# An Analysis of Income and Poverty in South Africa

by

Jeanine Elizabeth Malherbe

*Assignment presented in partial fulfilment of the requirements for the degree of Master of Commerce at Stellenbosch University*

Study leaders:

Prof. T de Wet   Dr. H Viljoen   Dr. A Neethling

March 2007

# Declaration

I, the undersigned, hereby declare that the work contained in this assignment is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature: ..............................
J.E. Malherbe

Date: ..................................

# Abstract

The aim of this study is to assess the welfare of South Africa in terms of poverty and inequality. This is done using the Income and Expenditure Survey (IES) of 2000, released by Statistics South Africa, and reviewing the distribution of income in the country. A brief literature review of similar studies is given along with a broad definition of poverty and inequality. A detailed description of the dataset used is given together with aspects of concern surrounding the dataset. An analysis of poverty and income inequality is made using datasets containing the continuous income variable, as well as a created grouped income variable. Results from these datasets are compared and conclusions made on the use of continuous or grouped income variables. Covariate analysis is also applied in the form of biplots. A brief overview of biplots is given and it is then used to obtain a graphical description of the data and identify any patterns. Lastly, the conclusions made in this study are put forward and some future research is mentioned.

# Uittreksel

Die doel van hierdie studie is om welstand van Suid-Afrika se burgers te ondersoek in terme van armoede en inkomste-ongelykheid. Die ”Income and Expenditure Survey (IES)” van 2000 word hiervoor gebruik en die verdeling van inkomste word ondersoek. ’n Kort oorsig oor soortgelyke studies word gegee, tesame met ’n breë definisie van armoede en inkomste-ongelykheid. ’n Indiepte verduideliking word gegee van die datastel wat gebruik gaan word, asook enige kwessies van belang aangaande die datastel. ’n Analise van die data word gemaak met behulp van ’n kontinue asook ’n kategoriese inkomste veranderlike. Die resultate van die verskillende datastelle word vergelyk en gevolgtrekkings aangaande die gebruik van kontinue of kategoriese inkomste veranderlikes word gemaak. Kovariaat analise word toegepas in die vorm van ’n biplot. ’n Kort verduideliking van ’n biplot word gegee en dit word gebruik om ’n grafiese verspreiding van die data te verkry, asook om enige patrone in die data te identifiseer. Laastens word die gevolgtrekkings wat in hierdie studie gemaak is gegee, sowel as ’n aantal moontlikhede vir verdere ondersoek.

# Acknowledgements

I would like to express my sincere gratitude to the following persons:

- My parents, for all the love and support in giving me the opportunity to get this far.
- My study leaders, Prof. T. de Wet, Dr. H. Viljoen and Dr. A. Neethling, for all their patience and advice.
- Prof. N.J. le Roux, for inspiring me to always give my best.
- And lastly, Dirko van Schalkwyk, for all the love and inspiration without which I could not have achieved this. And, of course, for all the help with LaTeX.

# Contents

# List of Figures

# List of Tables

# Acronyms

The following abbreviations were used in this assignment:

| | |
|---|---|
| CPI | Consumer Price Index |
| CVA | Canonical Variate Analysis |
| EA | Enumeration Area |
| EPRI | Economic Policy Research Institute |
| GDP | Gross Domestic Product |
| GE | Generalised Entropy |
| HDI | Human Development Index |
| HPHC | Home Production for Home Consumption |
| IES | Income and Expenditure Survey |
| LFS | Labour Force Survey |
| OHS | October Household Survey |
| PCA | Principle Components Analysis |
| PIMD | Provincial Index of Multiple Deprivation |
| PSU | Primary Sampling Unit |
| RDP | Reconstruction and Development Program |
| SRS | Simple Random Sampling |
| Stats SA | Statistics South Africa |
| UNDP | United Nations Development Program |

# Chapter 1

# Introduction

The analysis of income data in South Africa gives an insightful view into the distribution of poverty among the people of this country. It is not only an interesting subject but is of vital importance for a country in which the average percentage of poor is estimated at 58% when using a 'cost of basic needs' approach to define poverty (Hoogeveen & Özler, 2005). The study of inequality is of equal importance. Questions on the progress made since the end of Apartheid in 1994 can be linked directly to how inequality in South Africa changed between then and now. The new government of South Africa introduced the Reconstruction and Development Program (RDP) with the specific goal of dealing with aspects of poverty and inequality in the country. It is thus of great importance to them and other private researchers to know the extent of poverty and inequality in the country.

Poverty or Inequality cannot only be judged in terms of lack of monetary assets. Poverty/inequality can be lack of housing, it can be to lack adequate education and many other aspects of everyday life, some of them not even measurable. Hence, the importance of not only defining someone as being poor in terms of his/her monetary wealth is emphasized, but also looking at other variables that will have an impact on how an individual is classified. Many definitions of poverty exist emphasizing the multiple dimensions of poverty. In this study the focus will be on those variables defined to influence the classification of poverty.

## 1.1 Problem statement

The objective of this study will be three-fold, the first will be an overview of previous studies on poverty and income distribution in South Africa. The focus will be on the various definitions of poverty and income inequality. A description of the most common indicators used to measure poverty and inequality will be given, while a more in depth look will be taken at the various forms of deprivation.

The second objective will be to identify which techniques used in previous studies can successfully be applied to the current dataset to obtain a model for the distribution of income in the country. The focus will be on variations of the current dataset that are commonly used, how these datasets are manipulated into usable data and what techniques are used to obtain meaningful results from the data. The study will also attempt to answer questions concerning the type of income variable used in surveys, whether continuous and grouped income variables give matching outcomes.

Lastly, attention will be given to the influence of other variables on the prediction of poverty. General biplot techniques will be used to assess which variables are significantly linked to poverty. The aim is to identify those variables that will give a good indication as to the welfare of a person.

## 1.2 Study outline

The layout of the assignment is as follows: In the next chapter an overview is given of previous studies on poverty and inequality in South Africa. Certain technical aspects like poverty lines and poverty and inequality measurements are explained briefly. A broader definition is also given on poverty and inequality and an in depth look is taken at all aspects of deprivation. Chapter 3 gives a description of the dataset used in this study. It also explains the sampling design and weighting system used by Statistics SA to obtain the dataset. A brief overview is also given of how the smaller dataset was obtained and any adjustments made to it. In chapter 4 the actual analysis of the dataset is put forward, an in depth look is taken at the income distribution and conclusions are made on poverty and inequality in South Africa. Comparisons will be drawn between using a continuous and grouped income variable. Chapter 5 uses biplots to detect patterns or correlations between the different forms of deprivation. Conclusion of the results obtained in this study are given in Chapter 6 as well as some topics for future study.

# Chapter 2

# An Overview

Over the past decade many studies on poverty and income in post-Apartheid South Africa have been done on national and sub-national level with data more readily available in the form of Census 1996 and 2001, the Labour Force Surveys (LFS), Income and Expenditure Surveys (IES) of 1995 and 2000 and the October Household Surveys (OHS). In this chapter attention will be given to previous studies using these datasets and particular attention will be given to aspects pertaining to this study.

This chapter gives a broad definition of poverty, together with the different aspects of poverty. Measures of poverty is briefly explained, as well as poverty lines. Poverty indicators like the FGT family of indicators, HDI and Sen index is given. Income inequality is also defined together with its indicators, namely, the GE class of measurements, the Gini coefficient and the Decile dispersion ratio. The five elements of deprivation is summarized and a brief conclusion is made on what this chapter contains.

## 2.1 Defining poverty

No unanimous definition of poverty exist. The Concise Oxford Dictionary provides the following composite definition:

> Poverty is the state of lacking adequate means to live comfortable and want of things or needs indispensable to life.

This immediately highlights the various dimensions of poverty. These dimensions can be given by three main aspects of poverty:

- objective versus subjective,
- temporary versus chronic, and
- absolute versus relative.

**Objective versus subjective** ”Determining the extent or level of poverty requires a comparison between an observed and normative condition” (Boltvinik, 2001). This comparison can be made objectively or subjectively. In South Africa both objective and subjective indicators are used in defining poverty. Objective indicators include *Economic Deprivation*, deprivation in terms of income, expenditure/consumption or asset possession, *Educational Deprivation* and *Biological Deprivation*, either suffering from malnutrition, chronic disease or a disabling condition. These indicators usually refer to quantitative measures, whereas subjective indicators are generally associated with qualitative measures.

According to Govender *et al.* (2006) three subjective poverty dimensions are identified. The first being physical or social isolation due to peripheral location, lack of access to goods and services, ignorance or illiteracy. Secondly powerlessness within existing social, economic, political and cultural structures and thirdly,vulnerability to a crisis or the risk of becoming even poorer.

Hence, in the South African context, poverty is perceived by the poor to include alienation from the community, food insecurity, crowded houses, usage of unsafe and inefficient forms of energy, lack of jobs that are adequately paid and/or secure, and fragmentation of the family.

**Temporary versus chronic** Being poor is not a static condition. Individuals or households that can move between poor and non-poor over time are classified as *temporarily* poor entities, while *chronically* poor entities are observed as being poor at each successive observation.

**Absolute versus relative** Absolute poverty is determined without reference to the relative level of wealth of peers. It is claimed by Woolard *&* Leibbrandt (1999) to be an objective, scientific determination as it is based on the minimum requirement needed to sustain life. Relative poverty on the other hand is determined relative to the living standards of a society.

The above mentioned dimensions of poverty gives the general scope of poverty indicators. In the South African context, what indicators really give a reliable estimate of poverty? This is one of the main questions considered by the Government as it is vital information when trying to eradicate poverty. The multi-dimensionality of poverty was also asserted in the Reconstruction and Development Program (RDP):

> It is not merely the lack of income which determines poverty. An enormous proportion of very basic needs are presently unmet. In attacking poverty and deprivation, the RDP aims to set South Africa firmly on the road to eliminating hunger, providing land and housing to all our people, providing access to safe water and sanitation for all, ensuring the availability of affordable and sustainable energy sources, eliminating illiteracy, raising the quality of education and training for children and adults, protecting the environment, and improving our health services and making them accessible to all (African National Congress, 1994).

And more recently it has been argued that poverty should be seen :

> . . . in a broader perspective than merely the extent of low income or low expenditure in the country. It is seen here as the denial of opportunities and choices most basic to human development to lead a long, healthy, creative life and to enjoy a decent standard of living, freedom, dignity, self-esteem and respect from others (Statistics South Africa, 2000*b*).

Thus poverty is more than a physical state of deprivation, but is also perceived as a mental or psychological state of deprivation by the people of this country. To be poor also means to be alienated from your community. Measuring poverty using physical deprivation hence only attends to one aspect of poverty, but because of the difficulty in measuring these other aspects of poverty, this study will be based only on the measurable aspects of poverty.

Using the 1996 Census data Statistics SA has evolved two development indices, namely the *Household infrastructure index* and the *Household circumstance index*, to describe the extent of development of the different areas in South Africa (Hirschowitz *et al.*, 2000). A theoretically plausible list of relevant indicators were defined as:

- living in formal housing;
- access to electricity for lighting from a public authority or supply company;
- tap water inside the dwelling;
- a flush or chemical toilet;
- a telephone in the dwelling or a cellular telephone;
- refuse removal at least once a week by a local or district authority;
- level of education of the head of household;
- average monthly household expenditure;
- unemployment rate;
- average household size; and
- the proportion or children in the household under the age or five years.

In chapter 5 biplots will be used to try and identify which variables have an influence on the income level obtained by an individual or household.

## 2.2 Measures of poverty

It was said by Govender *et al.* (2006) that ”In order to measure poverty, there are a number of steps to be followed. Firstly, the concept of poverty being measured needs to be defined. Secondly, a poverty line - relative to the concept of poverty adopted – needs to be specified. Finally, the appropriate poverty measurements need to be selected”.

The diverse definitions of poverty naturally leads to a diversity of approaches to the measurement of poverty. Measures of poverty can be approached from two perspectives, one focusing on desired outcomes that are defined to characterize not being poor and the other considering the inputs necessary to eradicate poverty. The second approach proves to be the easier and more obtainable measure. The focus is thus on money-based measures, specifically measuring economic deprivation, although it is important to realize that this measure does not necessarily capture the full context of poverty. It does, however, give a good indication of the level of poverty.

Having acknowledged such money-based measures as an acceptable measure of poverty, the debate moves into the consideration of the relative merits of the income and expenditure methods. Measuring expenditure is a preferred approach for several reasons, the first being that it is a better measure of consumption than income, reflecting more directly the degree of commodity deprivation. Secondly, income tends to vary more over time than expenditure, thus expenditure gives a more smoothed and reliable picture of consumption and thirdly, income is less reliably reported in surveys, than expenditure.

The measure of poverty chosen can be analyzed at an individual or at a household level. In general the household level is preferred for the following reasons: (EPRI, 2001)

- Income and expenditure data is usually derived from household surveys and it is therefore difficult to break down further to an individual level. This is particularly the case with expenditure.

- The household is often considered to be the level at which economic decisions are taken. Income from individuals within a household is also often pooled, especially in the case of the poor.

But how can households of different sizes and composition be compared? The simples approach is to determine the household per-capita income/expenditure, determined by dividing the total household income/expenditure by the number of household members. This however does not allow for the economies of scale within households and thus requires a more complex form of normalization in order to compare households. This, however, falls outside the scope of this study and will therefor only be mentioned.

## 2.3 Poverty lines

The level (of the concept of poverty chosen) considered to be necessary to attain in order to be considered as not being poor is defined as the *poverty line*. Govender *et al.* (2006) defines a poverty line as: "A poverty line is the welfare (usually income/expenditure) level below which people are regarded as being poor". Poverty lines can be either absolute or relative. An absolute poverty line is defined with respect to the income/expenditure needed to attain a minimum standard of living, while a relative poverty line is defined by reference to others in the population. Absolute poverty lines are generally used and focuses on food/caloric needs. However, there is always an element of arbitrariness in poverty lines, despite the 'science' that exists in determining an appropriate level. The main use of poverty lines should thus be to assess changes in poverty over time, rather than the absolute extent of poverty at a particular time (Deaton, 2004, 2003; EPRI, 2001).

One of the most well known poverty lines is the $1 a day poverty line. It is used by the United Nations to measure extreme poverty across countries. Surveys by the United Nations (2005) using the $1 a day poverty line were taken in 1990 and 2001. Percentage wise the number of people living in extreme poverty in Asia has dropped drastically, reduced by at least 25% in more than 30 countries. It is, however, a totally different picture in Sub-Saharan Africa where the percentage has gone up from 44.6% in 1990 to 46.4% in 2001.

## 2.4 Poverty indicators

Having decided on the concept of poverty and critical level (poverty line) of this concept, it is necessary to define the indicators that will provide an indication of the level of poverty in the population under consideration.

### 2.4.1 Principles of defining a poverty measurement tool

There are certain accepted principles for providing a sound indicator of poverty. The four key principles that should be aimed for were put forward by Sen (1976):

- *Monotonicity axiom* - If the income of a poor individual falls (rises), the index must rise (fall).
- *Transfer axiom* - If a poor individual transfers income to someone less poor than herself (whether poor or non-poor), the index must rise.
- *Population symmetry axiom* - If two or more identical populations are pooled, the index must not change.
- *Proportion of poor axiom* - If the proportion of the population which is poor grows (diminishes), the index must rise (fall).

### 2.4.2 FGT family of indicators

The two most commonly used poverty measurement tools are the *headcount index* and the *poverty gap index*, both of these indices being special cases of the FGT class of poverty measures put forward by Foster *et al.* (1984). Woolard (2001) maintains that the headcount index measures the proportion of the population under consideration that is poor and the poverty gap index measures the average distance that a poor person is from the poverty line - the depth of poverty among the poor. A formulation of the FGT class of measures can be given as: (Woolard & Leibbrandt, 1999)

$$P_\alpha = \frac{1}{n}\sum_{i=1}^{q}[\frac{(z-y_i)}{z}]^\alpha \quad \text{for } \alpha \geq 0, \tag{2.4.1}$$

where

z is the poverty line,

$y_i$ is the welfare measure/indicator of the ith individual/household,

$\alpha$ is the "aversion to poverty" parameter,

n is the total individual/household population size,

q is the number of "poor" individuals/households.

When $\alpha = 0$, the FGT class yields the headcount index and when $\alpha = 1$, the outcome is the poverty gap index.

### 2.4.3 Other indicators

A number of widely quoted poverty/development indices are in use, based on a variety of different combinations of welfare measures and poverty lines. Two of the best known are the United Nations Development Program (UNDP) Human Development Index (HDI), and the Sen Index.

**HDI** The HDI measures the welfare across countries using three basic dimensions of human development: (Bhorat *et al.*, 2004)

- A long and healthy life, as measured by life expectancy at birth index.
- Knowledge, as measured by an education index, measuring both adult literacy and the general enrollment in primary, secondary or tertiary education.
- A decent standard of living, as measured by the Gross Domestic Product (GDP) per capita index.

**Sen Index** Another index proposed by Sen (1992) is a combination of the headcount index, the poverty gap index and the Gini coefficient. It is an attempt to reflect the degree of inequality in the distribution of income/expenditure amount the poor, and is calculated as the average of the headcount index and poverty gap index weighted by the Gini coefficient of the poor. As a formula it is given by Govender *et al.* (2006) as:

$$S = [H * G] + P * [1 - G], \tag{2.4.2}$$

where

H is the population headcount index,

P is the population poverty gap index,

G is the Gini coefficient of the poor.

Refer to section 2.5.2 for a definition of the Gini coefficient. When $G = 0$ the Sen index is simply the same as the poverty gap index and when $G = 1$ the Sen index would simply be the same as the headcount index. In other words, Sen's index takes into account the numbers of the poor, their shortfall in income/expenditure relative to the poverty line, and the degree of inequality in the distribution of their income.

## 2.5 Income and inequality

A second definition of welfare often considered in analysis is that of income inequality. According to Coudouel *et al.* (2002) poverty measures depend on the average level of income or consumption in a country and the distribution of income or consumption. Based on these two elements, poverty measures therefore focus on the situation of those individuals or households at the bottom of the distribution. Inequality is a broader concept than poverty in that it is defined over the entire population, not only below a certain poverty line.

### 2.5.1 Definition of inequality

Inequality looks at variations in the standards of living across a whole population or region, it refers to any aspect of deprivation - deprivation in terms of income, assets, health etc. However the focus is usually on income inequality. Two types of inequality exist, namely *relative* inequality and *absolute* inequality. Relative inequality depends on the ratios of individual income to the overall mean, while absolute poverty refers to the absolute differences in the levels of income. However relative inequality is most commonly used in literature dealing with the analysis of inequality.

### 2.5.2 Measures of inequality

Income inequality looks at the distribution of income in a population. There are various ways of measuring this income inequality and a good measure should generally meet the following set of axioms: (Litchfield, 1999)

- *Pigou-Dalton Transfer Principle* - An income transfer form a poor person to a richer person should register as a rise (or at least not as a fall) in inequality and an income transfer from a richer to a poorer person should register as a fall (or at least not as an increase) in inequality.
- *Income Scale Independence* - The inequality measure should not depend on the magnitude of total income.
- *Principle of Population* - The inequality measure should not depend on the number of income receivers.
- *Anonymity* - It should only be affected by the incomes of the individuals. No other characteristics of the individual should affect the index.
- *Decomposability* - This requires overall inequality to be related consistently to constituent parts of the distribution, such as population groups.

Any measure that satisfies all of these axioms is a member of the Generalised Entropy (GE) class of inequality measures.

**GE class of measurements** Members of the GE class of measures have the general formula as follows: (Govender *et al.*, 2006)

$$GE(\alpha) = \frac{1}{\alpha^2 - \alpha}[\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i}{\overline{y}})^{\alpha} - 1], \tag{2.5.1}$$

where

n is the number of individuals in the sample,

$y_i$ is the income of individual $i$,

$i \in (1, 2, \ldots, n)$,

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the arithmetic mean income.

