***	0.0231	1.82
Marital status	0.0959***	0.0131	0.1092***	0.0178	1.74
Headship status	0.1406***	0.0141	0.1389***	0.0175	1.43
Formal	0.2594***	0.0191	0.2775***	0.0255	1.50
Union	0.2381***	0.0143	0.2133***	0.0201	1.74
Agriculture[c]	-0.5799***	0.0264	-0.5879***	0.0354	1.62

Mining	0.0912**	0.0286	0.0086	0.0454	
Utilities	0.3009***	0.0656	0.2545	0.0703	
Construction	-0.0556*	0.0328	-0.0938**	0.0451	
Trade	-0.1804***	0.0235	-0.1948***	0.0300	
Transport	0.0578*	0.0322	0.0239	0.0393	
Finance	0.0796**	0.0283	0.0819**	0.0303	
Services	0.0812**	0.0233	0.0382	0.0277	
Domestic services	-0.5063***	0.0520	-0.4948***	0.0613	
Managers[d]	0.5644***	0.0362	0.5557***	0.0526	1.51
Professionals	0.4252***	0.0385	0.4583***	0.0446	
Technicians	0.2941***	0.0300	0.3128***	0.0408	
Clerks	0.1536***	0.0279	0.1320***	0.0346	
Salesperson	-0.0591**	0.0275	-0.0708**	0.0346	
Skill agricultural workers	-0.1545***	0.0397	-0.1650**	0.0475	
Operators	-0.0632**	0.0242	-0.0830**	0.0296	
Elementary workers	-0.1869***	0.0236	-0.1907***	0.0296	
Domestic workers	-0.1951***	0.0548	-0.1801**	0.0643	
Eastern Cape Province[e]	-0.4884***	0.0266	-0.4570***	0.0391	2.10
Northern Cape Province	-0.3086***	0.0278	-0.3138***	0.0462	
Free State Province	-0.5560***	0.0265	-0.4992***	0.0412	
Kwazulu-Natal Province	-0.2025***	0.0256	-0.1792***	0.0374	
North West Province	-0.2253***	0.0271	-0.1776***	0.0383	
Gauteng Province	-0.0870***	0.0232	-0.0461	0.0327	
Mpumalanga Province	-0.2179***	0.0267	-0.1653***	0.0439	
Northern Province	-0.2476***	0.0285	-0.2292***	0.0389	
palphaConstant	5.9542***	0.0853	5.8331***	0.1105	1.54
Number of observations	17945		17945		
Number of strata			18		
Number of PSUs			2815		
Population size			7 042 100		
Model Chi2 (c.1) or F (c.2)	16 056		336.78		
Prob> Chi2 or F	0.00		0.00		

Notes: Associated standard errors are heteroscedastic-consistent. *** Statistically significant at the 1% level, ** the 5% level, * the 10% level. Reference category: (a) No education, (b) White, (c) Manufacturing, (d) Artisans and (e) Western Cape Province.

It should be noted that the coefficients in an interval regression are estimated by a pseudo-maximum-likelihood when survey design features are taken into account. As such, they are not directly interpretable since the coefficients predict the effects of changes in the exogenous variables on the latent variableas y^* as $\theta = \frac{\partial E(y_j^*)}{\partial x_j}$. For y, the marginal effect is expected to be smaller (see Maddala, 1983, 160). Despite this, comments can be made concerning the sign and relative size of the coefficients.

As most of the variables have a similar influence whether or not survey design features are accounted for, the general results of the two regressions are firstly considered. The block of educational dummies shows expected results. Schooling increases earnings, and the more educated workers are the higher the return of the year of schooling completed. Age and tenure have positive and decreasing returns on wages. We can thus conclude that all of these variables have an influence consistent with human capital theory.

Racial dummies are all significant and display the expected order. Other things being equal, Africans earn less than White workers, followed by Coloureds and Indians, corroborating similar results found by Hofmeyr (2000) on a 1993 sample, and consistent with South Africa's racially divided past. For further investigation of the estimates of racial discrimination, the residual difference methodology employed by Oaxaca (1973) should be utilised (see for instance Allanson et al (2000) and Rospabé (2002)).

