Felix Ritchie (University of the West of England)

Improving data quality: the value of the user as detective

Topic 2 – Learning more from what we already know

Keywords: quality, non-compliance, minimum wage, measurement error

Introduction

Statistical quality is often measured using statistical tools: distributions, outliers, comparability with similar observations or in different periods, and so on. In recent years, the preponderance of 'statistical editing' has encouraged the use of statistical tools, often automated, to clean up the data. Much of this is driven by a need to reduce the cost of data collection.

This automation goes against the increasing demand for granularity in aggregate statistics, and microdata for direct re-use by researchers, both of which require high quality data at the micro level. As Ritchie (2004) notes, this presents a problem for national statistical institutions (NSIs): not only do cost pressures encourage a focus on macro accuracy, but micro accuracy often requires a knowledge of how the data is to be used, knowledge which is not necessary held by the NSI. Finally, the detective work necessary to maintain high-quality microdata can be labour-intensive.

Nevertheless the gains from labour-intensive detective work can be substantial: a simple rounding error, for example, changed aggregate statistics by 3%. Moreover, this work should be seen as an investment, not an ongoing expense, as we are trying to understand how systematic impacts on quality occur.

Methods / Problem statement

To illustrate the value of the user-based detective study, we consider the question of non-compliance with statutory minimum wages in the UK. Many countries have a statutory minimum wage for employees. There is a strong policy interest in knowing the degree of compliance with the law, and quantitative analysis seems ideally suited to this. In practice, distinguishing genuine underpayment of wages from reported data is complicated by statistical factors: sampling, weighting, survey design, measurement error, timing, sample sizes, and processing errors can all contribute to the under- or over-estimation of the true level of compliance.

While these factors can raise concerns in any quantitative analysis, the potential for inappropriate analysis is exacerbated by the ‘yes-no’ nature of compliance. However, the binary nature of compliance also raises opportunities to study in detail how very small problems in the data can lead to much larger apparent changes in aggregate statistics: a statistical 'butterfly effect'. We exploit this to study quality problems in data which would be hidden from analyses that rely upon continuous distributions.

Results / Proposed solution

UK minimum wages have been extensively studied, using large-scale high-quality datasets whose characteristics are well understood; and previous studies have shown that household-based surveys are subject to substantial measurement error around the minimum wages.

We focus particularly on apprentices; the minimum wage for apprentices was introduced in 2010, and non-compliance rates for some groups of apprentices appear to be high as 50%. Ritchie et al, (2016) provide a detailed discussion of the issue; the focus here is on the operational implications for NSIs. We supplement

this analysis with wider examples of the user-centred approach to data analysis in business and social datasets, to draw out common themes across data collection.

Conclusions

Bringing together results from three years of commissioned analysis on minimum wages, we identify several problems leading to under- and over-estimation of the non-compliance rate. We also note that only some of those are amenable to statistical solutions.

By augmenting the analysis of non-compliance with more general work looking at data quality from the research users' perspective, we show that the problems identified are often systematic, and hence amenable to changes in operating procedures. Hence, this sort of detective work should be seen as an investment in data quality, not a recurring expense.

While concentrating on non-compliance with a statutory minimum wage, the paper has some wider lessons for the analysis of government data. In particular, we demonstrate the value of a very detailed knowledge of the data at crucial points in the distribution, and the importance of triangulation for understanding the reliability of estimates.

Finally, we note the importance of making microdata available to researchers, as one of the cheapest and most effective quality-control measures available to NSIs.