The value of GE ranges form 0 to $\infty$, with zero representing an equal distribution and higher values representing higher levels of inequality. The parameter $\alpha$ in the GE class represents the weight given to distances between incomes at different parts of the income distribution, and can take any real value. The GE measures with parameters 0 and 1 become two of Theil's measures of inequality, the mean log deviation and the Theil index: (Litchfield, 1999)

**The mean Log Deviation**

$$GE(0) = \frac{1}{n}\sum_{i=1}^{n} log\frac{\overline{y}}{y_i} \tag{2.5.2}$$

**The Theil Index**

$$GE(1) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\overline{y}} log\frac{y_i}{\overline{y}} \tag{2.5.3}$$

Both of these measures are widely used because of their property of decomposability.

**Gini Coefficient** The Gini coefficient is the most widely used measure of income inequality. It varies between 0 (when there is perfect equality and all the individuals earn equal income) and 1 (when there is perfect inequality and one individual earns all the income and the other individuals earn nothing). The Gini coefficient is calculated from the Lorenz curve, which plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest households. Figure 2.1 provides a hypothetical example of a Lorenz curve. The Gini coefficient measures the area between the Lorenz curve and the hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. The only drawback in using the Gini coefficient is that it is not easily decomposable.

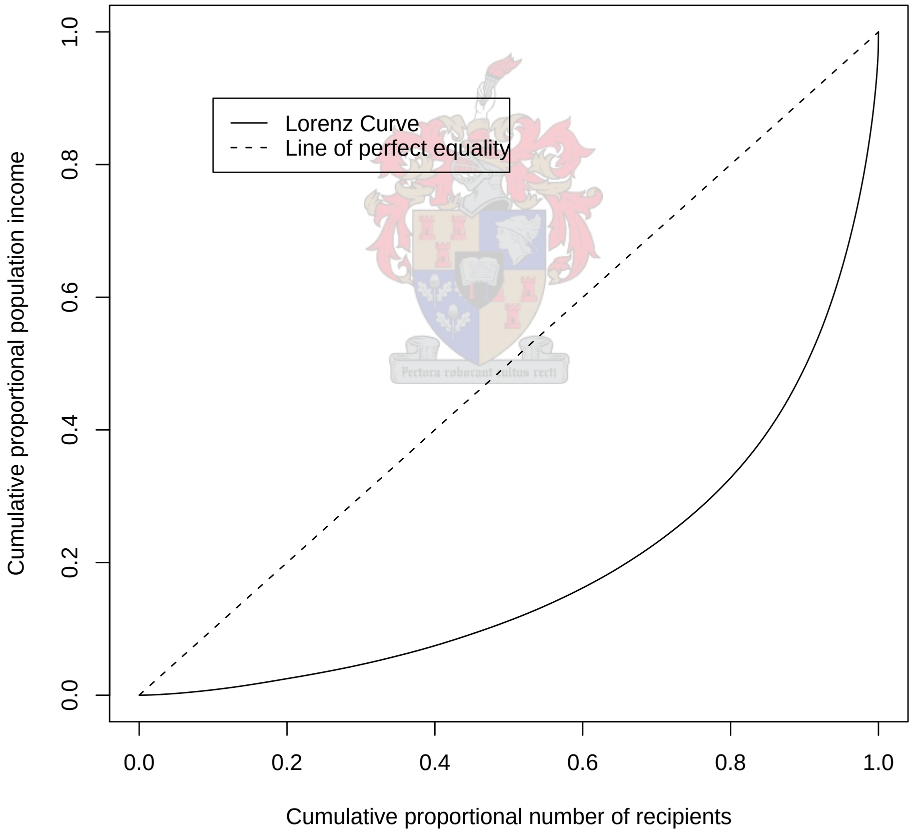


**Figure 2.1:** Lorenz curve.

An important aspect of the Lorenz curve is that it is mainly used to compare inequality between two distributions, drawing the respective Lorenz curves one can conclude that inequality is unanimously higher in one distribution if its Lorenz curve is everywhere below the curve of the other distribution. If the curves cross, the ranking is indeterminate. Methods exist to estimate the empirical Lorenz curves from the sample data, however these methods do not apply to the tails of the Lorenz curves since the tails contain to few observations. However, the tail behaviour is of considerable interest, and it is precisely in the tails where crossings often occur in practice (Schluter *&* Trede, 2002). Hence, Extreme Value Theory can be used to overcome this problem.

**Decile dispersion ratio** The decile dispersion ratio is also an inequality measure that is sometimes used. It represents the ratio of the average consumption or income of the richest 10 percent of the population divided by the average income of the bottom 10 percent (Coudouel *et al.*, 2002).

## 2.6 A closer look at deprivation

Poverty as defined above is a measure involving multiple deprivation. According to a study by Noble *et al.* (2006) multiple deprivation is a combination of unidimensional domains of deprivation which are combined using appropriate weighting. They identified five domains of deprivation using the Census 2001 data and used this to form an index of multiple deprivation for each province. The five domains were: Income and material Deprivation, Employment Deprivation, Health Deprivation, Education Deprivation, and Living Environment Deprivation. Each domain was presented as a separate domain index reflecting a particular aspect of deprivation. For each domain index a number of indicators were identified. A brief summary of their conclusions follow:

### 2.6.1 Income and material deprivation domain

The purpose of this domain is to capture the proportion of the population experiencing income and/or material deprivation in an area. Income deprivation is a good proxy for general material deprivation and is included in this domain alongside two *direct* measures of material deprivation. The indicators are:

- Number of people living in a household that has a household income that is below 40% of the mean equivalent household income; or
- Number of people living in a household without a refrigerator; or
- Number of people living in a household with neither a television nor a radio.

The income deprivation aspect of this domain is represented by the number of people in a ward living in households with equivalent income of less than 40% of the national mean. When combining the indicators a simple proportion of people living in households experiencing one or more of the deprivations was calculated.

There were some issues when considering income deprivation since all the income values of Census 2001 were reported in 12 bands (or income level) and reported at individual level. The problem was overcome by assigning income values (in most cases the logarithmic mean) to the bands. Another area of difficulty was the large numbers of missing values. Stats SA imputed values for the missing cases using a variety of techniques (e.g. logical or 'hot deck'). For those households with either missing values or 'implausible' zero values, multiple imputation techniques were employed to validate Stats SA's imputations.

### 2.6.2 Employment deprivation domain

This domain measures employment deprivation conceptualized as involuntary exclusion of the working age population from the world of work. The indicators are:

- Number of people that are unemployed (using the official definition); and
- Number of people that are not working because of illness or disability.

Stats SA uses two definitions of unemployment. According to the (international) official or strict definition, the unemployed are those people within the economically active populations who (a) did not work in the seven days prior to Census night, (b) wanted to work and were available to start work within a week of Census night, and (c) had taken active steps to look for work or start some form of self-employment in the four weeks prior to Census night. A person who fulfills the first two criteria above but did not take active steps to seek work is considered unemployed according to the expanded definition. The domain was calculated as the proportion of the economically active population (15 to 65 year olds inclusively) plus people not working due to illness or disability that were unemployed or not working due to illness or disability.

### 2.6.3 Health deprivation domain

The purpose of this domain is to identify areas with relatively high rates of people who die prematurely. There is only one indicator:

- Years of Potential Life Lost.

For the measure of premature deaths used in each of the PIMDs[1], Years of Potential Life Lost (YPLL), the level of unexpected mortality is weighted by the age of the individuals who has died, see Blane & Drever (1998).

### 2.6.4 Education deprivation domain

This domain is to capture the extent of deprivation in education qualification in a local area. The primary focus for this measure is adults aged 18 to 65 years. The single indicator is:

- Number of 18 to 65 year olds (inclusive) with no schooling at secondary level or above.

### 2.6.5 Living environment deprivation domain

The purpose of this domain is to identify deprivation relating to poor quality of the living environment. It has several indicators:

- Number of people living in a household without piped water inside their dwelling or yard or within 200 meters; or
- Number of people living in a household without a pit latrine with ventilation or flush toilet; or
- Number of people living in a household without use of electricity or lighting; or
- Number of people living in a household without access to a telephone; or
- Number of people living in a household that is a shack; or
- Number of people living in a household with two or more people per room.

A simple proportion of people living in households experiencing one or more of the deprivations was calculated.

[1] Provincial Index of Multiple Deprivation (PIMD)

## 2.7 Conclusions

The above topics give a broad view of some general knowledge on poverty, income inequality and multiple deprivation. It is clear that no fixed definition of poverty exists. To some degree the choice of poverty line, the line which we use to define poverty, is a subjective choice and also dependent on the particular dataset that is in use. There are positive as well as negative aspects to this approach. On the positive side each dataset and sampling population needs to be evaluated for its own distribution and it thus leads to a more accurate choice of poverty line for that specific dataset. On the other hand, difficulties arise in comparing poverty in cases were the poverty lines and indicators differ. Bias can also be introduced by an analyst choosing a poverty line based on what outcome he/she wants to achieve and not on what the dataset presents.

Viewing poverty from the opposite angle in terms of deprivation other than Income Deprivation, the choice of deprivation indicators becomes an important choice. These deprivation indicators are often difficult to assess or measure. People also perceive deprivation on different levels, hence no universal deprivation model will apply to all individuals. Deprivation indicators are thus constrained to those measurable elements like housing and access to piped water etc.

Poverty indicators like poverty lines and the FGT family of indicators have been defined and will be used in chapter 4 to asses poverty in South Africa using the IES 2000 dataset. The Gini coefficient and Theil index used to measure income inequality will also be applied. In chapter 5 the aspects of deprivation will be used to obtain biplots of the data and identify any correlation between certain types of deprivation. In the next chapter, however, attention will be given to the data set to be used in this particular study and how the above topics relate to this dataset. Some explanation will be given on what route of analysis will be followed and what methods will be used. Other points of interest concerning the dataset will also be discussed.

# Chapter 3

# Describing The Data

The previous chapter gave a brief overview of the most relevant aspects concerning poverty and income inequality analysis. In this chapter a more in dept approach will be followed with reference to the dataset to be used. The description of the dataset will be given in section 3.1 as well as the techniques used to refine the dataset into the format that will be used in analysis, in section 3.2. The chapter is started with a summary of how Stats SA conducted the IES 2000 survey, this includes the survey design, clustering and stratification and the weighting used. The smaller dataset to be used is described, together with any deviations from the original dataset. The debate around continuous and grouped income variables is briefly given and a summary of what is to be done with the IES 2000 dataset.

## 3.1 Income and Expenditure Survey 2000

The dataset that will be used in this analysis is the 2000 Income and Expenditure Survey (IES). The IES is a five-yearly household survey. This survey is used by Statistics South Africa to measure income and expenditure in the country. It measures the detailed income and expenditure of households. These surveys were originally designed and are still used to determine weights for the South African Consumer Price Index (CPI). Recently, however, it has become better known for showing the earning and spending capacity and expenditure patterns of South

African households. The survey is done by means of interviews with household heads or responsible adults and the questionnaire is completed by the enumerator during this interview. The information is then used to obtain a picture of the welfare or the citizens of South Africa.

The metadata file published with the IES 2000 provides a description of the data, the sample design, the sampling weights and the variables contained in the dataset. The raw data are published in four ASCII text files with each line representing a record or observation, in this case a household or person depending on whether it is a person- or household-level file. The first file, *person.txt*, contains person-level data of all members in the household, allowing for a maximum household size of 25 members. The file contains variables such as gender, age, race, work status and income from employment for each household member. The second file, *worker.txt*, contains information on domestic workers employed by households. The third file, *homegrownproducts.txt*, contains information on home production for home consumption (HPHC) of farm produce and livestock at the household level. This information is included in the income and expenditure sides of the applicable households and takes into account the market values of goods produced, the amount consumed, and the values of excess production sold, taking into account input costs. Finally, *general.txt* contains all the general income and expenditure data. The file is the largest of all the data files and contains the majority of the information collected for the IES 2000 (Provincial Decision-Making Enabling Project (PROVIDE), 2005).

### 3.1.1 Survey design

The design of household surveys is usually based on the most recent Census. In the case of the IES 2000 the sample was based on a master sample using the South African 1996 Population Census of enumerator areas (EA's). An EA consists of approximately 100-150 dwelling units. In some cases EA's are added to the original EA to ensure that the minimum requirement of 100 dwelling units is met.

The IES 2000 is a two-stage stratified sample using probability proportional to size principles. In the two-stage sampling design, clusters are first selected randomly from a list of clusters covering the entire population. Next, households are selected from each of the sampled clusters. This generates a final sample in which households are not randomly distributed over the population, but are grouped geographically. Some reasons for using clustering is that it is more cost-effective and sometimes the only available approach to use. The 1996 Census forms the basis for clustering in the IES 2000 sample. The 3000 primary sampling units (PSU's) in the IES 2000 are drawn systematically from the list of census enumeration areas (EA's) (see Statistics South Africa, 2000*a*).

Household income and expenditure surveys generally distinguish between provinces and area type (urban and rural). Therefore, in the case of the IES 2000, explicit stratification of the PSU's based on the nine provinces and by location (urban or rural) is applied, giving 18 explicit strata in total. Within each explicit stratum, the PSU's are also implicitly stratified according to Magisterial District or District Council, and then by average household income (in the case of formal urban areas or hostels) or EA. In each stratum the predetermined number of EA's were systematically selected with probability proportional to the number of dwelling units in that EA. Ten households were then systematically selected from each of the stratified PSU's. As a result 30 000 dwelling units were selected. Of this sample 26 265 households completed the questionnaires, thus giving a response rate of 87,55% (Provincial Decision-Making Enabling Project (PROVIDE), 2005).

### 3.1.2 Weighting

Statistics South Africa defined their initial weights (household weights) to be equal to the inverse of the probability of selection, based on the sample design. That is:

$$Household\ weight = \frac{1}{P_1 P_2}, \tag{3.1.1}$$

where

$$P1 = \frac{(Census\ number\ of\ households\ in\ PSU) * (number\ of\ PSU's\ in\ stratum)}{Census\ total\ number\ of\ households\ per\ stratum} \tag{3.1.2}$$

$$P2 = \frac{Sample\ size\ [that\ is,\ 10\ dwelling\ units\ per\ PSU]}{Number\ of\ dwelling\ units\ in\ the\ selected\ PSU} \tag{3.1.3}$$

The initial weight for each member of the household is the same as the weight for the household itself. Further adjustment factors were then calculated within the PSU's to account for non-response.

## 3.2 Dataset for analysis

A smaller dataset was created from the original Income and Expenditure Survey 2000 dataset by only keeping those variables deemed important to the study of income distribution and poverty. This was done by the Department of Economics and the Bureau for Economic Research at Stellenbosch University. This is the dataset that will be used in chapter 4 when analyzing income analysis and poverty.

### 3.2.1 Adjustments

The Department of Economics made some minor adjustments to the original dataset. These adjustments mostly related to mistakes made by Stats SA in the original dataset. One of these is Total Household Expenditure where Stats SA counted the expenditure on 'cereal' twice. This was corrected by the Department of Economics. The variables relating to household size, education level, age, race, gender etc. were included. The variables deemed most important in terms of spending and income were also included. On the expenditure side the variable 'grain-food' was included, measuring the amount spent on grain food. For income all the variable contributing to Total Household Income were included, these were items like renumeration, interest, property etc.

## 3.3 Continuous versus grouped income

There exists a lengthy debate on the subject of earnings brackets. Should variables measuring income be given as a continuous variable or should it be given as a grouped income variable? Many people are reluctant to give an exact income variable or don't know their income to the nearest Rand. This leads to a loss of information and also possible bias in the data collected. An alternative method of collecting earnings information is thus needed and earnings brackets provide a solution.

Instead of giving an exact income figure, respondents are asked to indicate to what predefined income intervaly he/she belongs. This leads to a significantly greater response rate for income variables, hence a better dataset is created with possibly more correct results. However, this leads to questions about the accuracy of the indicators obtained from this grouped income data. In a study by Von Fintel (2006) he found that results obtained using either a continuous income variable or a grouped income variable were equally accurate, when using a dataset containing both.

The IES 2000 dataset contains income as a continuous variable. The objective is now to test what the effect of a grouped income variable will be on poverty and income indicators. A dataset containing income as a grouped income variable thus needs to be created. Income levels will be specified using the income levels as defined for Census 2001.

## 3.4 What needs to be done?

The dataset to be used is now known, as well as the adjustments made to it. The sampling technique used to obtain the dataset has been mentioned, as well as how the household weights were obtained. The problem of continuous versus grouped income variables has been given, together with a bit of background on the subject. What remains to be done is the analysis of this problem. This will be done in the next chapter. The dataset first has to be cleaned from aspects like missing values and zero income. The grouped income dataset must then be created using the income levels defined for Census 2001. In order to analyze this grouped income dataset it then has to be made continuous again using three approaches, namely, the midpoint approach, the interval regression approach and the random midpoint approach.

The four datasets will be compared in terms of poverty lines, extreme tail distributions and income inequality. Do the four datasets give the same results? Is there differences between results obtained from continuous or grouped income variables? These are the questions that will be answered in chapter 4. A quick look will also be taken at poverty and income inequality between provinces. This, however, will be done using the continuous income dataset only.

# Chapter 4

# Analyzing Income and Poverty

In Chapter 3 a brief description was given of how the dataset that will be used was obtained. This will be referred to as the Revised IES dataset. In this chapter the issue of missing values is first addressed. This is followed by the problem of unrealistic zero incomes. Two methods are described for dealing with these zero incomes. The methods are compared and a decision is made of which one to continue the study with. Next, a grouped income dataset is created from the continuous Revised IES 2000 dataset. Three methods of making this grouped income dataset continuous are discussed. The three generated continuous datasets are then compared with the Revised IES 2000 dataset in terms of poverty lines, the extreme tail distributions and income inequality. Conclusions are made on the accuracy of each of the datasets in terms of predicting poverty and inequality. A brief analysis of poverty and inequality between provinces is included. The Revised IES 2000 dataset is used throughout.

The analysis will be based on annual income. Household income is used, unless specifically specified otherwise. In the case of per capita income the total household income was divided by the number of individuals in the household. This per capita income was then taken as a per capita 'household' income and not weighted by the number of individuals in the household.

## 4.1 Dealing with missing values

Stats SA dealt with missing values in the following manner. For each variable a code was given for missing/unspecified values. However, for Expenditure and Income the missing values were not coded. The same technique will thus be applied to those entries in the dataset still containing missing values. In other words, where Stats SA provides a code for missing values, this code will be used and where such a code is not applicable the term ”NA” will be used to indicate a missing value. However, problems arise when dealing with ”NA” values, thus where there are missing values for total household income the value will be put equal to zero. There are only two cases in which no value for income is given. The next section will explain what is to be done with zero income shown unrealistically as zero. This is the case where a household shows expenditure but claims to have zero income. In the rest of this chapter when there is reference to zero income it will refer to such an unrealistic zero income.

## 4.2 Dealing with zero income

Although it is quite reasonable to assume that there are individuals or households having zero income, contradictions arise when there is expenditure greater than zero but no income. There were no households claiming zero income and zero expenditure. There were, however, households claiming zero expenditure but not zero income. This difficulty will not be pursued further in this study since it will focus on income distribution and poverty.

The problem at hand is one of dealing with those households claiming zero income, but not zero expenditure. There were 254 households out of a total of 26217 (approximately 0.97%) claiming zero income. There are two ways of dealing with this problem. The first, method 1, takes total household income as equal to total household expenditure and the second approach, method 2, uses an imputation method of approximating missing values.

### 4.2.1 Method 1

This method was used by Stats SA in analyzing the IES 2000 dataset. It involves setting total household income equal to total household expenditure in cases where the income is given as zero. It is however not a good method of approximating income as it does not take into account other factors pertaining to the income level of a household such as education level of the head of the household or the number of individuals in the household.

### 4.2.2 Method 2

Imputation is commonly used to assign values to missing items. A replacement value, often from another observation in the survey that is similar to the item nonrespondent on other variables, is imputed for the missing value (Lohr, 1999). It is this property of imputation that will be used to deal with zero income, as described in the next section. We will refer to this as the imputation approach.

