Chapter 1
Introduction

This book is concerned with the measurement and quality of employee income from household survey (micro) data. The empirical applications are based on South African household surveys compiled by the national statistics agency (Statistics South Africa). Despite this specificity, the insights are generalisable to any household survey concerned with measuring income.

Data quality is a central theme in any data compilation effort. However, it is often very difficult to diagnose where exactly in the data production process data quality falters. Data quality is a concern for both macro- and microdata. For macro-economic data, the International Monetary Fund (IMF) presides over the process of ensuring standards are developed for the production of national economic statistics associated with the System of National Accounts (see IMF, 2003 for the latest such framework). For household survey data, there are data quality frameworks for surveys themselves (see Statistics Canada, 2003, 2009 and Statistics South Africa, 2006a; 2006b), and for specific themes like income.

In all household survey data, several forms of error are present in different magnitudes, including coverage error, sampling error, nonresponse error, adjustment error, processing error, measurement error and validity. These components of error form part of the total survey error paradigm (Groves et al., 2004), and can often be exacerbated by poor data quality management within statistical organisations. To understand data quality therefore requires some understanding of the practises inside statistical organisations with respect to data quality control. Examples of such data quality control elements from Statistics Canada (2003) and Statistics South Africa (2006a) include relevance, timeliness, accessibility, interpretability, coherence, integrity, methodological soundness and accuracy.

For income data measured in household surveys, the Canberra Group's (2001; 2011) recommendations on household income statistics is the main reference. The Canberra Group was a group of national statistics and other data compilation agencies from over fifteen countries, plus representatives from many international agencies, whose main objective was to "... enhance national household income statistics by

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_1

developing standards on conceptual and practical issues related to the production of income distribution statistics" (Canberra Group, 2001, xi). The global level of importance accorded to this task was noteworthy, for it coincided with the adoption of the Millennium Development Goals, the first of which was to halve absolute poverty, defined as all those living below US$1.00 per day in constant purchasing power parity (PPP) adjusted terms, between 1990 and 2015.[1]

The income distribution has been a central preoccupation of economists since the inception of the discipline due to its positive correlation with individual and societal welfare. An important formalisation of the work on income distributions was made by Vilfredo Pareto in the nineteenth century, who found after analyses of empirical income data on several European countries that the probability distribution of income was right-skewed (Kirman, 2008). More detailed analyses of income distributions since then led to the realisation that several possible statistical distributions have valid application to income over different ranges of the variable (see (Cowell, 2000) for discussion).

As long as people have analysed income distributions there have been debates about the data utilised for this purpose. Income is measured both in the national accounts and with household survey data. However, the methodologies used to collect and aggregate this data renders income measured in the national accounts to be quite a different construct to income measured in household surveys (Havinga et al., 2010). This book is concerned with income measured in household surveys only.

1.1 The Income Construct in Household Surveys

Generally, when income distribution is discussed, the debate concerns the distribution of *total* income. But total income is comprised of many components. The Canberra Group (2001, 18) distinguish the following types of income that together sum to total income:

- Employee income, plus
- Income from self-employment, plus
- Income from rentals, plus
- Property income, plus
- Current transfers received.

This book is primarily concerned with employee income. Employee income is considered to be a form of cash income that is easily and accurately measured relative to property income and cash transfers (Canberra Group 2000, 13). However, the employee income question in household surveys is complicated by a feature that is designed to increase the probability that a respondent answers the question. That is, a second, bounded income bracket question is presented to respondents as a follow-up to the exact income question in the event that they refuse to answer or state that

[1] See http://www.un.org/millenniumgoals/poverty.shtml.

they don't know. This leads to an income variable with a continuous distribution for exact income responses and a discrete, grouped continuous distribution for bounded income bracket responses.

Respondents can also refuse to answer the follow-up question, or once again state that they don't know their income or that of the proxy respondent on whose behalf they are reporting. Consequently, there is also nonresponse to the employee income question. How researchers treat the many issues that confront them with income data in public-use household surveys can often be very different, leading to different estimates of parameters of the income distribution from the same dataset.

The advantage of having a follow-up income question with a lower level of information disclosure is that it reduces the social sensitivity of the question, but can also aid respondent recall. Consequently, some form of follow-up question that bounds the range of income is often also asked in household surveys for other components income, including income from self-employment, rentals, property and transfers. Therefore, while the emphasis in this book is on employee income, the insights are generalisable methodologically to any component of income that is measured in a similar way.

The overall quality of household surveys also has an important bearing on the accuracy of individual income statistics. In South Africa (SA), nationally representative sample surveys have only been compiled by Statistics South Africa (SSA) since the early 1990s. Before 1994, the geopolitical borders of SA included the Bantustans, considered separate by the Apartheid Government to the state of SA. Consequently, in the national statistics community in the mid 1990s, more emphasis was placed on creating new sampling frames for the democratic SA than refining questionnaire design for constructs like employee income. This necessary trade-off in the data production process led to poorer quality income data initially that gradually improved as other operational aspects of the household surveys themselves improved.

