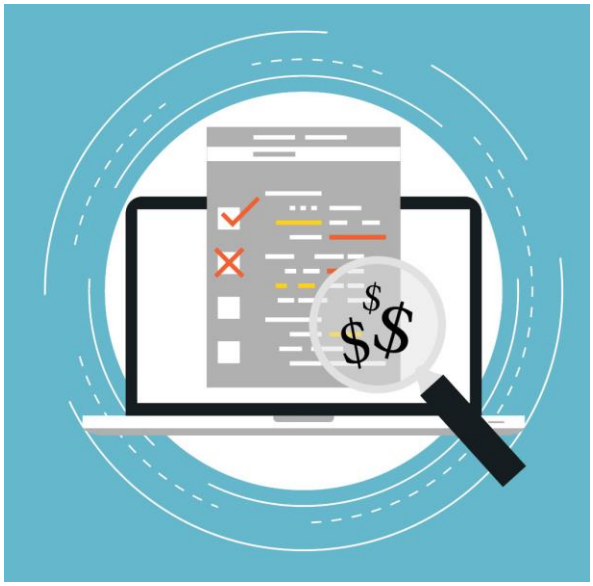
Methodological News

A Quarterly Information Bulletin by ABS Methodology Division

SEPTEMBER 2017 EDITION

Contents



Microsimulation and Non-Linear Cost Modelling to Predict Effects of Increasing Online Response on Survey Cost and Optimal Clustering

Online collection of household survey data is becoming an increasingly important mode for official statistics. However, predicting the cost of household multistage surveys with a meaningful online response component can be challenging with traditional linear cost models, as the relationship between cost and online response levels is expected to be non-linear. A lack of available data on costs incurred under similar levels of online uptake in previous collections can also be limiting.

Microsimulation of survey enumeration can be useful for predicting possible cost outcomes and has previously been used by the ABS for predicting optimal cluster sizes. The ABS has recently explored the use of microsimulation to predict the costs of survey enumeration at a range of online response levels under a typical cross-sectional multistage household survey design. Microsimulations considered the likely distribution of mode choices by respondents nationally, and interviewers' behaviour as they virtually enumerated their allocated samples, including decisions about approaching households, time and distance costs incurred, and the results of each approach. Predicted cost savings increased non-linearly with increasing online uptake for several cost components related to - interviewers' travel time and distance. That is, substantial cost savings did not occur until relatively high levels of online response for some cost components. However, when there was also high telephone interviewing, cost savings were more substantial at lower levels of online response.

The ABS has also recently developed non-linear cost models to predict both total cost savings and optimal cluster size, given online response rate, within block homogeneity, and ratio of costs between first and second stages. Within block homogeneity is a measure of how similar households within a cluster are to each other in key survey characteristics, e.g. households in the same neighbourhood may tend to report similar levels of income. The cost ratio is a measure of expenses incurred by an interviewer in reaching a cluster of selected houses (first stage), compared to the cost of enumerating each household once they are already in the neighbourhood (second stage). In the models, each household's choice to respond online was treated as a random event, so the probability that an entire cluster responds online was expressed as the likelihood of online response, raised to the power of the cluster size. Predicted first stage enumeration

costs were reduced by the number of whole clusters that do not require an interviewer visit, and second stage enumeration costs were reduced by the online response rate.

Cost savings estimated from these models agreed very closely with the results of the microsimulations. Optimal cluster size predictions showed some interesting features. Declustering was generally not cheapest until relatively high levels of online response (e.g. 70-90%). Where the cost ratio was lower, as in an urban area, the predicted optimal cluster size was smaller and declustering was efficient at lower levels of online response. In a cost profile more typical of a rural area, larger clusters were preferred and declustering was not cheapest until quite high levels of online response. Similarly where clusters were less homogeneous, a larger cluster size was preferred and declustering was not ideal until very high levels of online response. At up to moderate levels of online response, optimal cluster sizes were reasonably stable. However, if the within block homogeneity was low and/or the cost ratio was high, optimal cluster sizes could be somewhat larger at moderate compared to low levels of online response, before declustering became efficient.