Important decisions need to be made on what variables will be used for the imputation approach. The role of these variables will be to form cells (classes) of similar households and allow zero income values to be imputed. The education level of the head of the household and household size seem to be the most appropriate variables. After the households are divided into these cells, the unknown income of a household in a specific education level - household size cell can be imputed by the average (known) income of the households in that cell. This method is called cell mean imputation (Lohr, 1999). For details see Appendix A. In cases where total household expenditure was greater than imputed total household income, income was made equal to expenditure. The values for per capita income were recalculated using the imputed total household income values whereafter the net profit was recalculated as the difference between total household income and total household expenditure. In cases where expenditure was more than the imputed income the value of net profit was taken as zero.

### 4.2.3 Comparing the two methods

This study will use only one of the above methods. Hence the more appropriate of the two methods needs to be identified. A quick look at the weighted population mean and total household income distribution was taken and the following results were obtained. Note that total household income, as given in the Revised IES dataset, was converted to a logarithmic scale in order to obtain a better view of income distribution. This was done to limit the weight of the extremely large income values.

**Method 1** The estimate of the population mean total household income using method 1 is obtained as:

$$\begin{aligned} \overline{y}_{str} &= \frac{\sum_{h=1}^{H} \sum_{j \in \mathbf{S}_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in \mathbf{S}_h} w_{hj}} \qquad (4.2.1) \\ &= \mathbf{R}\ 37512.1. \end{aligned}$$

Here

$w_{hj}$ is the household weight for household j in stratum h;

$y_{hj}$ is the total household income of household j in stratum h;

H is the number of strata; and

$\mathbf{S}_h$ is the set of all the households belonging to stratum h in the sample.

Figure 4.1 is a histogram of the log-transformed total household income using methods 1 & 2.

**Method 2** The estimate of the population mean household income using the imputation approach is obtained as:

$$\begin{aligned} \overline{y}_{str} &= \frac{\sum_{h=1}^{H}\sum_{j\in \mathbf{S}_h} w_{hj}y_{hj}}{\sum_{h=1}^{H}\sum_{j\in \mathbf{S}_h} w_{hj}} \qquad (4.2.2) \\ &= \mathbf{R}\ 37694.83. \end{aligned}$$

In Figure 4.1 the dashed line represents the histogram of the log-transformed total household income using the imputation approach.

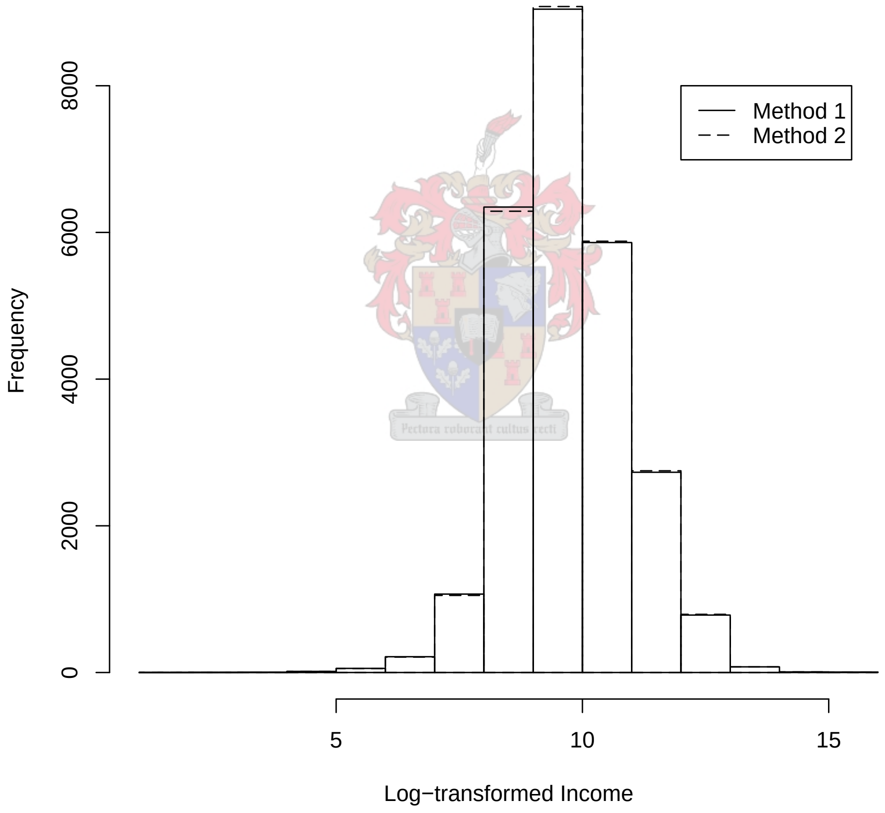


**Figure 4.1:** Histogram of the log-transformed total household income using both methods.

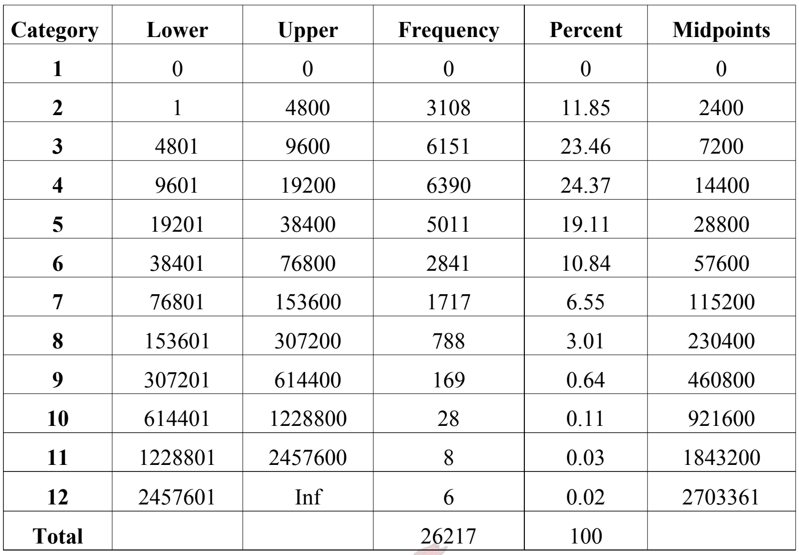
It is clear from the above that there is little difference between the two methods in terms of income distribution. Either one of the two is thus a good choice to continue with, although it still seems the better choice to take other variables into account. Hence the Imputation Approach will be used. It should however be mentioned that both these methods are only approximations to total household income and that further study is needed to more accurately impute values for zero income. It is however outside the scope of this study and will thus not be pursued further.

## 4.3 Creating a grouped income dataset

The debate surrounding continuous and grouped income variables has already been mentioned in chapter 3. But how do the results obtained from these two types of variables compare? To test this, a grouped income dataset is created from the Revised IES 2000 dataset. The impact of using either of these two types is tested using poverty and income inequality indicators. First a grouped income dataset needs to be created using the continuous dataset available. Only the income variables of the original dataset will be changed, while all other variables will remain unchanged. The grouped income dataset is created using annual total household income before tax and using the income intervals of Census 2001 to define the income levels.

In order to work with grouped income data, it first needs to be made continuous for the purposes of this analysis. The methods that will be used to assess poverty and inequality are based on continuous data. Various approaches to this problem exist, the main two being the midpoint method and the interval regression method. Both methods will be implemented, as well as a variation of the midpoint method, effectively creating four datasets to be analyzed. The first, the original continuous dataset, the second a continuous dataset using the midpoint method for analysis, the third also a continuous dataset but using the interval regression method and the fourth a continuous dataset using a random midpoint method. Table 4.1 gives the income levels together with the frequencies of households lying within each bracket and the midpoint for each bracket.

**Table 4.1:** Income Levels, Frequencies and Midpoints.



| Category | Lower | Upper | Frequency | Percent | Midpoints |
|---|---|---|---|---|---|
| **1** | 0 | 0 | 0 | 0 | 0 |
| **2** | 1 | 4800 | 3108 | 11.85 | 2400 |
| **3** | 4801 | 9600 | 6151 | 23.46 | 7200 |
| **4** | 9601 | 19200 | 6390 | 24.37 | 14400 |
| **5** | 19201 | 38400 | 5011 | 19.11 | 28800 |
| **6** | 38401 | 76800 | 2841 | 10.84 | 57600 |
| **7** | 76801 | 153600 | 1717 | 6.55 | 115200 |
| **8** | 153601 | 307200 | 788 | 3.01 | 230400 |
| **9** | 307201 | 614400 | 169 | 0.64 | 460800 |
| **10** | 614401 | 1228800 | 28 | 0.11 | 921600 |
| **11** | 1228801 | 2457600 | 8 | 0.03 | 1843200 |
| **12** | 2457601 | Inf | 6 | 0.02 | 2703361 |
| **Total** | | | 26217 | 100 | |

### 4.3.1 Midpoints

The midpoints method is simple and widely implemented by researchers (see for example Von Fintel, 2006). For this method, it is assumed that each person who supplies his/her income interval earns the interval midpoint. Since no upper bound exists for the top income level, it is assumed that the midpoint exceeds the lower bound by 10%.

The midpoint method was implemented using a program written in R/S-Plus (see Appendix B.1). It began by creating the grouped income dataset using the Census 2001 income levels. It then calculated the midpoint of each income level and assigned the midpoints to the households within each income level (see Appendix B.1). Table 4.1 gives the income levels, together with the frequency of households lying within each income level as well as the midpoint for each level. The midpoint for the top level in Table 4.1 is $2457601 \times 1.1 = 2703361$.

### 4.3.2 Interval regression

The second method used to obtain a continuous dataset from the grouped income dataset, is interval regression. Interval regression tries to fit a model to the grouped income dataset using some well chosen variables that will have an impact on the level of income each household receives. Using this model, it then predicts what income each household will have based on the variables used to fit the model. Thus, the interest is in household income, but variables relating to an individual, the head of the household, are used to predict the household income.

In order to use interval regression, we first need to create dummy variables for education level (edlev). This is done to indicate the level of education reached by the head of the household. Six dummy variables are created using the following definition for each:

- NOEDUC - Respondents having no education;
- PRIMARY - Respondents having primary school education or incomplete primary school education;
- INCSECOND - Respondents having an incomplete secondary school education or an NTC I or II certificate;
- MATRIC - Respondents having matric or an NTC III certificate;
- TERTIARY - Respondents having any form of tertiary education;
- MISSING - Respondents not specifying their education level or not knowing their education level.

The Mincerian Earnings Model will be used for specifying a model to be fitted. This model tries to predict what an individual's income will be based on his/her education and experience. It fits a model to the grouped income dataset and then predicts total household income using the following formula (Reilly, 2007):

$$LnY_i = b_0 + b_1 School_i + b_2 Exp_i + b_3 Exp_i^2 + e_i, \qquad (4.3.1)$$

where

$Y_i$ represents the grouped income of household i,

$School_i$ is the years of schooling of the head of household i,

$Exp_i$ represents experience in the labour market of the head of household i.

In the context of this study years of schooling will be taken as education level using the dummy variables defined above. Experience in the labour market will be approximated by the age and the squared age of the i-th individual. The regression formula thus becomes:

$$\begin{aligned} LnY_i &= b_0 + b_1 NOEDUC_i + b_2 PRIMARY_i + b_3 INCSECOND_i + \\ &\quad +b_4 MATRIC_i + b_5 TERTIARY_i + b_6 MISSING_i + \\ &\quad b_7 AGE_i + b_8 AGE_i^2 + e_i \end{aligned} \tag{4.3.2}$$

Next the model is fitted to the grouped income data and income is predicted using STATA/SE (see Appendix B.2). The household weights are taken into account when fitting the model. Table 4.2 contains the results of the interval regression. The coefficient for each variable is given as well as the standard error and 95% confidence interval. By looking at the weight (coefficient) each variable carries it is clear that tertiary education is most important in predicting income followed closely by whether an individual has matric or not.

It should, however, be said that this model does not give a good fit for the data. The fit was tested by grouping the predicted income and calculating the percentage of misfits, in other words, the percentage of predicted income intervals differing from the original income intervals. The model miss-predicted 71.22% of the income data. The model tends to under-predict the extremes of the data, the very small and very large incomes. If we assume that the interval regression uses the midpoint of the income interval to fit the model, the $R^2$ statistic can be estimated. This gives an indication of the fit of the model, with values lying between 0 and 1, where a value

**Table 4.2:** Interval Regression Results.

| | Coefficient | Stand.Error | z | P>z | 95% CI | |
|---|---|---|---|---|---|---|
| **primary** | .32 | .02 | 14.77 | .00 | .28 | .36 |
| **incsecond** | .78 | .02 | 31.29 | .00 | .73 | .83 |
| **matric** | 1.78 | .03 | 57.24 | .00 | 1.71 | 1.84 |
| **tertiary** | 2.72 | .06 | 47.81 | .00 | 2.61 | 2.83 |
| **missing** | .69 | .07 | 9.84 | .00 | .56 | .83 |
| **age** | .02 | .00 | 25.98 | .00 | .02 | .02 |
| **age2** | .00 | .00 | -25.39 | .00 | .00 | .00 |
| **constant** | 8.23 | .04 | 198.69 | .00 | 8.14 | 8.31 |

of 1 indicates a 100% fit. The formula for obtaining $R^2$ is:

$$\begin{aligned} R^2 &= 1 - \frac{\sum (X_i - Y_i)^2}{\sum (X_i - \bar{X})^2}, \qquad (4.3.3) \\ &= 0.13142857 \end{aligned}$$

where

$X_i$ is the midpoints for the income intervals;

$\bar{X}$ is the mean of the midpoints of the income intervals;

$Y_i$ is the predicted income value.

The ability of the interval regression dataset to predict the total household income is thus inadequate. This could explain the differences between this dataset and the original continuous and midpoint datasets, as seen later in this chapter.

### 4.3.3 Random midpoint dataset

Another method to create a continuous dataset is a variation of the midpoint method. The random midpoint method uses the midpoint of a income level and then distributes the households falling within the income level randomly across the level. Assuming that $f_i$ represents the frequency of households falling within income level $i$ and $x_i$ represents the midpoint of income level $i$, the following model is applied to obtain the random midpoint dataset:

$$Y_{ij} \;=\; x_i + sign_{ij} U_{ij}, \tag{4.3.4}$$

where

> $Y_{ij}$ is the new random midpoint income value for income level i and household j, j=1,2, $\ldots, f_i$;
>
> $x_i$ is the midpoint for income level i;
>
> $sign_{ij}$ is the sign for income level i and household j, where
>
> $$sign_{ij} = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2; \end{cases}$$
>
> $U_{ij} \;\sim\; Uniform(lowerbound_i, x_i)$;
>
> $lowerbound_i$ is the lower bound of income level i.

R/S-PLUS was used to obtain the continuous random midpoint dataset, for details see Appendix B.3.

## 4.4 Continuous versus grouped income

How do continuous data and data considered continuous but approximated from grouped income data compare when looking at general poverty and income inequality indicators? In this section an answer will be sought by looking at these indicators using each of the three datasets created above and comparing the results. The percentage of individuals below some well known poverty lines will be assessed. Extreme value theory will then be used to obtain thresholds and fit models to those individuals in the region of being extremely poor. Income inequality will be measured using the Gini coefficient. However, income will be taken as per capita and not as total household income as this gives more comparative results.

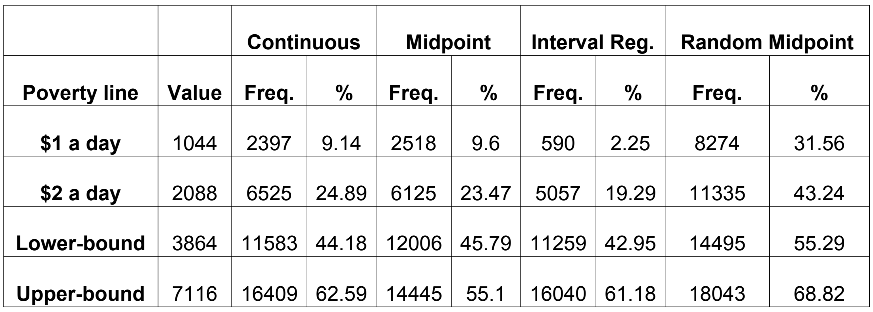
### 4.4.1 Poverty lines

In an article by Hoogeveen *&* Özler (2005) they use four well-known poverty lines to assess poverty in South Africa. These poverty lines are the \$1 a day, \$2 a day, lower-bound and upper-bound poverty lines. Hoogeveen *&* Özler (2005) obtained the \$1 and \$2 a day poverty lines by calculating the value of \$1 and \$2 in 2000 and multiplying it by the number of days in a month to obtain a montly poverty line. According to Ravallion (1994, 2001) a reasonable poverty line for South Africa, in terms of the cost of basic needs, must lie between R322 (lower-bound poverty line) and R593 (upper-bound poverty line) per capita per month in 2000 prices. Converting these monthly poverty lines to yearly income by multiplying them by 12 gives the following four respective poverty lines:

- \$1 a day = **R**1044
- \$2 a day = **R**2088
- lower-bound = **R**3864
- upper-bound = **R**7116

The poverty lines are all in 2000 rand values so as to relate to IES 2000 income values. For this section per capita household income was used for analysis, that is the total household income was divided by the size of the household. We first compare the three datasets in terms of the percentage of individuals lying below each of these poverty lines. These results are given in Table 4.3. Table 4.4 is similar to Table 4.3 but weighs the frequencies and percentages by the number of individuals per household. That is, for each household lying below a certain

**Table 4.3: Households** lying below the four poverty lines.



| | | Continuous | | Midpoint | | Interval Reg. | | Random Midpoint | |
|---|---|---|---|---|---|---|---|---|---|
| **Poverty line** | **Value** | **Freq.** | **%** | **Freq.** | **%** | **Freq.** | **%** | **Freq.** | **%** |
| **$1 a day** | 1044 | 2397 | 9.14 | 2518 | 9.6 | 590 | 2.25 | 8274 | 31.56 |
| **$2 a day** | 2088 | 6525 | 24.89 | 6125 | 23.47 | 5057 | 19.29 | 11335 | 43.24 |
| **Lower-bound** | 3864 | 11583 | 44.18 | 12006 | 45.79 | 11259 | 42.95 | 14495 | 55.29 |
| **Upper-bound** | 7116 | 16409 | 62.59 | 14445 | 55.1 | 16040 | 61.18 | 18043 | 68.82 |

poverty line, the number of individuals in the household is measured. Looking at the results obtained it is clear that the random midpoint method overestimates the number of households/individuals lying below the 1% and 2% a day poverty lines. This dataset is thus not useful for analyzing poverty and inequality in terms of poverty lines and extreme tail distributions. Hence, it will not be used further in this study as it does not yield meaningful results.

Before using these results to make statements about the percentage of poor in the country, it is necessary to test how good these estimates are in terms of confidence intervals. This can be done using two methods. The first is to obtain a formula for the confidence interval of the estimate and the second is using bootstrap techniques to obtain a standard error for the estimate and hence a confidence interval. It should, however, be mentioned that the two methods that are used to obtain the confidence intervals assume that the data was obtained through 'simple random sampling (SRS)'.

Table 4.4: **Individuals** lying below the four poverty lines.

| | | Continuous | | Midpoint | | Interval Reg. | | Random Midpoint | |
|---|---|---|---|---|---|---|---|---|---|
| **Poverty line** | **Value** | **Freq.** | **%** | **Freq.** | **%** | **Freq.** | **%** | **Freq.** | **%** |
| **$1 a day** | 1044 | 15071 | 14.49 | 16210 | 15.59 | 6643 | 6.39 | 32621 | 31.37 |
| **$2 a day** | 2088 | 38060 | 36.60 | 37177 | 35.76 | 38452 | 36.98 | 44706 | 43.00 |
| **Lower-bound** | 3864 | 60527 | 58.21 | 60235 | 57.93 | 67860 | 65.27 | 57384 | 55.19 |
| **Upper-bound** | 7116 | 77434 | 74.47 | 72679 | 69.9 | 82801 | 79.63 | 71308 | 68.58 |

It thus ignores the complex sampling used to obtain this data as well as the unequal weights in the dataset. The results are thus only approximations.