The male dummy has a positive and significant influence on wages. Whereas this result can partly be explained by the fact that males and females don't benefit equally from the same contract of employment, an unknown part of the coefficient also reflects potential gender wage discrimination. As expected, the number of hours worked on average during a month positively influences earnings.

The results for the locational variables were also expected to some extent. Firstly, living in an urban area increases earnings. Secondly, the outcomes for provincial dummies show that earnings are lower for workers who are located in any other province other than the Western Cape. However, the coefficient for Gauteng is not significant when survey design is considered.

Being married and being the head of a household confers some advantages to workers, which indicates that these two variables could be a motivational signal for employers. Alternatively, it could also be due to confounding marriage with earning potential and age.

Turning to the impact of sectors on earnings, estimates show that workers in the formal sector earn higher wages than in the informal sector. This result is not unexpected as the formal dummy also reflects the effects of firm size and welfare contributions, which are likely to be larger in the formal sector. If we consider the results taking survey-design into account, we can also see that there are a few industrial sectors that provide significantly higher wages

than manufacturing, exemplified by the utility and finance sectors. However, other industries such as agriculture, trade and domestic work pay less than the manufacturing sector.

Union members earn significantly more than non-union members. This result is common in the literature on the union wage premium and highlights the strong bargaining power of South African unions over wages. Similar results have already been found in previous studies (Butcher and Rouse (2001), Moll (1993), Mwabu and Schultz (1998)), though for African workers only. As far as white workers are concerned, the premium is often found to be insignificant. Therefore, it should be expected that if the results were disaggregate by race, the conclusions would be quite different.

The results for the block of occupational dummies also display the expected wage hierarchy, where artisans were used as the base category. Estimates show that managers, professionals, technicians and clerks earn significantly more than artisans, whereas workers perceived as less skilled receive lower wages.

In the following section, we compare the results of the earnings interval regressions estimated under simple random sampling and when survey design was accounted for.

4.3 The Influence of Survey-Design

At first glance, there are no obvious differences between the results of the estimates with or without integrating survey-design. As expected, the standard errors increase when clusters are included into the analysis, since the simple random sampling regression overstates precision by ignoring the dependence of observations within the same PSU. Ignoring clustering leads to a rise in the probability of committing a type I error. The design effects (*deff*) are large for race and province variables, exceeding 2 on average. However, whether or not survey-design is accounted for, the probability of committing a type I error remains zero in both cases, except for Gauteng where the coefficient becomes insignificant under cluster sampling.

The interpretation of the results doesn't change too much except for the industries. Coefficients for mining, transport and service dummies are significantly different from zero, at least at the 10 per cent level in the case of simple random sampling estimates. However, they become insignificant when survey-design is taken into account. A Wald test shows that given survey-design features, we cannot reject the joint significance of the industrial dummies.

To some extent, the sizes of the coefficients differ when the data are weighted. Differences are small for human capital variables, urban locations and gender, but are larger for some industries, occupations and provinces. As the extent of the impact of each variable on earnings is difficult to interpret in the case of pseudo-likelihood, so are the effects of the variations in the size of the coefficients between simple random sampling and survey design.

In summary, it is evident that there were not large differences between the weighted relative to the unweighted coefficients. This indicates that the survey's sampling methodology was sound, capturing information from the sampled population that was not too different from the total population. However, the fact that the coefficients were different themselves, regardless of the magnitude of this difference, indicates that without incorporating the weights the coefficients would be incorrect. As far as the variance is concerned, it was evident that, with the exception of Other Race, every standard error in the regression results increased. Consequently, it is fundamental that survey design features be accurately incorporated into the variance formulae.

5. Conclusion

This paper has estimated an earnings function from coarsened data using the interval regression model based on a pseudo-maximum likelihood estimation procedure. The analysis used the 1999 OHS and took into account both point and interval income observations, as well as the design features of the survey including stratification, multi-stage sampling and weights.

