Chapter 5
Conclusion: How Data Quality Affects Our Understanding of the Earnings Distribution

Household survey data are subject to multiple forms of survey error that can have a direct bearing on data quality, influencing end-user estimates of parameters of interest in unpredictable ways. This book has focussed specifically on employee income, but the insights are generalisable to any component of income.

Chapter Two developed a framework for investigating microdata quality that was based largely on the total survey error (TSE) paradigm, but that also included specific data quality control elements. The TSE framework decomposes survey error into coverage error, sampling error, nonresponse error, adjustment error, processing error, measurement error and validity. We focussed on adapting the total survey error framework to shed light on which aspects of data quality researchers can observe and do something about. This framework then served as the basis for evaluating the evolution of data quality in Statistics South Africa's labour market household surveys from the early 1990s to 2007.

It was argued that efforts to improve data quality should involve a virtuous interaction between producers and consumers of microdata and should be considered an evolving process. For producers of data, the preparation and publication of detailed data quality frameworks was emphasised, and two such examples were reviewed (the Statistics Canada and SSA Data Quality Frameworks). These frameworks are also excellent documents to inform users about issues of relevance to survey organisations and how these may affect the overall quality of the public-release data. For example, the late 1990s would have been an excellent time for the national statistics office (SSA) to inform users to expect variation over the repeated cross-sections of survey data due to non-sampling errors, and to explain that process in some detail. However, data quality frameworks were not in use by SSA at that time.

For consumers of data, judicious analyses of the univariate, bivariate and multivariate relationships in public-use versions of the datasets shed light on different components of survey error in variables of interest. Any problems associated there-

R. C. Daniels, *How Data Quality Affects our Understanding of the Earnings Distribution*,
https://doi.org/10.1007/978-981-19-3639-5_5

with should be communicated back to survey organisations. However, this does not make the analysis task any easier, and comparisons of repeated cross-sections of income data are particularly vulnerable to components of survey error directly under the control of the survey organisation. Ultimately, it was noted that improving data quality for income in particular is about improving data quality for household surveys in general.

Chapter Three isolated questionnaire design and item nonresponse for the employee income question in two South African labour market surveys: the October Household Survey (1997–1999) and the Labour Force Survey (2000–2003). The choice of time period isolated a period of changing questionnaire design for the employee income question. Between 1997 and 2000, the income question gradually included new response options for the respondent to state that they don't know or refuse to answer the question. We used sequential logistic response models to evaluate how improvements to the income question improved the capacity to understand the nonresponse and bounded response mechanisms. The use of these models represents an important contribution to the literature, for they can be used to evaluate the response process regardless of whether the bounded response question is in the form of a showcard, an unfolding bracket or a respondent generated interval.

It was found that the probability of initial nonresponse to the exact income question was correlated with income, but when the second follow-up bounded income question was presented to respondents, final nonresponse was no longer repeatedly associated with predictors of income. This suggested that the bounded income question overturned initial nonresponse to the exact income question and included more high income earners in the observed response subset. The addition of refuse and don't know response options to the employee income question played a very important role in improving the understanding of the nonresponse process, but in the final analysis of this chapter at least, respondents who refused to answer the employee income question were no longer significantly different to those who stated that they didn't know their income, at least as far as predictors of income were concerned. Rather, correlates of the knowledge of income became significant, with self-reporters and those cohabiting with romantic partners having the most consistently higher odds of refusing over time.

Chapter Four was concerned with conducting univariate multiple imputations for coarse response subsets of the employee income question. An analysis of the interrelationship between the exact income and bounded income variables released in the public-use data revealed a non-trivial degree of processing and/or measurement error for certain survey years between 1997–2003. We identified two forms of error that had to be dealt with effectively before multiple imputation could be performed. We also noted an idiosyncratic feature of the bounded employee income question in all of SSA's household surveys, namely the existence of a zero bracket. This was left in the data and not imputed because it was deemed to be a reasonable response value to the income question given the fact that employees could state they were not working due to being ill.

Once these features of the public-use data were effectively treated, we then conducted multiple imputations for coarse income observations using four differ-

ently specified models to test the sensitivity of imputed draws of income to misspecification in the imputation algorithm. It was found that a combination of response propensity and Mincerian earnings function covariates led to imputed draws that were the least likely to be extreme values in the income distribution, relative to alternative specifications. This has very important implications for more complex multiple imputation algorithms that seek to simultaneously impute income and covariate coarse data, an exercise that will require much initial data preparation and analysis before the integrity of the algorithm can be validated.

We then also evaluated the point estimates of quantiles and moments of the multiply imputed income distributions as the number of imputations increased, where it was found that stability in the estimates and inferences was achieved after only two imputations. This was likely a product of both the low percentage of item missing data and the restricted ranges of plausible imputed draws for the bounded income respondents. However, despite the low percentage of item missing data, it was found that imputed draws for refusals always had higher values than don't know respondents. This was a very important finding that was not discernible in Chapter Three, where predictors of the refuse subset no longer seemed to be different to the don't know subset on variables correlated with income.

The coarse data framework proved to be very useful in Chapter Four in guiding the approach to multiple imputation, not only because it informed the use of an interval censored regression algorithm, but also because it led to the decision rule to exclude unspecified responses in the LFS from being imputed in the primary analysis. When we then conducted a separate imputation process for these unspecified responses in 1999 and 2000, it was found that imputed draws were very differently distributed compared to imputed draws for don't know and refuse responses. This suggested that unspecified responses was an altogether different error process to item nonresponse on the employee income question, and should be treated as such.

Taken in combination, Chapters Three and Four show the necessary steps that researchers need to take when preparing the data for final estimates of univariate parameters of the earnings distribution. Post-imputation, poverty and inequality estimates can only then be thought of as accurate to the maximum degree possible given the data. This is true regardless of the country for which data is collected, which makes the methodology generalisable to any context. Limitations could still exist though, to the extent that unobservable components of survey error, such as frame error and sampling error, remain material.

In summary then, the presence of multiple sources of survey error in microdata need not impose undue constraints to the reliable estimation of parameters of the income distribution. What is required is that each source of survey error's potential impact on that distribution is known, even though nothing can be done about some of those components of error after public release of the data. For those components of error that can be observed, statistically rigorous methodology has to inform the approach to univariate and multivariate analyses, and researchers need to be explicit about their treatment of each relevant component of error.