**Method 1: Approximation** Let $\hat{p}$ indicate the estimate of the proportion of individuals lying below a certain poverty line. It is assumed that $\hat{p}$ is approximately normally distributed with mean p (the actual proportion of individuals lying below a certain poverty line), and standard deviation $\frac{p(1-p)}{n}$, where n is the number of individuals in the sample. That is,

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \tag{4.4.1}$$

An approximate $(1-\alpha)$ confidence interval (CI) is then obtained from:

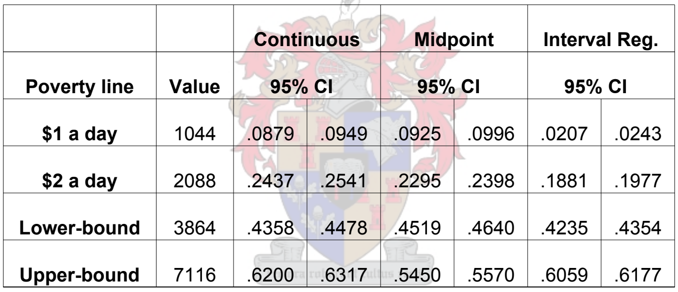
$$\begin{aligned} 1-\alpha &\doteq P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) \\ &\doteq P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right), \end{aligned}$$

$as:$

$$\left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right], \tag{4.4.2}$$

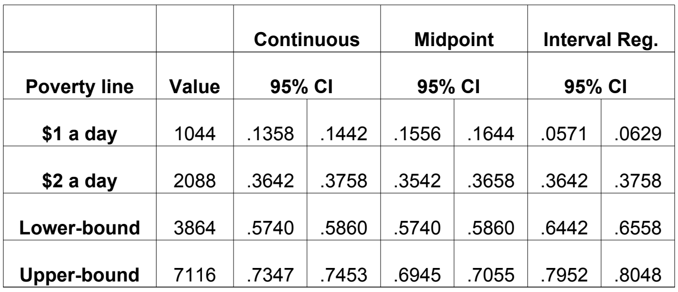
where in the standard deviation we also estimate p by $\hat{p}$. Table 4.5 contains the 95% confidence intervals for the estimated proportion of households lying below a certain threshold for each of the three datasets. Table 4.6 is similar to Table 4.5, but gives the confidence intervals for the estimated proportion of individuals lying below a certain poverty line.

**Table 4.5:** Confidence Intervals for **households** using method 1.



| | | Continuous | | Midpoint | | Interval Reg. | |
|---|---|---|---|---|---|---|---|
| Poverty line | Value | 95% CI | | 95% CI | | 95% CI | |
| $1 a day | 1044 | .0879 | .0949 | .0925 | .0996 | .0207 | .0243 |
| $2 a day | 2088 | .2437 | .2541 | .2295 | .2398 | .1881 | .1977 |
| Lower-bound | 3864 | .4358 | .4478 | .4519 | .4640 | .4235 | .4354 |
| Upper-bound | 7116 | .6200 | .6317 | .5450 | .5570 | .6059 | .6177 |

**Table 4.6:** Confidence Intervals for **individuals** using method 1.

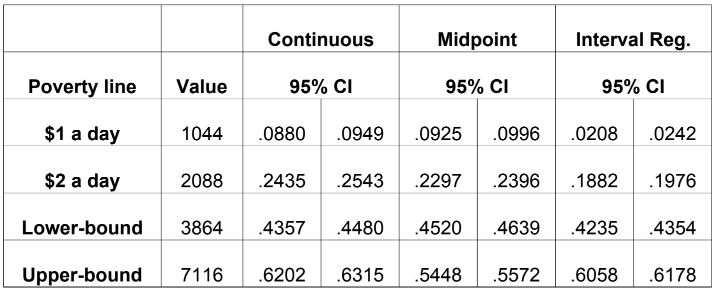


| | | Continuous | | Midpoint | | Interval Reg. | |
|---|---|---|---|---|---|---|---|
| Poverty line | Value | 95% CI | | 95% CI | | 95% CI | |
| $1 a day | 1044 | .1358 | .1442 | .1556 | .1644 | .0571 | .0629 |
| $2 a day | 2088 | .3642 | .3758 | .3542 | .3658 | .3642 | .3758 |
| Lower-bound | 3864 | .5740 | .5860 | .5740 | .5860 | .6442 | .6558 |
| Upper-bound | 7116 | .7347 | .7453 | .6945 | .7055 | .7952 | .8048 |

**Method 2: Bootstrap** The bootstrap method works by drawing a random sample with replacement from a dataset, where the sample is of the same size as the dataset. The parameter that was estimated for the original dataset is then estimated for the new sample. The process is repeated a large number of times, say B times. The B estimates of the parameter are then used to obtain an approximate standard error. This can then be used to approximate the confidence interval. R was used to carry out this bootstrap procedure. See Appendix C for details. Tables 4.7 and 4.8 contain respectively the confidence intervals for households and individuals using the bootstrap method.

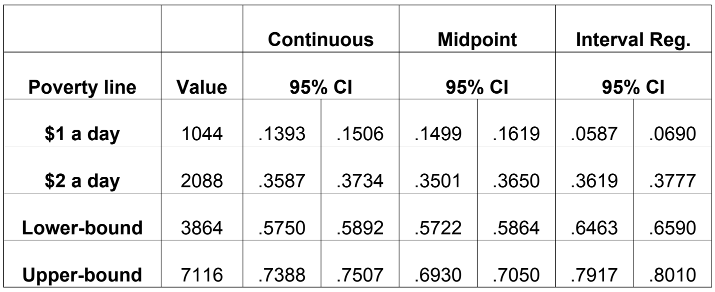
Clearly methods 1 and 2 give very similar confidence intervals. For both methods the confidence intervals obtained are quite short indicating high accuracy of estimation. Deductions can thus be made about the percentage of poor in the country. The continuous and the 'midpoint' datasets give similar results for the first three poverty lines, although they differ for the upper-bound poverty line. The interval regression dataset, on the other hand, is clearly out of line with these two methods. It overestimates the income of households or individuals in the region of being poor. This is not unexpected given the poor fit of the regression function obtained. By

**Table 4.7:** Confidence Intervals for **households** using Method 2.



| | | Continuous | | Midpoint | | Interval Reg. | |
|---|---|---|---|---|---|---|---|
| **Poverty line** | **Value** | **95% CI** | | **95% CI** | | **95% CI** | |
| **$1 a day** | 1044 | .0880 | .0949 | .0925 | .0996 | .0208 | .0242 |
| **$2 a day** | 2088 | .2435 | .2543 | .2297 | .2396 | .1882 | .1976 |
| **Lower-bound** | 3864 | .4357 | .4480 | .4520 | .4639 | .4235 | .4354 |
| **Upper-bound** | 7116 | .6202 | .6315 | .5448 | .5572 | .6058 | .6178 |

**Table 4.8:** Confidence Intervals for **individuals** using Method 2.



| | | Continuous | | Midpoint | | Interval Reg. | |
|---|---|---|---|---|---|---|---|
| **Poverty line** | **Value** | **95% CI** | | **95% CI** | | **95% CI** | |
| **$1 a day** | 1044 | .1393 | .1506 | .1499 | .1619 | .0587 | .0690 |
| **$2 a day** | 2088 | .3587 | .3734 | .3501 | .3650 | .3619 | .3777 |
| **Lower-bound** | 3864 | .5750 | .5892 | .5722 | .5864 | .6463 | .6590 |
| **Upper-bound** | 7116 | .7388 | .7507 | .6930 | .7050 | .7917 | .8010 |

looking at the percentages of individuals lying below the $1 and $2 a day poverty lines, it is seen that the values are dramatically smaller than those of the other two datasets. They compare reasonably for the other poverty lines.

### 4.4.2 Extreme tail estimation

Another method of analyzing poverty is by looking at the lower-tail of the income distribution. This is done here by using Extreme Value Theory. Per capita income was used, in other words, the total household income divided by the household size. The data was log-transformed to limit the effect of the extremely large values. The data was also made negative because it was more convenient to work with the maximum. The statistical software used is also limited to finding the maximum.

The threshold model will be used to analyze the data. It uses the Generalized Pareto Distribution to approximate the excess distribution. It is based on the assumption that the observations, $X_1$, $X_2$, $\ldots$, are a sequence of independent and identically distributed and the extreme events are those observations exceeding some high threshold $u$ (say). That is, for large enough $u$, the distribution function of (X-$u$), conditional on X $>$ $u$, is approximately: (Coles, 2001)

$$H(y) \;=\; 1 \;-\; (1+\xi y/\bar{\sigma})^{-1/\xi}, \tag{4.4.3}$$

where

$\bar{\sigma} = \sigma \;+\; \xi(u-\mu)$;

$\xi$ is the shape parameter; and

$\sigma$ is the scale parameter.

Mean residual life plots are used to find a reasonable threshold $u$. Above a threshold $u_0$ (say) at which the generalized Pareto distribution provides a valid approximation to the excess distribution, the mean residual life plot should be approximately linear in $u$. We thus choose our threshold as the value of $u$ above which the mean residual life plot is linear. Figure 4.2 is the mean residual life plot of the continuous income dataset. The figure was obtained using the R software package

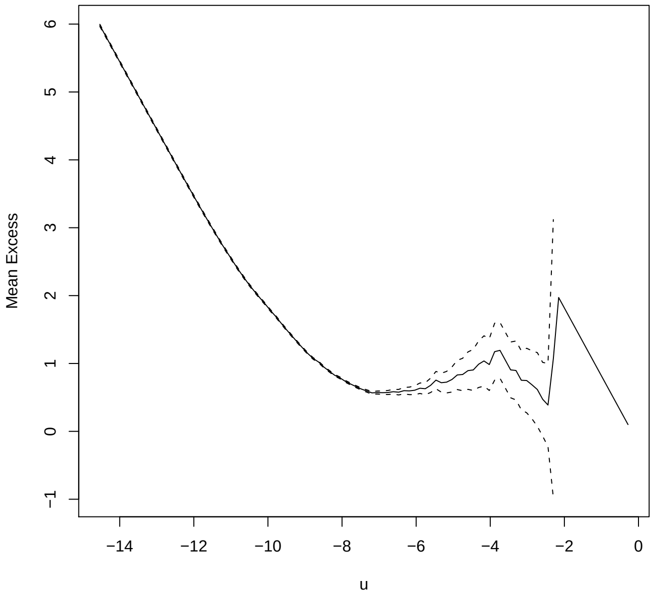


**Figure 4.2:** Mean residual life plot using the continuous income dataset.

'ismev' created by Coles & Stephenson (2006), all other calculations were also done using this software. The plot becomes approximately linear from $u = -6$ onwards until it becomes unstable. We thus took $u = -6$ as the threshold. This converts to approximately R403.50 per person per year. The rate of exceedences of the threshold $u$ can be calculated as the fraction of individuals earning less than the current threshold $u$. Hence, using the continuous income dataset we obtain an exceedance of $\frac{441}{26217} = 0.01682115$. Figure 4.3 is similar to Figure 4.2, but uses the grouped income midpoint dataset. Again looking at the figure it is clear that it starts being approximately linear at $u = -8$ giving a threshold of approximately R2981 per person per year. We again calculate the exceedences and obtain a rate of $\frac{9781}{26217} = 0.3730785$. Figure 4.4 represents the mean residual life plot of the interval regression dataset. The figure becomes linear at $u = -7.5$, thus individuals earning less than R1808 per year. This relates to an exceedance rate of $\frac{3710}{26217} = 0.1415112$.

The Generalized Pareto Model can now be fitted using these thresholds. Return levels, the probability that an individual will earn less than a certain value, can be calculated. Extreme quantiles can also be calculated. Extreme quantiles determine

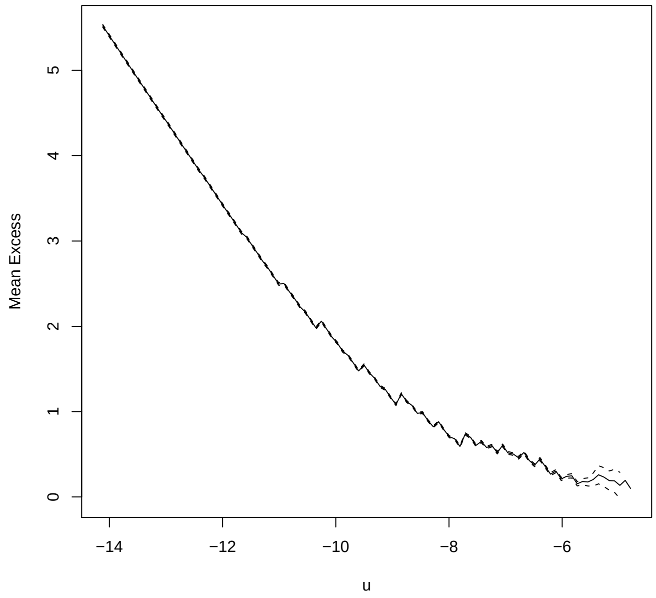


**Figure 4.3:** Mean residual life plot using the midpoint income dataset.

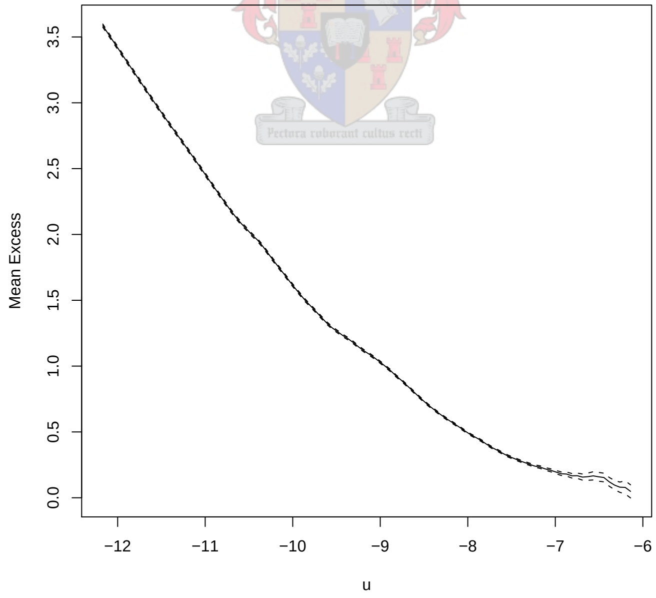


**Figure 4.4:** Mean residual life plot using the interval regression income dataset.

the value below which a certain proportion of the individuals lie. For calculating extreme quantiles Coles (2001) first writes:

$$Pr(X > x) \;\doteq\; \zeta_u \left[1 + \xi \left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi}, \tag{4.4.4}$$

and for this equal to $\alpha$ (say), then solves the right hand side for x, to obtain:

$$x_\alpha \;\doteq\; u + \sigma log(\frac{1}{\alpha}\zeta_u), \tag{4.4.5}$$

where

$\zeta_u = Pr(X \geq u).$

It should be mentioned that the generalized Pareto model is only applicable for income values lying below the respective threshold. Hence, it will not be possible to estimate the probability of individuals to have less income than the four poverty lines defined above. The values to be tested should lie below the threshold, but we also want to compare the results from the three datasets. A value will thus be chosen that is below all three thresholds. Table 4.9 contains the probabilities for individuals to have less income per year than the three given levels. The continuous income dataset has the highest probabilities of individuals lying below these values. The midpoint dataset's probabilities are smaller than those of the continuous income dataset, while for the interval regression dataset the probabilities are all zero. The conclusion can thus be made that the interval regression dataset overestimates the income of individuals lying in the extreme tail of the income distribution. Estimating $\zeta_u$, $\sigma$ and $\xi$ in the above, we obtain:

$$\hat{H}^{-1}(1 - \alpha) \;=\; u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(\frac{N}{n\alpha}\right)^{-\hat{\xi}} - 1\right], \tag{4.4.6}$$

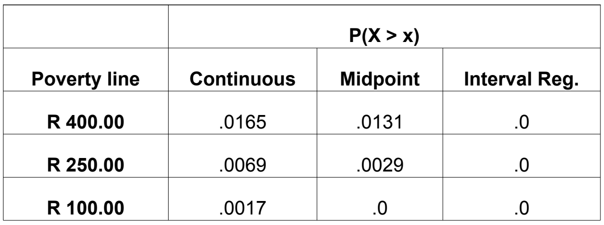
where

N is the number of observations; and

n is the number of observations exceeding the threshold $u$.

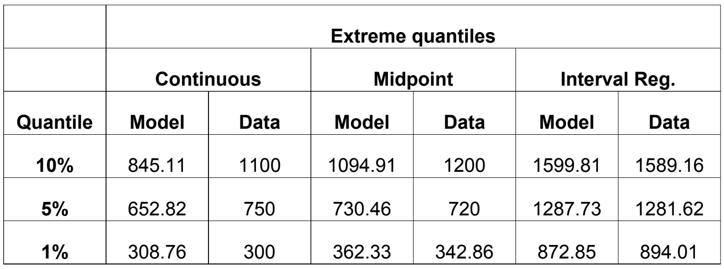
Table 4.10 contains the 1%, 5% and 10% quantiles for the three datasets. Each value

**Table 4.9:** Exceedance probabilities for individual yearly income.



| | P(X > x) | | |
|---|---|---|---|
| **Poverty line** | **Continuous** | **Midpoint** | **Interval Reg.** |
| **R 400.00** | .0165 | .0131 | .0 |
| **R 250.00** | .0069 | .0029 | .0 |
| **R 100.00** | .0017 | .0 | .0 |

**Table 4.10:** Extreme quantiles for individual yearly income.



| | Extreme quantiles | | | | | |
|---|---|---|---|---|---|---|
| | **Continuous** | | **Midpoint** | | **Interval Reg.** | |
| **Quantile** | **Model** | **Data** | **Model** | **Data** | **Model** | **Data** |
| **10%** | 845.11 | 1100 | 1094.91 | 1200 | 1599.81 | 1589.16 |
| **5%** | 652.82 | 750 | 730.46 | 720 | 1287.73 | 1281.62 |
| **1%** | 308.76 | 300 | 362.33 | 342.86 | 872.85 | 894.01 |

represents the income level below which x% of the sample lies. The quantiles were obtained by fitting the model to individuals earning less than the threshold. For comparison the sample quantiles were also obtained for each dataset as well. It is clear that the 'midpoint' and 'interval regression' data correspond closely to these. However, for the continuous data there is quite a difference at the 10% and 5% levels. This is a result of the small number of individuals earning below the chosen threshold. The continuous and 'midpoint' datasets do, however, give comparible sample quantiles. The trend for the interval regression method to overestimate the income of individuals is again clear when looking at the extreme quantiles. The values obtained are much higher than those of the other two datasets.

### 4.4.3 Income inequality

Income inequality is measured using the standard income inequality indicators. We used the total household income to measure income inequality between households. The Gini coefficient, Theil index and Mean log deviation are calculated for each of the three datasets defined above. The Gini coefficient is calculated using the Lorenz curve; it varies between 0 (when there is perfect equality) and 1 (when there is perfect inequality). The Theil index and Mean log deviation belong to the GE class of measurements described in section 2.5.2. A higher value for these two measurements imply a higher level of inequality. The indicators and Lorenz curves were obtained using the statistical software STATA and the package "Measures of Inequality" created by Whitehouse (1995) in STATA.

Figures 4.5, 4.6 and 4.7 represent the Lorenz curves for the continuous, midpoint and interval regression income data. Table 4.11 contains the income inequality measures for the three datasets. The income inequality indicators for the continuous and 'midpoint' datasets are very similar. The interval regression dataset again shows signs of overestimating income for individuals in the region of being poor. In a study by Hoogeveen *&* Özler (2005), also on the IES 2000 dataset, they found the Gini coefficient of expenditures to be 0.56. Since income and expenditure are highly correlated, it is quite reasonable to have a Gini coefficient of income of 0.63.