The ABS intends to use these modelling and microsimulation methods, along with intelligence from field experience where available, to inform expectations in survey planning of the possible implications of differing levels of online response for cost and survey design.

For more information:
Susan Shaw
Susan.Shaw@abs.gov.au

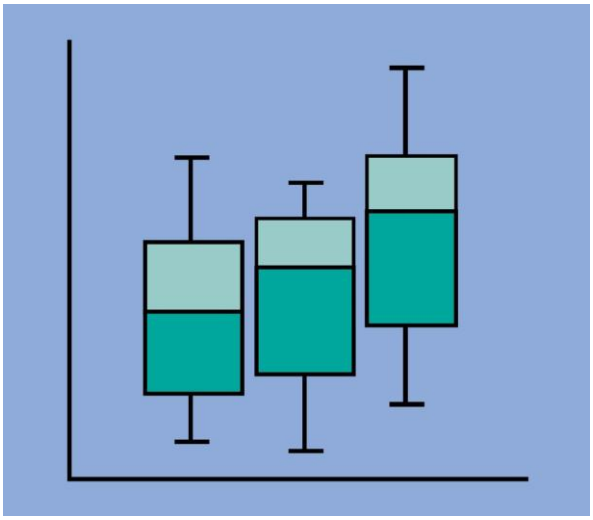
Replicate Variance Estimation for ABS Business Surveys using Permanent Replicate Identifiers

The transformed ABS output estimation capability will provide variances using a replication approach. There is little change to current methods for replicate variance estimation, except that they will be supported by a permanent set of replicate identifiers for each survey unit, rather than replicates being generated at the time of variance estimation. This approach simplifies the correct estimation of variances based on multiple time points and will allow for integrated outputs based on multiple surveys.

Most ABS household and business surveys are conducted using complex survey designs and many produce statistics that are complex functions of elementary estimates (e.g. ratios, differences of ratios). Re-sampling methods have been chosen as the preferred method of variance estimation for ABS household and business surveys, since the alternative linearisation method can be quite cumbersome to implement for complex survey designs, as it requires the derivation of separate variance formulae for each non-linear estimator.

In ABS business surveys the current variance estimation method used is the without replacement rescaled bootstrap procedure (Chipperfield and Preston, 2007), while in ABS household surveys the primary variance estimation method used is the delete-a-group jackknife method (Kott, 1998). While the bootstrap procedure is still considered the most suitable method of replicate variance estimation for ABS business surveys, the

current application of the rescaled bootstrap procedure has several less than ideal properties, including the requirement for multiple sets of replicate identifiers for various situations and the underestimation of the movement variances in particular situations.



An alternative application of the rescaled bootstrap procedure is that it will use a permanent set of replicate identifiers over time and across surveys. This permanent set of replicate identifiers will be generated from sets of replicate identifiers for a fixed number of replicate groups, since it will be more efficient from an operational perspective. The alternative application of the rescaled bootstrap procedure will overcome the issues associated with the current application of the rescaled bootstrap procedure.

The permanent sets of replicate identifiers for the fixed number of replicate groups in the bootstrap procedure will be generated from a Hadamard matrix H of order 256 to yield first-order and second-order balance. The scaled bootstrap weight method (Saavedra, 2001; Beaumont and Patak, 2012) will be incorporated into the bootstrap procedure to avoid negative bootstrap weights.

The delete-a-group jackknife method will also use permanent sets of replicate identifiers. One permanent set of replicates will be attached to Base Frame Units (BFU) on the area based private dwelling frame, and to dwellings on a list based non-private dwelling frame. Another permanent set of replicates will be attached to dwellings on the ABS address register.

For more information:
John Preston
John.Preston@abs.gov.au

Methods for making use of Transactions Data in the CPI

The Australian Consumer Price Index (CPI) is a robust measure of household inflation. The ABS produces the CPI each quarter using price quotations sourced via personal visits, online and telephone collection, and from administrative data sources (ABS, 2016a). One type of administrative data used is transaction records captured by retailers at the point of sale (transactions data).

