Enhancing household survey microdata accuracy using machine learning: literature survey

Nikhil Woodruff

Contents

1 Introduction **1**

2 Current approaches in economic surveys **1**
- 2.1 Under-coverage of high incomes 1
- 2.2 Adjustments using administrative tax data 2
- 2.3 Capital income imputation 2
- 2.4 Under-coverage of very low incomes . 3
- 2.5 Linking data directly to administrative data 3
- 2.6 Linear programming 4

3 Applicable machine learning techniques **4**
- 3.1 Gradient descent 4
- 3.2 Random forest models 4

4 Conclusion **5**

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 Current approaches in economic surveys

It is well known that household surveys produce inconsistent results to other data sources, such as administrative databases. Given the nature of how surveys are conducted (households must first consent to an interview, and secondarily must answer truthfully to questions asked), this inaccuracy can be introduced either by sampling error or measurement error (likely both, to some extent). Over the last few decades, household surveys have become the dominant tool in measuring and projecting economic impacts of policy changes, and as such, there has been a great deal of research into improving survey accuracy.

2.1 Under-coverage of 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.[8] 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 [4], 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.

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

2.2 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 [9]. They respond to some of the concerns raised by [4]:

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 actual 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.

2.3 Capital income imputation

The issue of income decomposition remained largely untackled until [17], 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.

2.4 Under-coverage of very low incomes

In [3], 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.

2.5 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 [16], 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 same time, around 62% of adults on the FRS who reported receiving Severe Disablement Allowance could not be identified in administrative data. There are multiple possible reasons for this, and they vary by

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

benefit: the recipient population is often confused or mistaken when answering questions about their benefits, and this is more acute for age- or disability-related benefits. This appears to provide additional evidence that measurement error is significant, at least at the low-income subset of the surveys.

2.6 Linear programming

Linear programming, a mathematical technique for solving linearly constrained optimisation problems, is commonly used to determine survey weight values, where the criteria are defined maximum deviations from top-level demographic statistics. In [14], linear programming methods are used to determine the optimal weights for the Family Resources Survey, according to limits on how far apart the FRS aggregates can be from national and regional population estimates. In both of [5] and [13], tax models apply a linear programming algorithm to solve for weight adjustments satisfying a combination of tax statistic deviation constraints, and weight adjustment magnitude limits.

3 Applicable machine learning techniques

There are several reasons why machine learning techniques are well-suited to the task of survey imputation. The most fundamental justification is in its context-agnostic nature: machine learning approaches do not require assumptions specific to the field they are applied in, unlike the current approaches to survey accuracy improvement (for example, the percentile adjustment methodology in [9], which explicitly partitions households into 'rich' and 'non-rich' using arguably arbitrary definitions). In other domains, for example image classificaion, a move away from prescriptive methods towards loss function minimisation has seen substantially improved accuracy and robustness.[15]

3.1 Gradient descent

Gradient descent is a technique for finding parameters which minimise a loss function, by iteratively updating the parameters in the direction of the steepest negative gradient.[1] This is a highly common technique in machine learning, and is used in a variety of contexts, most notably as the foundation for training artificial neural networks. It relies on no domain-specific assumptions other than those present in the definition of the loss function, enabling it to be applied to a wide range of problems.

Several variations of gradient descent have emerged over the years which achieve more efficient training procedures: stochastic gradient descent steps in the direction of an *estimate* of the gradient using individual training examples, rather than loading the full dataset.[20] Mini-batch gradient descent represents a compromise between batch (full-dataset) and stochastic gradient descent, by iterating parameters using fixed-size subsets of the training data.[11]

As well as gradient calculation methods, optimisation algorithms have revealed significant accuracy and efficiency improvements by defining behaviours for hyper-parameters such as the learning rate (the velocity at which parameters follow the gradient). These include Adam,[12] AdaGrad,[7] and RMSProp.