The Gini coefficient can be compared with that of other countries. This gives an

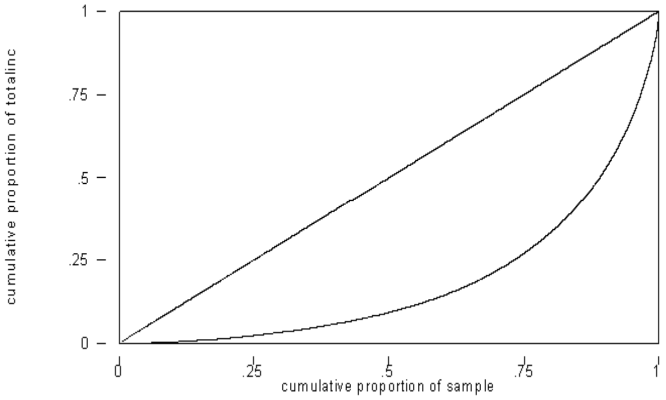


**Figure 4.5:** Lorenz curve using the continuous income data.

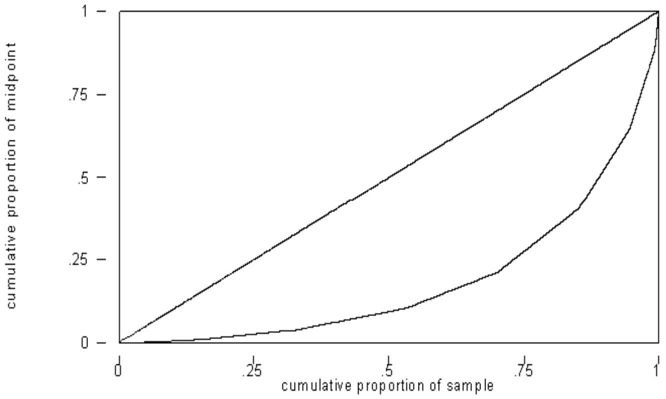


**Figure 4.6:** Lorenz curve using the midpoint income data.

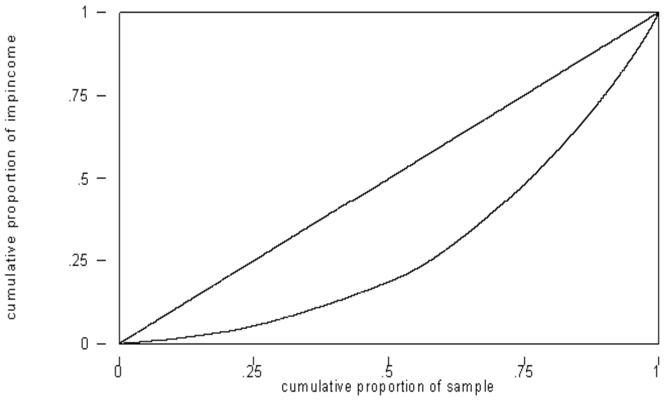
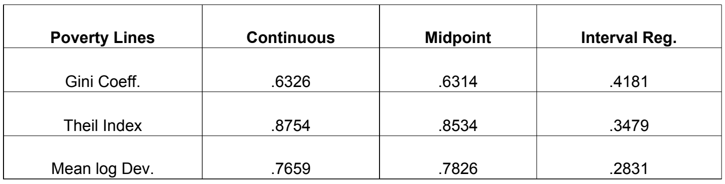


**Figure 4.7:** Lorenz curve using the interval regression income data.

**Table 4.11:** Income Inequality Indicators.



| Poverty Lines | Continuous | Midpoint | Interval Reg. |
|---|---|---|---|
| Gini Coeff. | .6326 | .6314 | .4181 |
| Theil Index | .8754 | .8534 | .3479 |
| Mean log Dev. | .7659 | .7826 | .2831 |

indication of how unequal income is distributed in South Africa relative to other countries. Table 4.12 includes the Gini coefficients for a range of countries as given by the United Nations (2006) for measuring income inequality. The Gini coefficients are measured for individuals. It can be seen that South Africa has one of the highest Gini coefficients for income inequality between individuals.

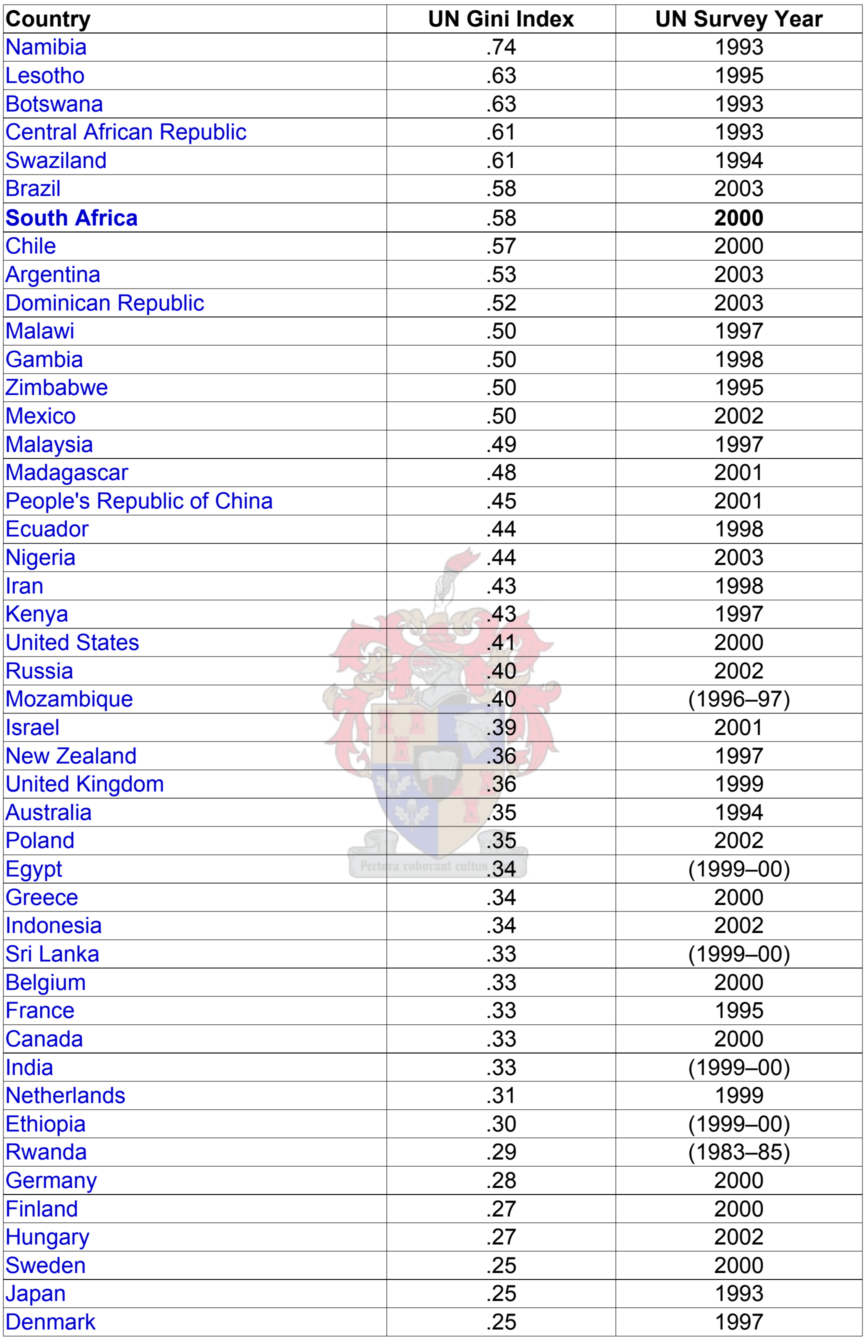
## 4.5 Provincial poverty and inequality

There are clearly differences of poverty and income inequality between provinces. Hence it is also important to study the distribution for the provinces individually and to compare the results. For this section only the continuous income dataset will be used. The same three techniques from the previous section will be used to analyze poverty and income inequality per province.

### 4.5.1 Poverty lines per province

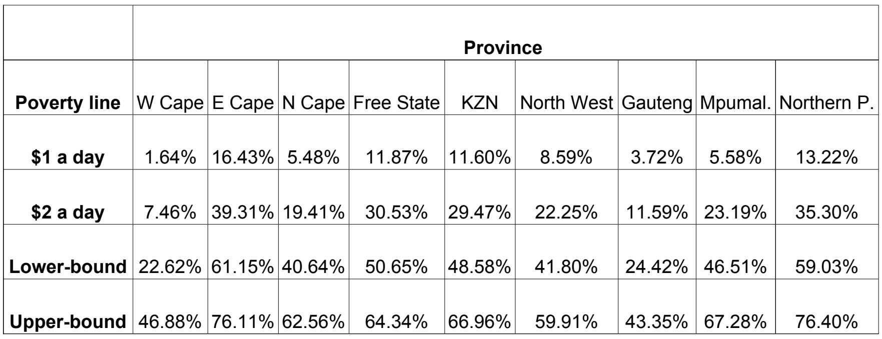
The same four poverty lines of the previous section were used to assess poverty per province. The poverty lines were evaluated against the per capita yearly household income. Table 4.13 gives the percentage of households earning less than a certain poverty line per province. Per capita income was used and the four poverty lines are those of the previous section. The Western Cape seems to be the province with the least poverty, followed closely by Gauteng. The Northern Province and Eastern Cape, on the other hand, have the highest rates of poverty.

**Table 4.12:** Gini Coefficients for a number of countries.



| Country | UN Gini Index | UN Survey Year |
|---|---|---|
| Namibia | .74 | 1993 |
| Lesotho | .63 | 1995 |
| Botswana | .63 | 1993 |
| Central African Republic | .61 | 1993 |
| Swaziland | .61 | 1994 |
| Brazil | .58 | 2003 |
| **South Africa** | .58 | **2000** |
| Chile | .57 | 2000 |
| Argentina | .53 | 2003 |
| Dominican Republic | .52 | 2003 |
| Malawi | .50 | 1997 |
| Gambia | .50 | 1998 |
| Zimbabwe | .50 | 1995 |
| Mexico | .50 | 2002 |
| Malaysia | .49 | 1997 |
| Madagascar | .48 | 2001 |
| People's Republic of China | .45 | 2001 |
| Ecuador | .44 | 1998 |
| Nigeria | .44 | 2003 |
| Iran | .43 | 1998 |
| Kenya | .43 | 1997 |
| United States | .41 | 2000 |
| Russia | .40 | 2002 |
| Mozambique | .40 | (1996–97) |
| Israel | .39 | 2001 |
| New Zealand | .36 | 1997 |
| United Kingdom | .36 | 1999 |
| Australia | .35 | 1994 |
| Poland | .35 | 2002 |
| Egypt | .34 | (1999–00) |
| Greece | .34 | 2000 |
| Indonesia | .34 | 2002 |
| Sri Lanka | .33 | (1999–00) |
| Belgium | .33 | 2000 |
| France | .33 | 1995 |
| Canada | .33 | 2000 |
| India | .33 | (1999–00) |
| Netherlands | .31 | 1999 |
| Ethiopia | .30 | (1999–00) |
| Rwanda | .29 | (1983–85) |
| Germany | .28 | 2000 |
| Finland | .27 | 2000 |
| Hungary | .27 | 2002 |
| Sweden | .25 | 2000 |
| Japan | .25 | 1993 |
| Denmark | .25 | 1997 |

**Table 4.13:** Households lying below the four poverty lines per province.



| | Province | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Poverty line** | W Cape | E Cape | N Cape | Free State | KZN | North West | Gauteng | Mpumal. | Northern P. |
| **$1 a day** | 1.64% | 16.43% | 5.48% | 11.87% | 11.60% | 8.59% | 3.72% | 5.58% | 13.22% |
| **$2 a day** | 7.46% | 39.31% | 19.41% | 30.53% | 29.47% | 22.25% | 11.59% | 23.19% | 35.30% |
| **Lower-bound** | 22.62% | 61.15% | 40.64% | 50.65% | 48.58% | 41.80% | 24.42% | 46.51% | 59.03% |
| **Upper-bound** | 46.88% | 76.11% | 62.56% | 64.34% | 66.96% | 59.91% | 43.35% | 67.28% | 76.40% |

### 4.5.2 Assessing the extreme events by province

Extreme value theory is again used to asses the lower-tail distribution of income, but now per province. The relevant threshold, $u$, is obtained for each province and the Generalized Pareto Model is fit to the data. From this the extreme quantiles and extreme probabilities can be calculated as in section 4.4.2.

Figure 4.8 gives the mean residual life plots for all 9 provinces. These plots are used to obtain the thresholds ($u$) and fit the generalized Pareto model. The threshold is again identified as the point, u, at which the mean residual life plot becomes approximately linear. The 5% and 1% quantiles were obtained for each province.

Table 4.14 contains the thresholds for each province, as well as the 5% and 1% quantiles. It is clear that the Western Cape and Gauteng provinces have the highest quantiles. This implies that on average a poor person in these two provinces will earn more than a poor person in the other seven provinces. Table 4.15 gives the

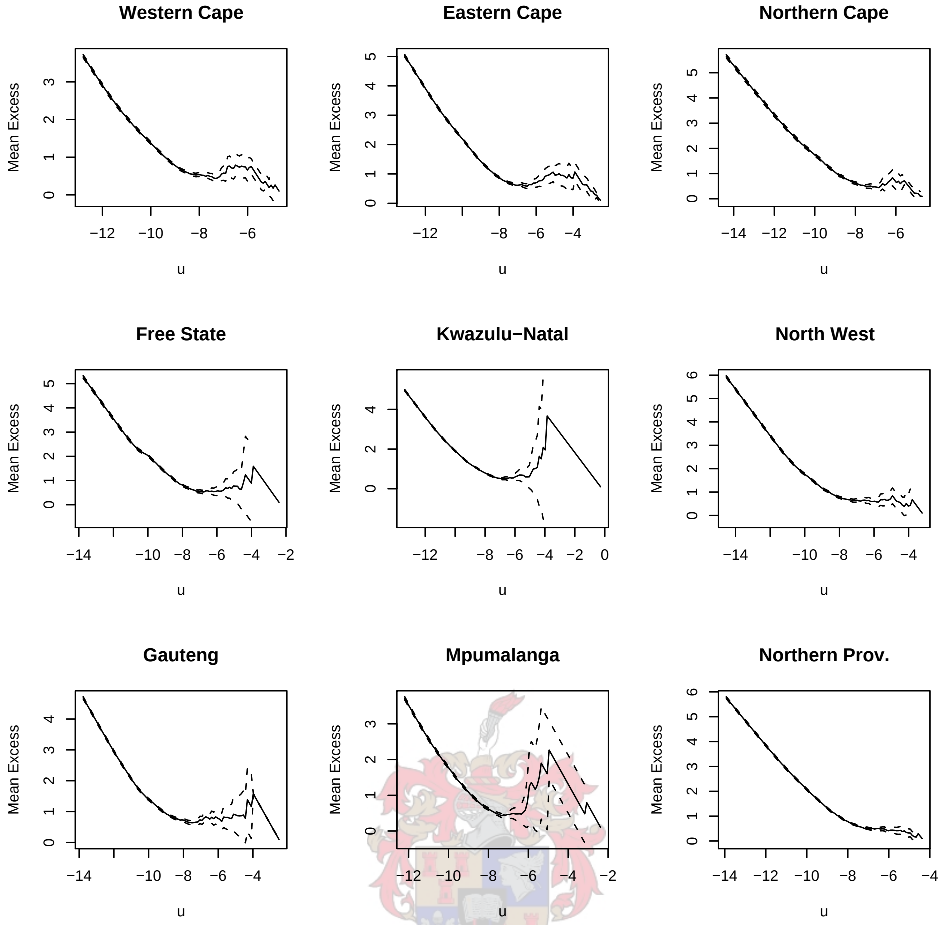


**Figure 4.8:** Mean residual life plots for all nine provinces.

probabilities that household will earn less than R1000 and R500 rand respectively. Again the Western Cape and Gauteng have the lowest probabilities for households to earn less than these two benchmarks.

**Table 4.14:** Thresholds and Extreme Quantiles per Province.

| | | | 5% Quantile | | 1% Quantile | |
|---|---|---|---|---|---|---|
| Province | Threshold | P(X<u) | Model | Data | Model | Data |
| Western Cape | 2980.96 | 0.14 | 1682.47 | 1620 | 718.41 | 926.46 |
| Eastern Cape | 1808.04 | 0.34 | 470.97 | 514.29 | 165.39 | 200 |
| Northern Cape | 8103.08 | 0.65 | 739.15 | 991.46 | 325.69 | 535.69 |
| Free State | 1096.63 | 0.13 | 646.69 | 660 | 261.06 | 289.23 |
| Kwazulu-Natal | 1808.04 | 0.26 | 658.71 | 685.71 | 259.84 | 300.06 |
| North West | 2980.96 | 0.33 | 678.36 | 740.75 | 244.95 | 283.66 |
| Gauteng | 1096.63 | 0.04 | 1289.69 | 1227.4 | 414.12 | 400 |
| Mpumalanga | 1808.04 | 0.19 | 988.03 | 1007.49 | 453.09 | 496.36 |
| Northern Prov. | 1096.63 | 0.14 | 638.99 | 650.06 | 297.78 | 300 |

**Table 4.15:** Exceedance probabilities for households.

| | | | P(X > x) | |
|---|---|---|---|---|
| Province | Threshold | P(X<u) | R 1,000.00 | R 500.00 |
| Western Cape | 2980.96 | 0.14 | .0189 | .0049 |
| Eastern Cape | 1808.04 | 0.34 | .0795 | .0290 |
| Northern Cape | 8103.08 | 0.65 | .0782 | .0254 |
| Free State | 1096.63 | 0.13 | .1099 | .0316 |
| Kwazulu-Natal | 1808.04 | 0.26 | .0998 | .0313 |
| North West | 2980.96 | 0.33 | .0855 | .0319 |
| Gauteng | 1096.63 | 0.04 | .0342 | .0129 |
| Mpumalanga | 1808.04 | 0.19 | .0513 | .0122 |
| Northern Prov. | 1096.63 | 0.14 | .1212 | .0302 |

### 4.5.3 The distribution of income within provinces

The income inequality within provinces can be measured using the Gini coefficient. The Gini coefficient measures income inequality between individuals. The Theil index and Mean log deviation are also given for reference. Table 4.16 gives the three income inequality measures per province. The Western Cape has the lowest Gini coefficient, indicating the most equal distribution of income within that province. However, the value obtained is still very high. The Northern Cape has the most unequal distribution of income.

**Table 4.16:** Income Inequality measures per province.

| Province | Gini Coeff. | Theil Index | Mean Log Dev. |
|---|---|---|---|
| Western Cape | .5638 | .5904 | .5861 |
| Eastern Cape | .6434 | .8619 | .8017 |
| Northern Cape | .6795 | 1.0665 | .8991 |
| Free State | .6785 | 1.2264 | .9109 |
| Kwazulu- Natal | .6201 | .7982 | .7180 |
| North West | .5888 | .8086 | .6615 |
| Gauteng | .6081 | .7825 | .7025 |
| Mpumalanga | .5687 | .6460 | .5722 |
| North Province | .6581 | 1.0903 | .8108 |

## 4.6 Conclusions

This chapter set out to compare results obtained from continuous income variables and grouped income variables. But first the dataset had to be cleaned in respect of missing values and unrealistic zero income. Two methods were used to deal with zero income. They were compared and the imputation approach was chosen as the method to be used in the analysis. The IES 2000 dataset only contained a continuous income variable and a grouped income variable was thus created from this continuous variable using the income levels defined in Census 2001. The methods of analysis used in this chapter relied on a continuous variable. This meant that the newly created grouped income variable needed to be made continuous again. This was done using three methods, namely the midpoint method, the interval regression method and the random midpoint method. It was clear from early analyses that the random midpoint dataset gave a bad approximation to the continuous income variable. That dataset was thus dropped from further analysis.

There were then three datasets to be analyzed, the continuous income dataset, the midpoint dataset and the interval regression dataset. We first looked at poverty lines and used four well-known poverty lines, the \$1 a day, \$2 a day, lower-bound and upper-bound poverty lines. The yearly per capita household income was measured against these poverty lines and the percentages of households lying below these thresholds were obtained for all three datasets. This was repeated using the number of individuals in a household as weights, hence obtaining the percentages of individuals lying below these four poverty lines. Confidence intervals were obtained for these results. These were obtained through two methods, by approximation and by bootstrap methods. Both methods gave very similar results, the confidence intervals were very short, implying a high accuracy of estimation. The continuous income and midpoint dataset gave comparable results, but the interval regression dataset underestimated the number of households/individuals lying below the first two poverty lines.

