Enhancing household survey microdata accuracy using machine learning: literature survey

Nikhil Woodruff

Contents

1 Introduction **1**

2 Under-coverage of very high incomes **1**

3 Adjustments using administrative tax data **2**

4 Capital income imputation **3**

5 Under-coverage of very low incomes **4**

6 Linking data directly to administrative data **4**

1 Introduction

Over the last few decade, there has been extensive research into the accuracy of household surveys for estimating socioeconomic and policy-related indicators, as well as methods of improving survey accuracy. Most of these studies have focussed on one particular mechanism by which surveys introduce inaccuracy (for example, by omitting top incomes or under-sampling low incomes), and examine a method of improving surveys which tackles this particular flaw. This literature survey aims to provide a comprehensive overview of the state of the art in improving survey accuracy, while also examining how and if these individual advancements complement each other.

2 Under-coverage of very high incomes

The Department for Work and Pensions is required by law to report on poverty and inequality metrics every year, and in meeting this requirement, it publishes a household-level dataset of disposable incomes, termed the Households Below Average Income (HBAI) dataset.[3] Since 1992, it has applied an adjustment to the disposable incomes of a subset of the dataset in order to make the coverage of top incomes more comparable with that of HMRC's Survey of Personal Incomes

(SPI) dataset - this adjustment termed the 'SPI adjustment'. In [2], the authors examine the methodology of this adjustment, as well as its performance against its original goals.

The authors document[1] the steps of the SPI adjustment, which involve first identifying a set of 'rich' households. The definition of rich applies a condition that a household's income must be above a certain threshold, where separate thresholds are used for pensioner and non-pensioner households. The target used to set thresholds is generally to ensure that around 0.5% of records are altered, varying by year. The HBAI 'rich' households are then modified by replacing gross incomes (an income measure which the SPI also contains) with the average values for records in the same group in the SPI. Finally, the survey weights are recalculated: in the original survey, weights are solved by matching administrative statistics on population sizes; under the SPI adjustment, population sizes of the 'rich' groups are included in the set of statistical targets to hit.

The authors find that the SPI adjustment has been successful in improving the coverage of top incomes in the HBAI dataset, but raise a number of issues:

Income decomposition The SPI adjustment is applied to a singular income variable, but the FRS contains a number of components. Modifying gross income, but not modifying employment income, savings income, etc. breaks the link between these variables, which prevents researchers from conducting decomposition analyses.

Stratification There is no obvious justification for separate thresholds for pensioners and non-pensioners (and further, between households in Great Britain and Northern Ireland). The authors suggest these stratification choices were made in order to minimise methodological changes over time, for example as the survey expanded to Northern Ireland.

SPI lag The Survey of Personal Incomes is not routinely available at the same time as the Family Resources Survey (from which the HBAI microdata is derived). Therefore the SPI adjustment is applied to the HBAI dataset using a lagged SPI dataset, which may introduce additional inaccuracy.

3 Adjustments using administrative tax data

For the 2019 edition of the Households Below Average Income series, the ONS published details of the methodology used to tune the dataset with the SPI in [4]. They respond to some of the concerns raised by [2]:

[1] Previously, the DWP had not published its research underlying the methodology of the SPI adjustment

Pensioner stratification The authors show that high-income pensioners and non-pensioners are both under-represented in their respective populations but comparing the ratios of incomes at different quantiles, finding that a common threshold for both groups would fail to ensure that pensioners (who have lower income, on average) are sufficiently affected by the SPI adjustment.

Choice of income threshold The authors discuss possible justifications for a particular income threshold, mostly based on the quantile at which divergence between the FRS and SPI 'became an issue'. However, the choice to use a binary variable (rather than, for example, phasing in an SPI adjustment) here is arbitrary, and the authors do not address the reasons why this choice was made.

SPI lag The authors acknowledge the issue of using SPI projections, rather than actuall outturn data, and examine the size of this effect. They find that revising recent survey releases with the actual SPI data later released changed the Gini coefficient of income inequality estimates by around 0.2 percentage points. This is considered to be small and therefore recommend against the need for the ONS to re-publish statistics when current SPI data becomes available.