Gradient descent could feasibly be applied to survey accuracy problems, since it requires only a loss function that is differentiable with respect to the parameters being optimised. In the context of survey accuracy, a loss function could be defined as the squared errors of individual aggregate statistics between official sources, and a survey, which would be continuously differentiable over the weights of individual household records.

3.2 Random forest models

Random forest models are a type of ensemble learning technique, which combine the predictions of multiple decision trees to produce a more accurate prediction than any individual tree.[2] The decision trees are trained on a subset of the training data, and the predictions of each tree are combined using a voting system. Although its introduction is far less re-

cent than more modern innovations in the field of neural networks (for example, artificial neural network variants[18] or transformers[19]), random forest models have shown consistently high accuracy across a wide range of domains, remaining competitive with the most recent techniques.

This type of model has been applied (to a limited extent) in the context of policy analysis, and have shown superior performance in prediction tasks to logit and other model types.[10]

There are several reasons why random forest models might outperform neural networks in predicting survey microdata values from other attributes (for example, predicting employment income from demographic variables), but the most natural reason is that tax-benefit law, which heavily influences financial decisions, is more similar in structure to a random forest than a neural network. For example, in [6] the authors found that capital gains variables are 'unnaturally' distributed in order to respond to incentives set by particular tax law parameters.

4 Conclusion

Current methods of enhancing surveys are largely effective at improving the accuracy of survey data on narrowly-defined subdomains (such as high-income analysis), but rely on explicit assumptions and are often not completely successful at bringing household surveys to the same level of accuracy as administrative data, especially at the low end of the income spectrum. Machine learning techniques such as random forest model imputation and gradient descent have shown promise in adjacent fields to public policy analysis, and could serve as more generalisable and effective replacements for the existing survey improvement methods.

The body of research on current methods for improving survey data is useful for examining how effective specific approaches were in improving a survey's answer to a narrow domain (for example, how adjusting income values improved the Gini index of income inequality), but there is little research on how each of the current methods, and any of the machine learning methods presented here, affects the overall picture of accuracy for a survey data. The reason for this is that many accuracy goals are orthogonal to each other: for example, improving the coverage of high taxable incomes might improve a survey's estimate of total income tax liabilities, but if it achieves this by overestimating employment income compared to dividend income, then a survey's estimates of payroll and dividend tax liabilities might each separately be reduced.

An implementation a general survey accuracy loss function that takes into account all (or as many as is feasibly possible in the scope of this project) of these accuracy targets, as well as implementations of both current and potential methods of data manipulation, would allow for a more comprehensive comparison of the effectiveness of each method.

References

[1] P. Baldi. Gradient descent learning algorithm overview: a general dynamical systems perspective. *IEEE Transactions on Neural Networks*, 6(1):182–195, 1995.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] 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.

[4] 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.

[5] Tax Policy Center. Tax model documentation. 09 2022.

[6] Tim Dowd and Robert McClelland. The bunching of capital gains realizations. *National Tax Journal*, 72(2):323–358, 2019.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of*

Machine Learning Research, 12(61):2121–2159, 2011.

[8] Office for National Statistics. Households below average income (hbai) statistics. 1992.

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

[10] Barbara Jarmulska. Random forest versus logit models: which offers better early warning of fiscal stress? *European Central Bank Working Papers*, 05 2020.

[11] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Mini-batch gradient descent: Faster convergence under data sparsity. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2880–2887, 2017.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Policy Simulation Library. Tax-data model documentation. *GitHub*.

[14] Charles Lound and Peter Broad. Initial review of the family resources survey weighting scheme. *Office for National Statistics*, 06 2013.

[15] Dengsheng Lu. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28:823 – 870, 03 2007.

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

[17] 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.

[18] Wolfgang Schöllhorn and Jörg Jäger. A survey on various applications of artificial neural networks in selected fields of healthcare. *Neural Networks in Healthcare: Potential and Challenges*, pages 20–58, 01 2006.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, and Jimmy Ba. An empirical study of stochastic gradient descent with structured covariance noise. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3621–3631. PMLR, 26–28 Aug 2020.