Next, the extreme tails of the three datasets were approximated by using Extreme Value Theory. A threshold was obtained for each dataset below which the Generalized Pareto Distribution was fitted to the data. Using this model, exceedance probabilities and extreme quantiles were then estimated for each dataset. The continuous and midpoint datasets gave similar results, but again the underestimation by the interval regression method was evident. These quantiles were also compared to the sample quantiles with good correspondence in some of the cases. Income inequality between households were compared between the three datasets. The Gini coefficient, Theil index and log mean deviation were calculated. Again, the midpoint and continuous income datasets gave very similar results. The interval regression dataset, however, was out of line with the previous two and gave an unrealistically low value.

On a more practical level, the distributions of poverty and income inequality between the provinces of South Africa were analyzed. This was done using only the continuous income dataset. The same three aspects of poverty and inequality as above were evaluated. Whether using poverty lines, the extreme tail distribution or income inequality measures, the same conclusions were drawn. The Western Cape and Gauteng are the wealthiest of the 9 provinces with the most equal distribution of income, although still highly unequal. The Eastern Cape is the worst off by far, with 5% of its households having less than R514 per capita income per year.

Having obtained all these results, what can be inferred about continuous or grouped income variables? From an analytical point of view, it would seem that the type of income variable used does not make that much of a difference in the results. The two factors that will have the biggest impact are firstly, the size of the income level, they should be small enough to capture the nature of the extreme events. The second factor is the method used to obtain a continuous dataset from the grouped income dataset, or to use methods of analysis that are compatible with grouped income data. Form a practical point of view the use of grouped income variables have the advantage that individuals are more likely to give their income in this form than as an exact amount. This will lead to a more reliable response in surveys. The optimal solution is to combine the two, give individuals the choice between giving either an exact income value, indicating in which income level they fall or indicate both. In this way more information can be gathered and better results obtained. This is because the income level data is more reliable, persons are more likely to indicate their correct income bracket, while the continuous income variable contains vital information for fitting a model to predict income.

# Chapter 5

# Multivariate analysis through biplots

Chapter 4 analyzed poverty and inequality using three datasets and then compared the results. In this chapter all 5 types of deprivation will be considered - monetary welfare will be measured against the other variables and those variables will be identified giving an indication of poverty.

The analysis in this chapter is done using total yearly household income. A description is given of the creation of the appropriate dataset to be used for plotting biplots. This is followed by a brief summary of the theory behind principle components analysis (PCA) and biplots are then drawn for the dataset using PCA biplots. Canonical variate analysis (CVA) is also used. The biplots are drawn for groups specified by race, province or area type (urban/rural).

## 5.1 Why biplots?

Gower & Hand (1996) says the following about biplots:

> Biplots are the multivariate analogue of scatter plots. They approximate the multivariate distribution of a sample in a few dimensions, typically two, and they superimpose on this display representations of the variables on which the samples are measured. In this way, the relationships between the individual sample points can be easily seen and,

> as we shall see, they can also be related to values of the measurements. Thus, like scatter plots, biplots are useful for giving a graphical description of the data, for detecting patterns, and for displaying results found by more formal methods of analysis.

It is this detecting of patterns that is of most interest. Valuable conclusions can be made about multivariate data without using complicated mathematical techniques. It also has the additional advantage of being easy to interpret, thus making the information available even to individuals without a statistical background. Biplots are especially convenient for the IES 2000 dataset in the same sense that it can show us the variables that are highly correlated as well as those variables that will have an impact on a household's income or expenditure.

## 5.2 Creating an appropriate dataset

Before any biplots can be used on the dataset it has to be cleaned in an appropriate manner. The decision of what variables should be included has to be made. The choice of variables included those falling into the categories of *income and material deprivation*, *education deprivation* and *living environment deprivation*. The following variables were included in the dataset:

- PROVINCE
- AREATYPE
- AGE
- RACE
- EDUCATION LEVEL
- HOUSEHOLD SIZE
- TOTAL HOUSEHOLD EXPENDITURE
- TOTAL HOUSEHOLD INCOME

Statistics SA used codes to indicate missing values. In order to overcome any bias this may introduce in the biplot, all missing values were recoded as 'NA'. When the actual biplot is drawn all missing values will be omitted. Another change made to the original dataset concerns the education level. In this chapter when there is reference to the education level of a household it will be taken as the education level of the head of the household. The categories used to create the dummy variable in the interval regression approach were used here as well to code education level into one of the following groups:

- 1 - No Education;
- 2 - Primary or incomplete primary school education;
- 3 - Incomplete secondary school education or NTC I or II certificates;
- 4 - Matric or NTC III certificate;
- 5 - Tertiary education;
- NA - Missing value or households that do not know their education level.

Hence, the dataset is created and all missing values are omitted. This gives a dataset of 25780 households or observations, which includes more than 98% of the original dataset.

## 5.3 Principle component analysis (PCA)

Principle components analysis is a dimension reducing technique. It is used to optimize the variation between the observations/households. Let $\mathbf{X}$ represent the dataset, hence it is of dimension n×p, where $n = 25780$ and $p = 8$. Each row of the matrix $\mathbf{X}$ represents a household and the columns represent the eight variables defined above. PCA takes the rows of the matrix $\mathbf{X}$ to give the coordinates of the n samples in a p-dimensional space $\mathcal{R}_p$. The distance $d_{ij}$ between a pair of points is given by Pythagoras' theorem and is referred to as a Euclidean distance. PCA chooses the $\rho$-dimensional subspace $\mathcal{L}$ of the p-dimensional space $\mathcal{R}$ that is best fitting in the least squares sense.

The subspace is found to be spanned by the first $\rho$ principle eigenvectors of $\mathbf{X}'\mathbf{X}$, namely $\mathbf{V}_\rho$. These $\rho$ eigenvectors define the set of orthogonal coordinate axes for the $\rho$-dimensional subspace. Relative to this, the coordinates of the projections of the samples onto the subspace $\mathcal{L}$ is given as $\mathbf{Z} = \mathbf{X}\mathbf{V}_\rho$ (Gower *&* Hand, 1996). All the biplots were obtained using programs written by Le Roux (2006) in R/S-PLUS.

### 5.3.1 PCA biplots grouped by race

PCA biplots were plotted for the dataset. Figure 5.1 gives all the households; the colours were coded by Race, though it had no further effect on the analysis. Figure 5.2 gives the mean for each race, as well as the 90% alpha bags. These bags are drawn so that 90% of the observations fall within the bag.

Race is coded as follows:

1 - Black

2 - Coloured

3 - Asian

4 - White

From the biplots some very interesting and important conclusions can be made. The fact that the LogInc and LogExp axes lie almost on top of each other points at a very high correlation between the two, which is to be expected since income and expenditure are linked. The white population of South Africa has the highest income and expenditure. It also has the highest level of education. The Asian population is the closest to the white population, while the Coloured population lies between the white and the black population. The black population has the lowest income and expenditure, as well as the lowest level of education. This points at a correlation between education level and the level of income and expenditure of a household. A point in the direction of the white population is clearly visible in the 90% alpha bag of the black population; this could be as a result of the upcoming black middle-class.

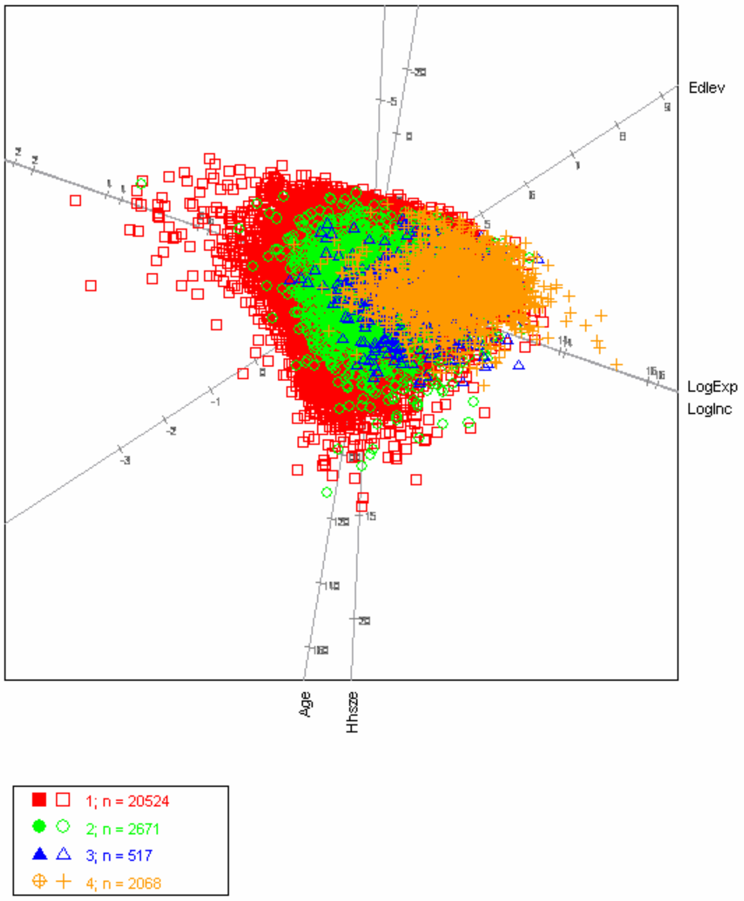


**Figure 5.1:** PCA Biplot by race.

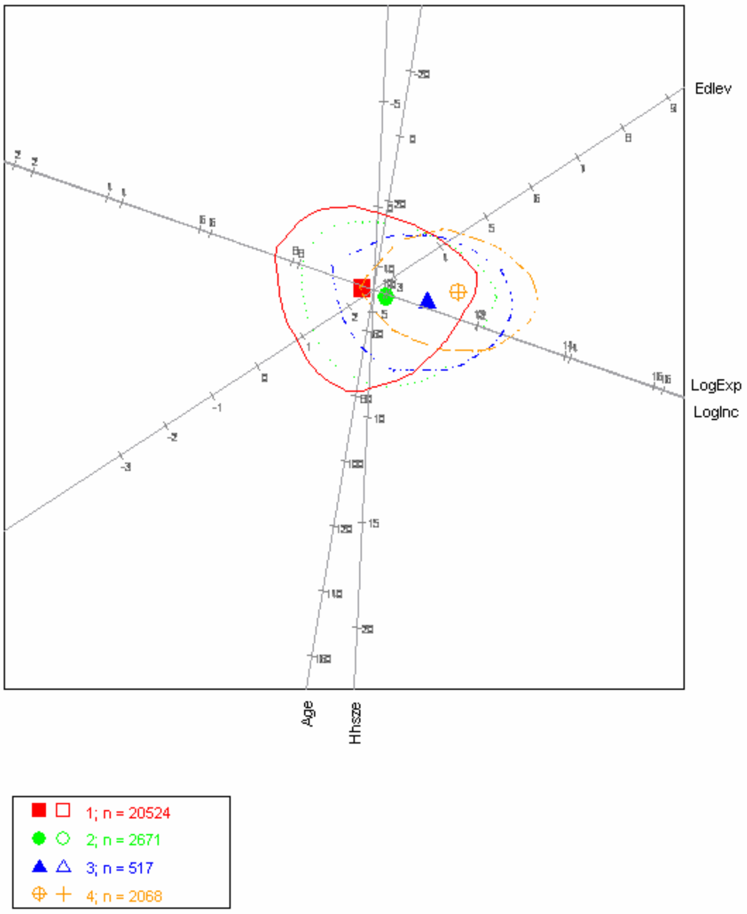


**Figure 5.2:** PCA Biplot with 90% bags.

## 5.4 Canonical variate analysis (CVA)

Canonical variate analysis is used to separate groups optimally. CVA is based on a dataset, say $\mathbf{X}_{n\times p}$, where $n$ is the number of observations/households and $p$ is the number of variables, that is partitioned into $g$ groups. In this study the $g$ groups will be represented by race, province and area (urban/rural). All the biplots were obtained using programs written by Le Roux (2006) in R/S-PLUS. CVA biplots were used for the remainder of this chapter since its property of optimal group separation is what we are interested in.

### 5.4.1 CVA biplots grouped by race

CVA biplots were plotted by race, but now race was taken into account when analyzing the data. The codes for race are the same as above, namely:

1 - Black

2 - Coloured

3 - Asian

4 - White

Figure 5.3 represents the population in groups of race, while Figure 5.4 gives the mean and the 90% alpha bags for the population in groups of race. The same pattern emerges as with the PCA biplots. The white population has greater income and expenditure, as well as higher education level. The black population has the lowest income and expenditure, as well as the lowest education level. The Asian population lies closest to that of the white population, with the Coloured population lying between the Asian and black population.

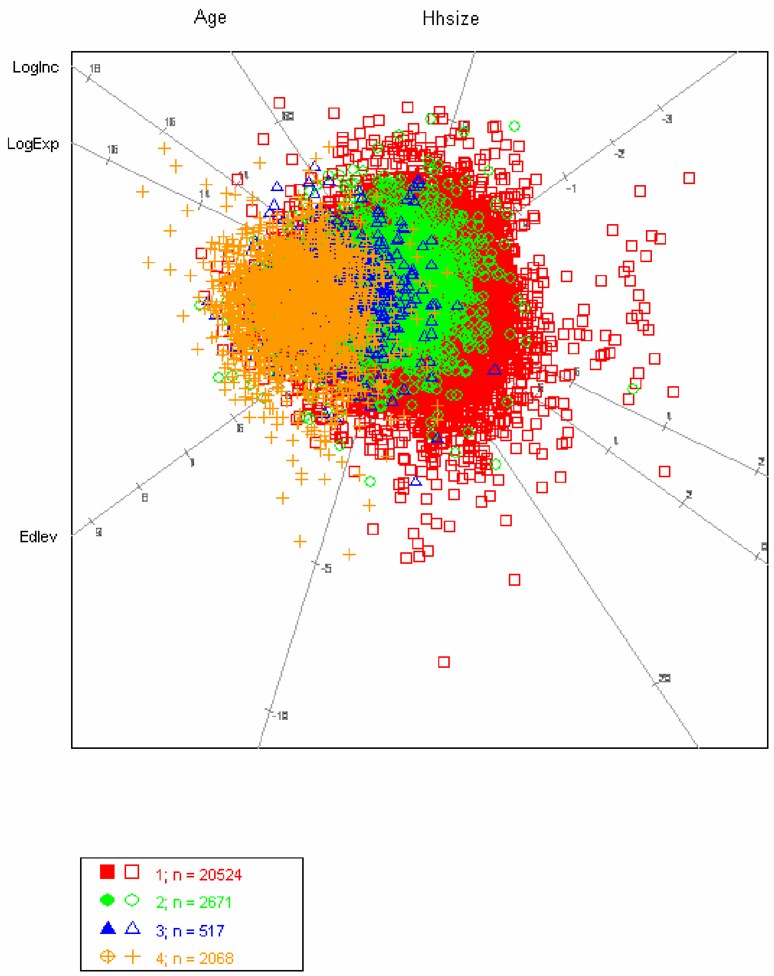


**Figure 5.3:** CVA Biplot by race.

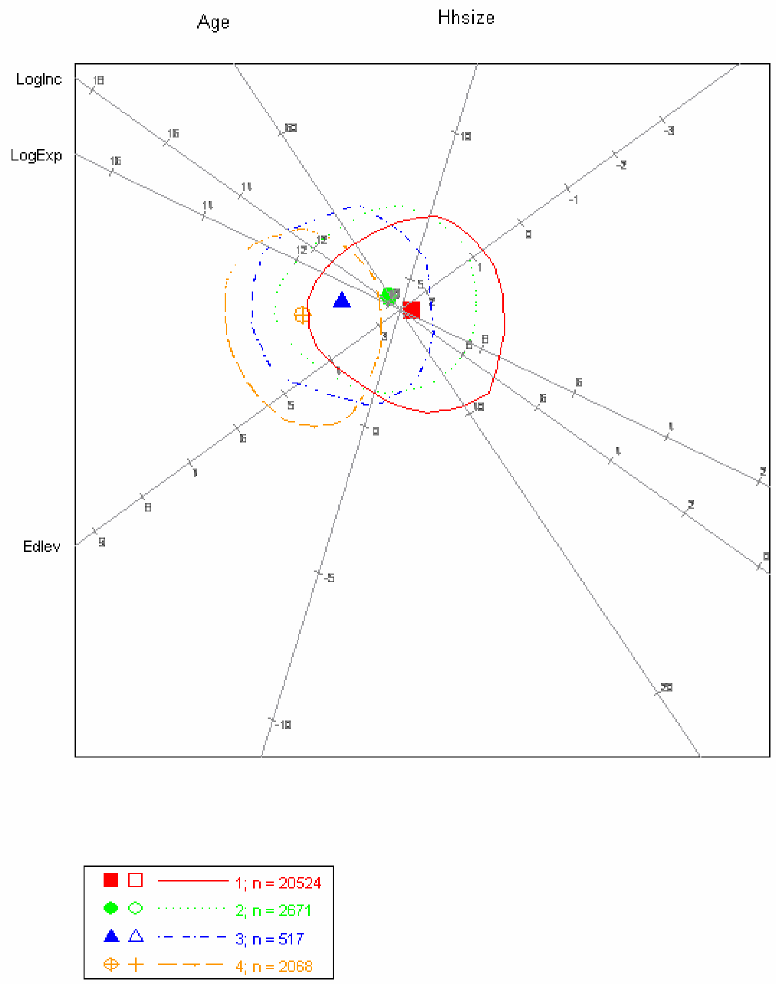


**Figure 5.4:** CVA Biplot with 90% bags.

### 5.4.2 CVA biplots grouped by province

CVA biplots were drawn for the data with the groups defined to be the provinces. This will give an indication of the welfare of the respective provinces. Provinces were coded as:

1 - Western Cape

2 - Eastern Cape

3 - Northern Cape

4 - Free State

5 - Kwazulu-Natal

6 - North West

7 - Gauteng

8 - Mpumalanga

9 - Northern Province

The biplots are also split for the white and black population of South Africa. These two groups lie at opposite ends of the income spectrum and the characteristics of each should be analyzed.

Figure 5.5 gives the means for each province together with its 90% alpha bags. The biplot for all the points are not given as it is not very informative. The provinces all lie in a tight group, although it is clear that the Western Cape and Gauteng are the most well-off in terms of income and expenditure. The worst-off provinces in terms of income and expenditure seems to be the Eastern Cape and Northern Province. The correlation between income and expenditure is apparent since the axes lie very close together. Another high correlation seems to be between education level and household size, though they are negatively correlated. This means that the higher the education level the lower the household size.