4 Capital income imputation

The issue of income decomposition remained largely untackled until [6], in which the authors attempt to improve the reporting of a specific component of gross income which is more severely under-reported in the FRS than others: capital income. They first establish that income under-reporting is mostly due to this particular category by comparing individual income sources between the FRS and SPI, finding that the aggregates of non-capital income are around 100% of the totals for the SPI, while capital income is only around 40% as represented.

The authors present a novel observation about the instances where capital income is under-reported: the capital share of income in individuals is far less represented in the FRS than in the SPI (specifically, the number of individuals with a 'high capital share'), rather than simply a lack of high-capital-income individuals. They introduce a new method to correct for this under-capture: adjust the weights of high-capital-share individuals in order to match the totals in SPI data.

The authors find that the new method is largely successful at correcting for under-capture of capital income, and increases the Gini coefficient of FRS data by between 2 and 5 percentage points (applying the methodology to historical FRS data releases). However, they do not measure the changes to how well the FRS ranks against other aspects of the SPI.

5 Under-coverage of very low incomes

In [1], the authors examine the other end of the income spectrum, finding that very low-income households tend to spend much more than moderately low-income households in the Living Cost and Food Survey (a household survey with similar administration to the FRS).

The authors report a variety of evidence that income at the low end is misreported in the survey:

Missing benefit spending By comparing total reported receipt of benefits by recipients with aggregate spending figures published by the DWP and HMRC, the authors find that the FRS and LCFS consistently under-report benefit income by around 5%, and that this figure has become worse over the last decade, rising from 2.5% in 2000.

Sub-minimum wage reporting In the LCFS, individuals report both hours worked and annual earnings, enabling researchers to calculate the implied hourly wage. For 10.5% of individuals in 2009, this was below the legal minimum wage. Although this does not guarantee a breach of employment law,[2] the proportion is substantial and implies that either earnings are under-reported or hours worked are over-reported.

The authors use a model of consumption smoothing to determine whether the overly high spending (compared to income) for low-income households can be explained by lifetime consumption smoothing, but find that this is not the case.

6 Linking data directly to administrative data

All of the previously covered research into survey inaccuracy has identified a common question: how much of the survey error is due to non-response bias, and how much is due to measurement error? In [5], the authors attempt to quantify the measurement error of the FRS by linking individual households with data from the DWP's administrative records, using non-public identifiers. The process of linking is not perfect: respondents are asked for permission to link their survey data with administrative data, and some (around 30%) refuse. However, for each benefit, the authors were able to find the percentage of reporting adults for whom a link to an administrative data record could be identified, the percentage of reporting adults recipients for whom no link could be found, and the percentage of adults represented only by administrative data.

The authors find that these splits vary significantly by benefit: recipient data on the State Pension (SP) is highly accurate in the FRS (96% of SP reported recipients were represented by the FRS, 1% were only on the FRS and not on administrative datasets, and 3% were only on administrative datasets). At the

[2] Employers can count some in-kind benefits as payment towards the minimum wage, and there are other legal exceptions.

same time, around 62% of adults on the FRS who reported receiving Severe Disablement Allowance could not be identified in administrative data.

This appears to provide additional evidence that measurement error is significant, at least at the low-income subset of the surveys.

References

[1] Mike Brewer, Ben Etheridge, and Cormac O'Dea. Why are households that report the lowest incomes so well-off? *The Economic Journal*, 127(605):F24–F49, 2017.

[2] Richard V. Burkhauser, Nicolas Hérault, Stephen P. Jenkins, and Roger Wilkins. Survey under-coverage of top incomes and estimation of inequality: What is the role of the uk's spi adjustment? *Fiscal Studies*, 39(2):213–240, 2018.

[3] Office for National Statistics. Households below average income (hbai) statistics.

[4] Office for National Statistics. Top income adjustment in effects of taxes and benefits data: methodology. 2020.

[5] Stephen McKay. Evaluating approaches to Family Resources Survey data linking.

[6] Tahnee Christelle Ooms. Correcting the underestimation of capital incomes in inequality indicators: with an application to the uk, 1997–2016. *Social Indicators Research*, 157(3):929–953, 2021.