In developing and applying the methodology, it was shown that researchers interested in analysing the determinants of income in a meaningful way need not be hampered by the presence of both point and interval observations, and can in fact account for these using a generalised Tobit model. By incorporating survey design features into the analysis of the variance, some changes were needed to the estimation procedure and this is where the pseudo-likelihood became useful. However, this then affected how the coefficients of the model were interpreted. Therefore, careful attention needs to be paid to the confluence of the model and its estimation procedures with survey design features.

The analysis of earnings was then undertaken both at the descriptive and analytical levels. In both instances, a comparison was made of the precision of survey design-based coefficients and variance estimates relative to their non design-based (simple random sampling) counterparts.

It was shown that the introduction of weights in the analysis significantly alters the size of the means of variables, and to a smaller extent, the size of the coefficients in an earnings regression. It was also observed that survey design features generally increase standard errors, as would be expected. In some cases, coefficients that were significantly different from zero under random sampling became insignificant when survey design was accounted for. These results point to the fact that adequate attention should be paid to features of complex survey data in order to yield both correct estimates of coefficients and their standard errors.

6. References

Allanson, P., Atkinks, J.P., and Hinks, T. (2000): "A multilateral decomposition of racial wage differentials in the 1994 South African Labour Market", *Journal of Development Studies*, 37 (1), 93-120.

Binder, D.A. (1983): "On the variances of asymptotically normal estimators from complex surveys", *International Statistical Review*, 51, 279-292.

Butcher, K. and Rouse C. (2001): "Wage effects of unions and industrial councils in South Africa", *Industrial and Labor Relations Review*, 54 (2), 349-74.

Deaton, A. (1997): *The analysis of household surveys: A microeconometric approach to development policy*, Baltimore: John Hopkins University Press.

Eliason, S.R. (1993): *Maximum likelihood estimation: logic and practice*, London: Sage Publications.

Greene, W.H. (2000): *Econometric Analysis*, Fourth Edition, New Jersey: Prentice Hall.

Heeringa, S.G., Little, R.J.A. and Raghunathan, T.E. (2002): "Multivariate Imputation of Coarsened Survey Data on Household Wealth", in Groves, R.M., Dillman, D.A., Eltinge, D.L., and Little, R.J.A. (eds): *Survey Nonresponse*, New York: John Wiley & Sons Inc.

Heitjan, D.F. and Rubin, D.B. (1990): "Inference from coarse data via multiple imputation with application to age heaping", *Journal of the American Statistical Association,* 85 (410), 304-314.

Heitjan, D.F. and Rubin, D.B. (1991): "Ignorability and coarse data", *Annals of Statistics,*19, 2244-2253.

Hofmeyr, J. (2000): "The changing pattern of segmentation in the South African Labour market", *Studies in Economics and Econometrics*, 24 (3), 109-128.

Lehtonen, R. and Pahkinen, E.J. (1995): *Practical Methods for Design and Analysis of Complex Surveys*, New York: John Wiley & Sons Inc.

Maddala, G.S. (1983): *Limited-dependent and qualitative variables in econometric*, Cambridge University Press.

Moll, P. (1993): "Black South African Unions: relative wage effects in international perspective", *Industrial and Labor Relations Review,* 46 (2), 245-61.

Mwabu, G. and Schultz, P. (1998): "Labor unions and the distribution of wages and employment in South Africa", *Industrial Labor Relation Review*, 51 (4), 680-703.

Oaxaca, R.L. (1973): "Male-female wage differentials in urban labor market", *International Economic Review*, 14 (3), 693-709.

Rospabé, S. (2002): "How did labour market racial discrimination evolve after the end of Apartheid?", *South African Journal of Economics*, 70 (1), 185-217.

StataCorp. (2003a): *Stata Statistical Software. Release 8.0 Base Reference Manual, Volume 4: S-Z"*, College Station, Texas: StataCorp LP.

StataCorp. (2003b): *Stata Statistical Software. Release 8.0 Survey Data Reference Manual*, College Station, Texas: StataCorp LP.

Statistics South Africa (1999): *October Household Survey*, Johannesburg: Statistics South Africa.

Sul Lee, E., Forthofer, R.N. & Lorimor, R.J. (1989) *Analyzing Complex Survey Data*, Sage London: Sage Publications.