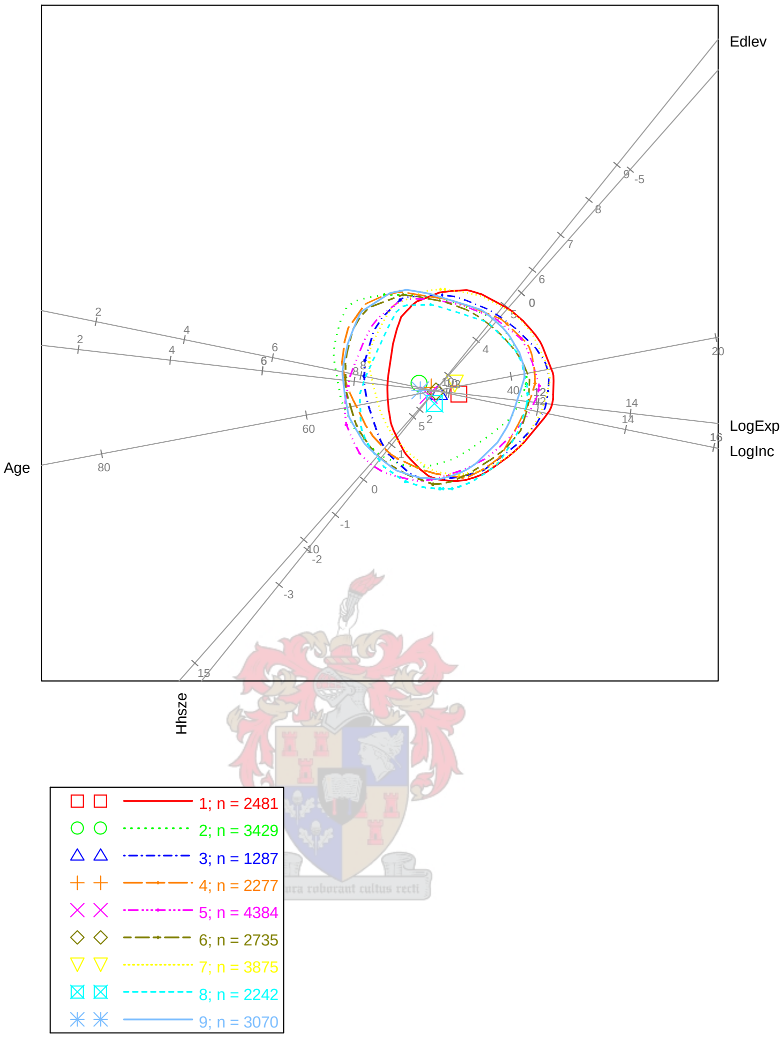


**Figure 5.5:** CVA Biplot by province with 90% bags.

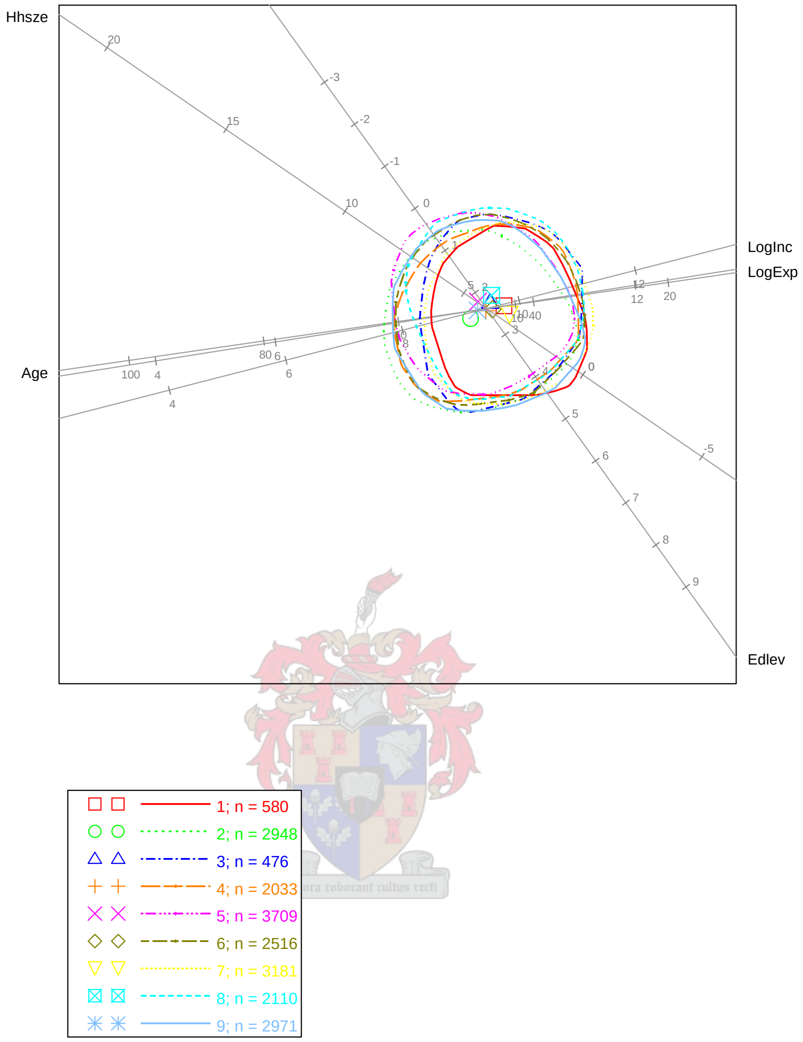


**Figure 5.6:** CVA Biplot with by province 90% bags for black households only.

Figure 5.6 is similar to figure 5.5; it now only gives the 90% alpha bags per province for the black population of South Africa. Here, the black population living in Gauteng is the most well-off, closely followed by those in the Western Cape. The black population of Mpumalanga has the highest education level. A strong negative correlation exists between expenditure and age, in other words, the higher your expenditure the younger you are. This could be a result of the end of apartheid with the younger black population having access to better education and employment opportunities.

Figure 5.7 represents the 90% alpha bags for the white population of South Africa. The correlation between income and expenditure for the white population seems to be less than for the black population. The Western Cape and Kwazulu-Natal have the highest income. Households form Mpumalanga had the highest number of individuals per dwelling, while households form North West had the lowest education level.

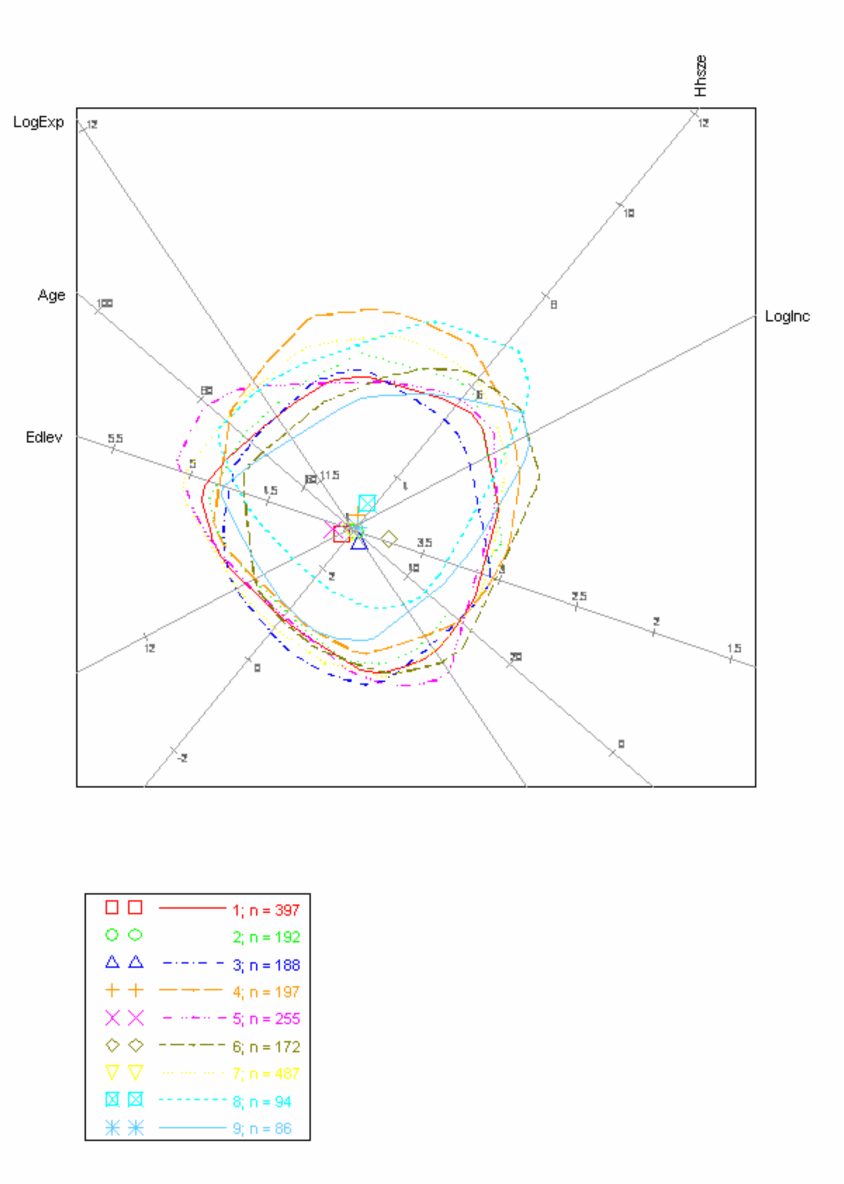


**Figure 5.7:** CVA Biplot with 90% bags for white households only.

### 5.4.3 CVA biplots grouped by area

CVA biplots were also drawn, grouped by area (urban/rural). The codes used were:

1 - Urban

2 - Rural

Figure 5.8 is a CVA biplot with the groups defined by type of area (urban/rural). There is again a strong correlation between income and expenditure. The population living in urban areas tend to have higher income and expenditure and a higher level of eduction. The part of the population living in rural areas has a significantly lower level of education, while it also has slightly lower levels of income and expenditure.

Figure 5.9 is similar to Figure 5.8, but only takes the black segment of the population into account. Again the strong correlation between income and expenditure is evident. The results are largely the same as for Figure 5.8 although the the difference in income and expenditure level between urban and rural is almost negligible. The difference in education level remains significant though.

## 5.5 Conclusions

A multivariate analysis of poverty and income distribution have now been done using biplots. It took into account factors like area of residence, province, education level etc. Throughout the analysis the correlation between income and expenditure became evident. The correlation between income and education level was also clear. Age did not seem to play a role in the level of income obtained, although household size did play a minor role.

The chapter started with a brief look at the advantages of using biplots. The details of how the appropriate dataset was created were given. The variables included in the dataset were province, area type, age, race, education level, household size, total household expenditure and total household income. The biplots were drawn using the total household income. Dummy variables were created for education

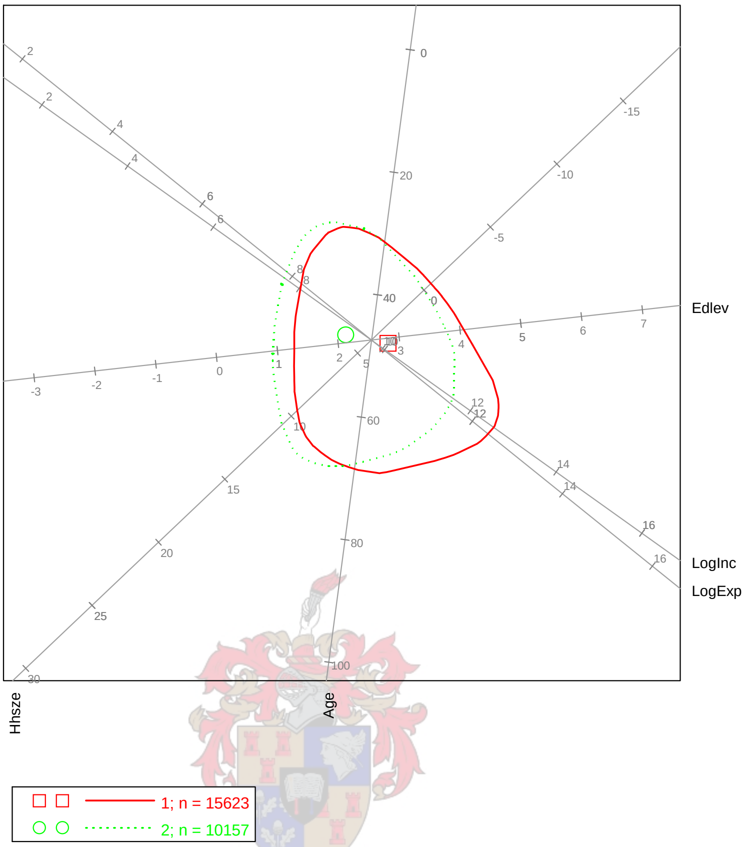


**Figure 5.8:** CVA Biplot by area with 90% bags.

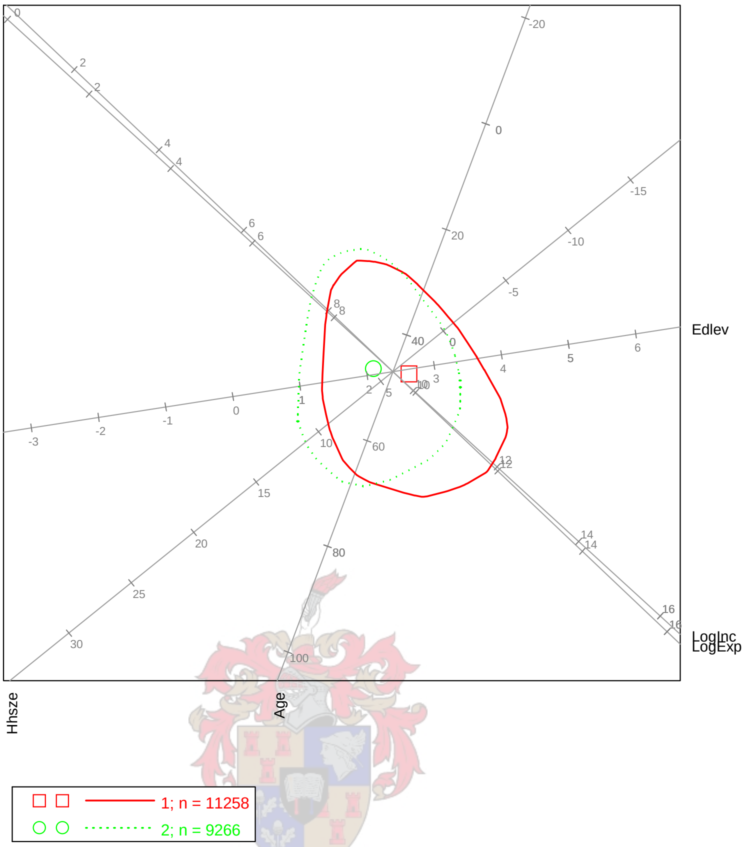


**Figure 5.9:** CVA Biplot by area with 90% bags for black households only.

level. Households (education level of the head of the household) could fall into one of 6 categories, either no education, primary or incomplete primary education, incomplete secondary education, matric, tertiary education or NA (missing value or households that do not know their education level). Principle components analysis (PCA) was then used to obtain biplots colour-coded by race. A brief summary was given on the theory behind PCA biplots. From these plots it is clear that the four races are separated from each other. This separation is due to large differences in the total household income and expenditure between races.

Canonical variate analysis (CVA) is also applied to the dataset. CVA works in such a manner that it separates groups optimally. We used three defined groups, first were defined race to be our group indicator, secondly we used province to be our group indicator and thirdly area type (urban/rural). The CVA biplot grouped by race again identified total household income and expenditure to be the main factors separating the groups. When grouped by province the welfare of each province could be judged relative to the other provinces. The Western Cape and Gauteng were clearly more well-off in terms of income and expenditure, while the Northern Province was the worst off. The same plots were again drawn for respectively black and white households only. Black households living in Gauteng were better off than black households living in the Western Cape. The Western Cape and Kwazulu-Natal were identified as the provinces where white households have the greatest income and expenditure. Area type (urban/rural) was also used as a group indicator. Households living in urban areas are better off than those in rural areas in terms of income and expenditure, although the difference is quite small. For black households this difference is almost negligible.

From the analysis in this chapter it is clear that there is a strong correlation between income and expenditure. The one can thus be used to predict the other. Another variable that features quite prominently is education level. This variable can also give a reasonable indication of the level of income a household will receive. Further study is recommended however. The variables used in this chapter are very limited. Using variables such as type of dwelling, access to clean water etc. would give a more accurate indication of income or expenditure.

# Chapter 6

# Conclusions and Further Study

## 6.1 Summary

As noted before, poverty and income inequality is of special interest in South Africa. It is necessary for government to know the extent of poverty in order to decide on ways of lowering the high rate thereof. It is also advantageous to know the factors associated with poverty, for example, lack of housing, lack of electricity etc. Hence, the correct method of measuring poverty and income inequality is of equal importance. Should income be considered in a continuous or grouped format? These are some of the topics this study touched on.

This analysis was started by discussing a number of definitions of poverty and income inequality from the literature. Various techniques used to explore poverty and income inequality were discussed. These included poverty measures like poverty lines, the FGT family of indicators, the HDI index and others. Income inequality measures included the GE class of measurements and the Gini coefficient. These measurements were described and most of them later used in the analysis of the dataset. A point was made of underlining the multidimensional nature of deprivation. It is not only deprivation in monetary terms, but also employment deprivation, health deprivation, education deprivation and living environment deprivation. Clearly there are lots of angles to measuring poverty. It can be measured on a monetary basis only or as a combination of aspects of deprivation. Also when measuring the monetary side of poverty, how does one define an individual or household as being poor?

The dataset used was the 2000 Income and Expenditure Survey (IES) obtained from Stats SA. This is a five-yearly household survey conducted by Stats SA to measure the welfare of the country. A brief summary was given of the survey design and the weighting used to obtain the dataset. A smaller dataset created from the original IES 2000 dataset was used for the analysis. This dataset was created by the Department of Economics at Stellenbosch University. It included only those variables deemed important to the analysis of income. Any mistakes made by Stats SA on the dataset were also corrected. The debate surrounding continuous versus grouped income variables was highlighted. Most people are reluctant to give an exact income figure or they do not know their income to the nearest Rand. It is thus important to find a way of lowering the rate of nonresponse. This is where grouped income variables come in, since people are more likely to give their income in terms of an income bracket than an exact amount. But how does the use of income as a grouped variable impact on the results obtain when measuring poverty or income inequality?

This was one of the main questions to be answered by chapter 4, but first the dataset needed to be cleaned in terms of missing values and zero income. This was done by dealing with missing values in the same way as Stats SA, by coding them in the correct manner. Unrealistic zero total household income was dealt with in two ways, putting income equal to expenditure or the imputation approach. The imputation approach imputed the unrealistic zero income values by evaluating the income values for households having equal household size and education level, where the education level of a household is the education level of the head of the household. The imputation approach was chosen as the most appropriate method to be used in further analysis. The IES 2000 dataset gives income as a continuous variable, so we needed to create a grouped income variable. This was done using the income intervals defined for Census 2001. The grouped income dataset was then created, but the methods used for analyzing poverty and income inequality made use of continuous variables. The grouped income dataset thus needed to be made continuous again. We used three methods to do this. The first was the midpoint method; each household was given as total household income the midpoint of the income interval to which it belonged. The second method used interval regression to predict the total household income. This model used the education level of the head of the household and age to predict income. The method did not give good results however. The third method was the random midpoint method. Income was randomly distributed over the income interval using the midpoint. Four datasets were thus created, the original continuous dataset and three 'continuous' datasets, created from a grouped income dataset for the analysis. We first chose four well-known poverty lines and fitted the data to them. These were the \$1 a day, \$2 a day, lower-bound and upper-bound poverty lines. The percentage of households and individuals lying below each poverty line were identified for all three datasets. The estimates obtained were then tested for accuracy using confidence intervals. These confidence intervals were obtained using two methods, a large sample approximation method and the bootstrap method. Both methods gave similar results in the form of short confidence intervals. This implied high accuracy for the estimates obtained. However, it must be pointed out that these confidence intervals were only approximations since they assumed the datasets were obtained through simple random sampling and not complex sampling, as is the case. The continuous and midpoint datasets gave very similar results for the poverty lines, while the interval regression dataset seemed to underestimate the number of poor households or individuals. The results obtained for the random midpoint dataset were not usable and that dataset was dropped from further analysis. The next step was to analyze the extreme tail distribution of the three datasets. Approximately 9.14% of households had less than R1044 per capita income per year, compared with approximately 14% of individuals having less than R1044 income per year. Approximately one quarter of South African households have less than R2088 per capita income per year.

Extreme Value Theory was used to fit a generalized Pareto distribution to per capita household income values lying below a predetermined threshold. This model was then used to predict quantiles and exceedance probabilities for the datasets. Again the midpoint and continuous datasets gave similar results. The interval regression dataset overestimates income for very low values, in other words, it predicts a too high income value for households lying in the lowest income level. Approximately 1% of households earn less than R300 per year, while approximately 10% earn less than R1100 using the continuous income variable data. This compares well with the $1 a day poverty line. Income inequality was measured using the Gini coefficient, Theil index and mean log deviation. The Gini coefficient for total household income is approximately 0.63, while the Gini coefficient for expenditure is given as 0.56 by Hoogeveen & Özler (2005). Overall the interval regression dataset underestimates income inequality in South Africa. The results obtained from the continuous dataset and the midpoint dataset were similar. Thus, in conclusion it is not so much the type of variable used, but the method used to approximate a continuous variable from a grouped income variable. It would however be recommended that income be measured in both ways. This would give a higher response rate as well as a continuous income variable to compare with a grouped income variable. This is because the income level data is more reliable, persons are more likely to indicate their correct income bracket, while the continuous income variable contains vital information for fitting a model to predict income.