1.2 Objectives and Chapter Typology

The main objectives of this book are:

- To develop a framework for investigating microdata quality and apply this framework to South African labour market household surveys that include a question on employee income.
- To investigate the relationship between questionnaire design for employee income and the respondents who choose to answer the question in different ways (including bounded income bracket responses, refusals and don't know responses).
- To formulate practicable solutions for researchers concerned with generating a derived employee income variable from public-use income variables with varying degrees of coarseness, using multiple imputation for this purpose.

Chapter Two is directed at understanding the universe of errors that can arise in household surveys and linking these to data quality protocols inside statistical

organisations. It identifies specific data quality metrics for each component of survey error that can arise. It then applies this framework to South African labour market household surveys. The chapter provides a general taxonomy for investigating data quality that can be useful to researchers whose aim it is to understand the relationship between survey error and data quality in public-use datasets. In order to demonstrate this, the individual income variable is reviewed for the employed population of South Africa, evaluated over multiple survey instruments and time periods, ranging from 1995–2007.

Chapter Three in this book isolates the design of the employee income question in household surveys and the propensity of respondents to provide a particular response type. Employee income is typically measured in a way that allows respondents to provide either an exact income value or an interval into which it falls. It is then the user's responsibility to generate a variable that combines these two response types appropriately. However, missing data is also present when respondents refuse to answer the question or state they don't know. Understanding the different subsets of respondents sheds light on the trade-offs of questionnaire design for employee income, and provides valuable insight into the response process that can inform single and multiple imputation exercises.

The final substantive chapter then goes on to investigate public-use employee income data with a mixture of continuously distributed income observations, grouped-continuous observations and item nonresponse. This mixture of data types is called coarse data in the literature, and has important implications for imputing plausible values for such data. In SSA's household surveys, we also find that there is a non-trivial degree of processing error in the two income variables released in the public-use dataset that must be treated appropriately before multiple imputation exercises can commence. We then conduct multiple imputation and discuss several aspects of the imputation algorithm – from the estimation method, to constraints on the bounds of the plausible draws, to the specification of prediction equations – all of which have a bearing on the reliability of imputed draws. Once these concerns have been addressed, multiple univariate imputations of employment income from coarse data can be obtained in a manner that allows researchers to account for the greater uncertainty inherent in that data. This then allows for the reliable estimation of univariate parameters of the income distribution.

The time-frame for the analysis spans the mid 1990s to the latter part of the 2000s for Chapter Two. However, for Chapters Three and Four, the time-frame is restricted to 1997–2003. This is because this period was associated with important changes in the way household surveys were conducted in Statistics South Africa. Between 1995–1999, the October Household Survey was a repeated cross-sectional survey that collected labour market data as well as more general household information. From 2000 onwards, this survey was split into the Labour Force Survey (LFS) and the General Household Survey. Only the LFS is analysed in this book. This allows us to identify the role of questionnaire design in improving the quality of income data.

The LFS was designed as a rotating panel survey whose explicit purpose was to obtain accurate estimates of employment and unemployment. In Chapters Three

and Four, only the September Waves of the Labour Force Survey are analysed in conjunction with the OHS 1997–1999. Because the LFS is a *rotating panel* household survey (see Cantwell, 2008 for a definition), a proportion of the respondents change with each Wave of the survey, ensuring that it is representative of the South African population at the time of going to field. Therefore, it is possible to analyse the cross-sectional OHSs in combination with individual waves of the rotating LFS panel.

The final chapter in this book concludes the discussion. Since each chapter contributes original insight into different aspects of data production and use, the Conclusion stresses the need to factor all of the issues discussed in this book into an overall set of guidelines for estimating parameters of the income distribution. The discussions in chapters three and four, in combination, provide particularly powerful insights about how to ultimately derive reliable points estimates about poverty and inequality.

References

Canberra Group. (2001). *Expert group on household income statistics: Final report and recommendations*. Ottawa: The Canberra Group

Canberra Group. (2011). *Canberra group handbook on household income statistics* (2nd ed.). Geneva: United Nations Economic Commission for Europe.

Cantwell, P. J. (2008). Rotating Panel Design. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods.* Thousand Oaks: Sage Publications.

Cowell, F. A. (2000). Measurement of inequality. In A. B. Atkinson & F. Bourguignon (Eds.), *Handbook of income distribution* (Vol. 1). New York: Elsevier.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley Press.

Havinga, I., Kamanou, G., & Vu, V. (2010). A note on the (mis)use of national accounts for estimation of household final consumption expenditures for poverty measures. In S. Anand, P. Segal, & J. E. Stiglitz (Eds.), *Debates on the measurement of global poverty.* London: Oxford University Press.

International Monetary Fund (IMF). (2003). Data quality assessment framework—Generic framework. IMF, Washington D.C.: Mimeo.

Kirman, A. (2008). Pareto, Vilfredo (1848–1923). In S. N. Durlauf & L. E. Blume (Eds.), *The new palgrave dictionary of economics* (2nd ed.). Palgrave Macmillan.

SSA. (2006a). *Draft data quality framework 001: South African statistical quality assessment framework*. Pretoria: SSA.

SSA. (2006b). *Data quality policy 001: Policy on informing users of data quality*. Pretoria: SSA.

Statistics Canada. (2003). *Statistics Canada quality guidelines* (4th ed.). Ottawa: Statistics Canada.

Statistics Canada. (2009). *Statistics Canada quality guidelines* (5th ed.). Ottawa: Statistics Canada.