Since 2014, prices from transactions data have accounted for about 25% of the weight of the CPI. At present, the ABS uses each retailer's transactions data to source prices for a sample of products that were formerly collected via personal visits to the corresponding retailer. This approach reduces the cost of (in-person) price collection. The prices derived from transactions data are also more representative of the prices actually paid by consumers than price quotes collected via one or a few personal visits.

The ABS has been assessing methods for maximising our use of the information on transactions datasets to enhance the CPI. Our goals have been to expand the product sample to incorporate all of the products that appear on these datasets, and to use the revenue information on these datasets to weight products in accordance with their economic importance. This information yields a more complete picture of the products purchased by consumers, and of how consumers substitute between products as prices and preferences change over time.

This information is also more dynamic than the samples traditionally used in the CPI, in the sense that products may appear and

disappear rapidly, and product revenues may change dramatically from one period to the next due to sales. In this context, the bilateral index (estimation) methods traditionally used in the CPI—which compare prices between two periods—can yield implausible results. Ivancic, Diewert and Fox (2011) suggest using multilateral index methods instead, which simultaneously compare prices over a window of more than two periods.

Over the last few years, the ABS has conducted extensive testing of a range of multilateral index methods. As there is no universally endorsed "best" multilateral method, we have developed a framework to guide this assessment. We have also assessed further methodological changes necessitated by the adoption of multilateral methods, including methods for splicing multilateral comparisons onto previously published index levels to extend the index, and minor changes to the index structure. To resolve methodological issues, we have collaborated with international experts and

with our counterparts at other statistical offices, and consulted with key stakeholders. Through this assessment, we have found the GEKS-Törnqvist multilateral method and the "mean splice" method for extending the index to be most suitable for our purposes—for more details, see (ABS, 2017).

The ABS plans to use multilateral methods to make greater use of transactions data in the production of the official CPI from the December quarter 2017 onwards.

References

Consumer Price Index: Concepts, Sources and Methods.
cat. no. 6461.0.
Australian Bureau of Statistics, Canberra

Making Greater Use of Transactions Data to Compile the Consumer Price Index.
cat. no. 6401.0.60.003.
Australian Bureau of Statistics, Canberra.

Holt, M., Webster, M & Pham, H (2017) An Implementation Plan to Maximise the Use of Transactions Data in the CPI.
cat. no. 6401.0.60.004.
ABS, Canberra.

Ivancic, L., Diewert, E. W. & Fox, K. J. (2011) Scanner data, time aggregation and the construction of price indexes.
Journal of Econometrics, 161, 24-35.

For more information:
Michael Webster
Michael.Webster@abs.gov.au

The Intelligent Coder: Developing a Machine-Learning Classification System

The ABS has developed the Intelligent Coder, a text classification application suited to the needs of a National Statistical Office (NSO) in classifying short free text responses to large classification hierarchies, such as ANZSCO or ANZSIC, the Australian and New Zealand Standard Classification of Occupation and Industry Classification respectively.

A large amount of effort is expended by NSOs in developing complete hierarchical descriptions of statistical classifications of interest, like industry, occupation, education, commodity, language or country of original. However it is unreasonable to expect survey respondents to be able to volunteer their relevant code. It falls to the NSO itself to receive respondents' descriptions of their relevant characteristics and map these descriptions to the standard classification code.