Next, the continuous dataset was used to assess poverty and inequality within provinces. The same three methods of analysis as above were used, namely poverty lines, extreme tail estimation and income inequality analysis. Through all three it was clear that the Western Cape and Gauteng had the highest level of welfare. In the Western Cape only 1.6% of households lie below the $1 a day per capita household income poverty line, whereas for the Eastern Cape more than 16% of households lie below this poverty line. The Western Cape also has the lowest level of income inequality.

The multivariate nature of poverty was also investigated using biplots. A brief description of the advantages of using biplots was given, together with how the dataset to be used was created. Principle components analysis (PCA) was used to obtain biplots colour-coded by race. It was clear that total household income and total household expenditure were highly correlated and were the major factors in the differences between the races. Canonical variate analysis (CVA) was also used to obtain biplots with groups indicated by race, province and area type (urban/rural) respectively. This again highlighted the correlation between income and expenditure. The education level of the head of the household was also identified as having an impact on the income level obtained.

## 6.2 Suggestions for possible improvement

There are a few aspects to this study that could be improved on. The method of dealing with unrealistic zero income was only an approximation; this should be done in more detail and more variables should be used relating to the households to predict the income level achieved. Also, better ways of converting a grouped income variable to a continuous income variable that will give a more accurate approximation of the distribution of income. The confidence intervals were obtained using methods that assumed simple random sampling (SRS) was used to obtain the dataset. This is not true for the IES 2000 dataset since complex sampling was used. The confidence intervals obtained were thus only approximations and using the correct methods for obtaining them is recommended. Another aspect of this study that could be improved on is the multivariate analysis in terms of biplots. Only a few variables were used in plotting the biplot. If more variables were used a better view could be obtained of the variables having an influence on the welfare of a household.

# Appendices

# Appendix A

# Imputation Approach

Table A.1 provides the values that were imputed for unrealistic zero income in households having a certain education level and a specific household size. These values were obtained by looking at those households having given their total household income and then computing the average income per education level of the head of the household and household size. There were however two deviations from this method, indicated below by an asterisk (*) and a double asterisk (**). The explanation for these deviations are:

* There were no observations having education level 15 and household size 11 and having given their income, the imputed value was then obtained by taking the average total household income of the observations having education level 15.

** These were observations having zero income and specifying their households size but not giving an indication of their education level. The problem was solved by taking the average of the households having given their income and having the respective household sizes.

**Table A.1:** The imputed values for zero income using household size and education level of the head of the household.

| | Household Size (hhsize) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Education level (edlev)* | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **1** | 11484 | 9871 | 10572 | 12227 | 15044 | 14145 | 15778 | | 17662 | | | |
| **4** | 10363 | 11176 | 11762 | 13138 | | | | | 15752 | | | 12605 |
| **5** | 13377 | 11383 | | | 16318 | 15385 | 19480 | 20263 | | 19664 | | |
| **6** | | 12213 | | 19200 | 20532 | | | | 20080 | | | 26126 |
| **7** | 14485 | 14125 | 13499 | 15650 | 14899 | 20332 | 23437 | | | | | |
| **8** | 13577 | 13885 | 13785 | 18967 | 16641 | | | 26810 | | | | |
| **9** | 13794 | 15857 | 18982 | 18871 | 19110 | 26340 | 21730 | | | | | |
| **10** | 14976 | 21073 | 21657 | 25834 | 27599 | 30352 | 25963 | | | 28636 | | |
| **11** | 14946 | 22551 | 21307 | 26947 | | 30884 | | | | | | |
| **12** | 20034 | 38563 | 36521 | | 44087 | 36376 | | | | | | |
| **13** | | 25716 | 26864 | 35419 | | 31624 | | | | | | |
| **15** | | | | | | | | | | | 69194* | |
| **16** | | | 69524 | | | | | | | | | |
| **17** | 27224 | 66834 | 70492 | 84380 | 81963 | | | | 48155 | | | |
| **18** | | 74201 | 102099 | | 87139 | | | | | | | |
| **19** | 58121 | 96861 | 106230 | 149118 | 146315 | 98808 | | | | | | |
| **20** | 95137 | 184391 | 172932 | | | 135216 | | | | | | |
| **21** | 103673 | | | 263127 | 155998 | | | | | | | |
| **23** | | | 23934 | | | | | | | | | |
| **24** | | 13664 | | 152594 | | | | | | | | |
| **NA**** | 20817 | 40518 | | | 40020 | | | | | | | |

# Appendix B

# Creating the Grouped Income Dataset

The R Source Code used to identify the income brackets:

```
program.brackets
function (data)
{# Program to create a catergorical dataset
 # Income brackets are defined as per Census 2001
 # Frequencies and percentages within each bracket is measured
 # Midpoints for each income bracket is calculated

## Income Brackets
inc.matrix <- matrix(NA,nrow=12,ncol=2)
rownames(inc.matrix) <- 1:12
inc.matrix[,1] <- c(0,1,4801,9601,19201,38401,76801,153601,307201,
                    614401,1228801,2457601)
inc.matrix[,2] <- c(0,4800,9600,19200,38400,76800,153600,307200,
                    614400,1228800,2457600,Inf)


## Identifying the brackets
inc.freq <- rep(0,12)
n <- nrow(data)
inc.data <- data[,'totalinc']

for(i in 1:n){
        for(j in 1:12){
                if(inc.data[i]>=inc.matrix[j,1] &&
                   inc.data[i]<=inc.matrix[j,2])
                                inc.freq[j] <- inc.freq[j]+1
        }
}

## Creating the Midpoints
inc.midpoint <- rep(0,12)
inc.midpoint[1] <- 0
for(i in 2:12){
        inc.midpoint[i] <- (inc.matrix[i,2]-inc.matrix[i,1]+1)/2
        +inc.matrix[(i-1),2]
}

## Create a table containing the results
Table <- cbind(inc.matrix,inc.freq,inc.midpoint)
colnames(Table) <- c('Lower','Upper','Frequency','Midpoints')
Table
}
```

## B.1 Creating the Midpoint Dataset

The R Source Code used to create the midpoint dataset:

```
program.midpoints
function (data)
{# Program to create a continuous dataset from
 # the grouped income dataset using midpoints

## Income Brackets
inc.matrix <- matrix(NA,nrow=12,ncol=2)
rownames(inc.matrix) <- 1:12
inc.matrix[,1] <- c(0,1,4801,9601,19201,38401,76801,153601,307201,
                    614401,1228801,2457601)
inc.matrix[,2] <- c(0,4800,9600,19200,38400,76800,153600,307200,
                    614400,1228800,2457600,Inf)


## Creating the Midpoints
inc.midpoint <- rep(0,12)
inc.midpoint[1] <- 0
for(i in 2:11){
        inc.midpoint[i] <- (inc.matrix[i,2]-inc.matrix[i,1]+1)/2
        +inc.matrix[(i-1),2]
}
inc.midpoint[12] <- 2703361    # Top income level is taken to be 10%
 above the lower bound #


## Creating the Midpoint Dataset
n <- nrow(data)
midpoint.inc <- rep(NA,n)
inc.data <- data[,'totalinc']
for(i in 1:n){
        for(j in 1:12){
                if(inc.data[i]>=inc.matrix[j,1] &&
                 inc.data[i]<=inc.matrix[j,2])
                                midpoint.inc[i] <- inc.midpoint[j]
        }
}

## Adding the other variables
data <- cbind(data[,-20],midpoint.inc)
data
}
```

## B.2 Creating the Interval Regression Dataset

The software package used was that of STATA. The following program was used to fit the model and predict income values:

```
# Generating the income levels
gen inccat=.
replace inccat=1 if totalinc==0
replace inccat=2 if totalinc>=1 & totalinc<=4800
replace inccat=3 if totalinc>=4801 & totalinc<=9600
replace inccat=4 if totalinc>=9601 & totalinc<=19200
replace inccat=5 if totalinc>=19201 & totalinc<=38400
replace inccat=6 if totalinc>=38401 & totalinc<=76800
replace inccat=7 if totalinc>=76801 & totalinc<=153600
replace inccat=8 if totalinc>=153601 & totalinc<=307200
replace inccat=9 if totalinc>=307201 & totalinc<=614400
replace inccat=10 if totalinc>=614401 & totalinc<=1228800
replace inccat=11 if totalinc>=1228801 & totalinc<=2457600
replace inccat=12 if totalinc>=2457601

# Generating the lower boundaries
gen lower=.
replace lower=0 if inccat==1
replace lower=1 if inccat==2
replace lower=4801 if inccat==3
replace lower=9601 if inccat==4
replace lower=19201 if inccat==5
replace lower=38401 if inccat==6
replace lower=76801 if inccat==7
replace lower=153601 if inccat==8
replace lower=307201 if inccat==9
replace lower=614401 if inccat==10
replace lower=1228801 if inccat==11
replace lower=2457601 if inccat==12
```

```
# Generating the upper boundaries
gen upper=.
replace upper=0 if inccat==1
replace upper=4800 if inccat==2
replace upper=9600 if inccat==3
replace upper=19200 if inccat==4
replace upper=38400 if inccat==5
replace upper=76800 if inccat==6
replace upper=153600 if inccat==7
replace upper=307200 if inccat==8
replace upper=614400 if inccat==9
replace upper=1228800 if inccat==10
replace upper=1457600 if inccat==11
replace upper=. if inccat==12

# Generating the dummy variables for education level
gen noeduc=0
replace noeduc=1 if edlev1==1

gen primary=0
replace primary=1 if edlev1==2|edlev1==3|edlev1==4|edlev1==5|
                     edlev1==6|edlev1==7|edlev1==8|edlev1==9

gen incsecond=0
replace incsecond=1 if edlev1==10|edlev1==11|edlev1==12|edlev1==13|
                       edlev1==14|edlev1==15
gen matric=0

replace matric=1 if edlev1==16|edlev1==17|edlev1==18|edlev1==19

gen tertiary=0
replace tertiary=1 if edlev1==20|edlev1==21|edlev1==22
```

```
gen missing=0
replace missing=1 if edlev1==23|edlev1==24

# Generating the age-squared variable
gen age2= age^2

# Generating the log income boundaries
gen lnlower= ln(lower)
gen lnupper= ln(upper)

# Fitting the model
intreg lnlower lnupper primary incsecond matric tertiary missing age age2
[pweight=hhweight], robust
predict imputation

# Transforming the predicted income values back
gen impincome=exp(imputation)
```

## B.3 Creating a Random Midpoint Dataset

The following program was written in R and used to create a random midpoint dataset:

```
 prog.uniform
function (data)
{# Program to create a catergorical dataset
 # Income brackets are defined as per Census 2001
 # Frequencies and percentages within each bracket is measured
 # Midpoints for each income bracket is calculated

## Income Brackets
inc.matrix <- matrix(NA,nrow=12,ncol=2)
rownames(inc.matrix) <- 1:12
inc.matrix[,1] <- c(0,1,4801,9601,19201,38401,76801,153601,307201,
                    614401,1228801,2457601)
inc.matrix[,2] <- c(0,4800,9600,19200,38400,76800,153600,307200,
                    614400,1228800,2457600,Inf)



## Identifying the brackets
inc.freq <- rep(0,12)
n <- nrow(data)
inc.data <- data[,'totalinc']

for(i in 1:n){
        for(j in 1:12){
                if(inc.data[i]>=inc.matrix[j,1] && inc.data[i]<=inc.matrix[j,2])
                                inc.freq[j] <- inc.freq[j]+1
        }
}

## Creating the Random Midpoints
```

```
inc.midpoint <- rep(0,12)
inc.midpoint[1] <- 0
for(i in 2:11){
        inc.midpoint[i] <- (inc.matrix[i,2]-inc.matrix[i,1]+1)/2+inc.matrix[(i-1),2
}
inc.midpoint[12] <- 2703361

new.midpoint <- NULL
toets <- 0
for(i in 2:12){
  rand.unif <- runif(inc.freq[i],inc.matrix[i,1],inc.midpoint[i])
  rand.sign <- rbinom(inc.freq[i],1,0.5)
  for(j in 1:inc.freq[i]){
        if(rand.sign[j]==1) {
                toets <- inc.midpoint[i]+rand.unif[j]
                new.midpoint <- c(new.midpoint,toets)}
        if(rand.sign[j]==0) {
                toets <- inc.midpoint[i]-rand.unif[j]
                new.midpoint <- c(new.midpoint,toets)}
        }
}

## Creating the Random Midpoint Dataset
n <- nrow(data)
inccat <- rep(NA,n)
inc.data <- data[,'totalinc']

for(i in 1:n){
        for(j in 1:12){
                if(inc.data[i]>=inc.matrix[j,1] &&
                                        inc.data[i]<=inc.matrix[j,2])
                                inccat[i] <- j
        }
}
data <- cbind(data,inccat)

nuwe.data <- NULL

for(i in 2:12){
       nuwe.data <- rbind(nuwe.data,data[data[,'inccat']==i,])
}

nuwe.data <- cbind(nuwe.data,new.midpoint)
nuwe.data

}
```

# Appendix C

# A Bootstrap program in R

The following program was written in R and used to obtain the standard error of the parameter to be estimated:

```
prog.bootstrap
function (data,B,povertyline,alpha=0.05)
{# Program to calculate the standard error of a parameter
 # using Bootstrap techniques

n <- length(data)
estimate <- prog.poverty.lines(data,povertyline)$Percentage
boot.est <- rep(0,B)

## Repeat B times
for(i in 1:B){

# Draw the bootstrap sample
smpl <- sample(data,n,TRUE)

# Calculate the parameter
boot.est[i] <- prog.poverty.lines(smpl,povertyline)$Percentage
}
# Calculate the standard error
boot.mean <- mean(boot.est)
boot.se <- sqrt((sum(boot.est-boot.mean)^2)/(B-1))

# Calculate the confidence interval
z.alpha <- -qnorm(alpha/2)
lower <- estimate-z.alpha*boot.se
upper <- estimate+z.alpha*boot.se
return(estimate,boot.se,z.alpha,lower,upper)
}
```

# Appendix D

# Creating a dataset for biplots

The statistical language R/S-PLUS was used to create the dataset to be used in plotting biplots. The biplots were also drawn using this software. The following R Source code was used to create the education level categories:

```
edlev.prog
function (data)
{# Function to convert education level to one of 6 categories
 # 1 - No Education
 # 2 - Primary School education or incomplete primary school education
 # 3 - Incomplete Secondary school education or NTC I and II certificates
 # 4 - Matric
 # 5 - Tertiary Education
 # 6 - Missing (will be coded with an ’NA’)

n <- nrow(data)

# Defining the categories
categories <- matrix(0,nrow=24,ncol=2)
categories[,1] <- 1:24
categories[,2] <- c(1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,5,5,5,NA,NA)
# Applying the categories to the dataset
edlev.cat <- rep(0,n)
for(i in 1:n){
   for(j in 1:24){
           if(data[i,'edlev']==categories[j,1])
           edlev.cat[i] <- categories[j,2]
        }
}

data <- cbind(data,edlev.cat)
data
}
```

# References

African National Congress. 1994. *Reconstruction and Development Program: A policy framework.* Johannesburg: Umanyano Publications.

Bhorat, H., Poswell, L. & Naidoo, P. 2004. Dimensions of poverty in Post-Apartheid South Africa: 1996 - 2001. A Poverty Status Report,. *Development Policy Research Unit (DPRU).* University of Cape Town.

Blane, D. & Drever, F. 1998. Inequality among men in standardised years of potential life lost, 1970-93. *British Medical Journal*, 317:255–256.

Boltvinik, J. 2001. Poverty Measurement Methods - An Overview. *United Nations Development Program.*

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values.* Springer Series in Statistics. London, UK: Springer-Verlag.

Coles, S. & Stephenson, A. 2006. *ismev: An Introduction to Statistical Modeling of Extreme Values.* R package version 1.2.
Available at: `http://www.maths.lancs.ac.uk/ stephena/`

Coudouel, A., Hentschel, J.S. & Wodon, Q.T. 2002. Poverty Measurement and Analysis. *World Bank Report.* Http://unstats.un.org/unsd/methods/poverty/edocuments.htm.

Deaton, A. 2003. Measuring poverty in a growing world. *NBER Working Paper*, nr. 9822. Cambridge, Massachusetts: National Bureau for Economic Research.

Deaton, A. 2004. Measuring poverty in a growing world. *The Review of Economics and Statistics, February 2005*, 87/01.

EPRI. 2001. Impact of the social security system on poverty in South Africa. *Economic Policy Research Institute Research paper.* Nr.19.

Foster, J., Greer, J. *&* Thorbecke, E. 1984. A Class of Decomposable Poverty Measures. *Econometrica 52*, no.3:761–765.

Govender, P., Kambaran, N., Patchett, N., Ruddle, A., Torr, G. *&* Van Zyl, N. 2006. Poverty and Inequality in South Africa and the World. *Actuarial Society of South Africa.* Convention Paper.

Gower, J. *&* Hand, D. 1996. *Biplots.* 1st edition. London: Chapman and Hall.

Hirschowitz, R., Orkin, M. *&* Alberts, P. 2000. Key baseline statistics for poverty measurement. *Measuring Poverty in South Africa.* Pretoria: Statistics South Africa.

Hoogeveen, J.G. *&* Özler, B. 2005. Not Separate, Not Equal: Poverty and Inequality in Post-Apartheid South Africa. *William Davidson Institute Working Paper*, (739).

Le Roux, N.J. 2006. *PCA and CVA biplots.* R package for drawing biplots.

Litchfield, J. 1999. Inequality: Methods and Tools. *Text for the World Bank's website on Inequality, Poverty and Socio-economic Performance from http://www.worldbank.org/poverty/inequal/index.htm.*

Lohr, S.L. 1999. *Sampling: Design and Analysis.* Duxbury Press. California, USA: Brooks/Cole Publishing Company.

Noble, M., Babita, M., Barnes, H., Dibben, C., Magasela, W., Noble, S., Ntshongwana, P., Phillips, H., Rama, S., Roberts, B., Wright, G. *&* Zungu, S. 2006. The Provincial Indices of Multiple Deprivation for South Africa 2001. *University of Oxford, UK.*

Provincial Decision-Making Enabling Project (PROVIDE). 2005. Creating a 2000 IES - LFS database in Stata. *Technical Paper Series*, (2005:1).

Ravallion, M. 1994. *Poverty Comparisons.* Harwood Academic Publishers.

Ravallion, M. 2001. Poverty Lines: Economic Foundations of Current Practices. The World Bank. Unpublished manuscript.

Reilly, Dr., B. 2007. Lecture one: Wage equation estimation. Lecture notes, University of Sussex.

Schluter, C. & Trede, M. 2002. Tails of Lorenz Curves. *Journal of Econometrics*, 109:151 – 166.

Sen, A. 1976. Poverty: an Ordinal Approach to Measurement,. *Econometrica*, 44:219–231.

Sen, A. 1992. *Inequality Re-Examined.* Cambridge, Harvard University Press.

Statistics South Africa. 2000*a*. Income and Expenditure Survey, 2000. *Pretoria: Statistics South Africa.*

Statistics South Africa. 2000*b*. Measuring poverty in South Africa. *Pretoria: Statistics South Africa.*

United Nations. 2005. The Millennium Develpment Goals Report. *United Nations Development Programme.* New York.

United Nations. 2006. Human Development Report 2006. *United Nations Development Programme*, 335.

Von Fintel, D. 2006. Earnings bracket obstacles in household surveys - How sharp are the tools in the shed? *Stellenbosch Economic Working Papers*, (08/06).

Whitehouse, E. 1995. STATA/SE package. OECD, Paris.

Woolard, I. 2001. Income Inequality and Poverty: Methods of Estimation and Some Policy Applications for South Africa. *Unpublished PhD thesis, University of Cape Town.*

Woolard, I. & Leibbrandt, M. 1999. Measuring poverty in South Africa. *Development Policy Research Unit (DPRU).* Working Paper 99/33, University of Cape Town.