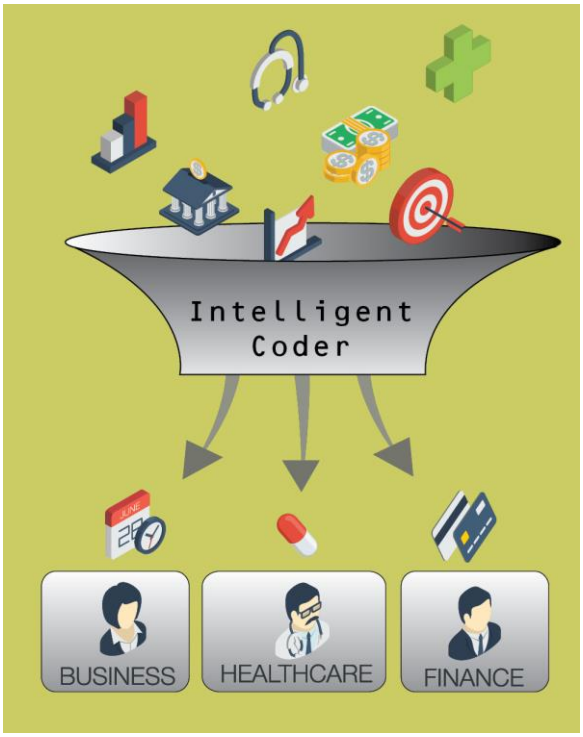
The original and most widely accepted way of mapping descriptions to classifications is using clerical coding: where officers are trained to have a full understanding of a set of classification hierarchies. These officers then manually assign classification codes to responses. Clerical coding is expensive and time-consuming, so automated solutions must be pursued.

The ABS has long used an index-based coder for automated classification. This involves the creation of an index file: a set of rules that map the presence of particular words and phrases to the code that should be assigned. This is an attempt to mechanise the heuristics

that a clerical coder might use to assign codes, and it succeeds in speeding up the classification process, as a large numbers of records are able to be classified very quickly.

The identification of patterns in responses and codifying these patterns is a procedure for which there are automated options; instead of manually creating an index file an automated procedure could be used which, given a set of examples, determines the optimal rules for classification. This allows the creation and refinement of index files to proceed much more quickly, as long as coded example records exist. The Intelligent Coder is this solution.

The Intelligent Coder represents text as points in vocabulary space, implements a hierarchical multi-class classification algorithm, and replicates the classification algorithm to ensure that generalisation to unseen data can be judged.



Text responses are processed to a numeric vector by the bag-of-words approach, where a vocabulary of all unique words is listed from the text data available. Then an individual text record is represented as a binary vector of the same length as the vocabulary list, with 1 for each vocabulary word that is contained in the record, and 0 for words that are not. This can be thought of as representing a record as a point in vocabulary space. The bag-of-words approach does not respect the order, the importance, or the context of words in a record, but the presence or absence of words captured by the bag-of-words approach probably captures most of the distinguishing information in records – descriptions provided by respondents tend to be semantically simple and terse.

In lieu of implementing a natural multiclass classification algorithm, the Intelligent Coder classifies records by combining a set of binary support vector machine classifiers. Specifically, a record begins at the root of the classification tree and is recursively classified to the most likely child node, where “most likely” is judged by two factors: the set of binary classifiers that combine to classify to the set of child nodes, the confidence that the coder has for that child node. The binary classifiers can be combined by creating a binary classifier for all pairs of child nodes, and a record is assigned to the child node that is assigned by the most classifiers.

The confidence that the coder has is created by bagging: resampling from the training data with replacement and creating an independent coder for each resample. These coders then vote on each record – this vote is used to evaluate the confidence of the classification.

The Intelligent Coder was trained and tested on a set of text responses to questions about occupation and industry collected between 2013 and 2015, which had been classified using an index-based coder with clerical coding for remaining records. Initial results showed that with little effort the Intelligent Coder could increase the rate of automated coding by 20% without a degradation in accuracy.

For more information:
Rory Tarnow-Mordi
Rory.Tarnow-Mordi@abs.gov.au

How to contact us and Email Subscriber List

Methodological News features articles and developments in relation to methodology work done within the ABS Methodology Division By its nature, the work of the Division brings it into contact with virtually every other area of the ABS. Because of this, the newsletter is a way of letting all areas of the ABS know of some of the issues we are working on and help information flow.

We hope the Methodological Newsletter is useful and we welcome comments.

If you would like to be added to or removed from our electronic mailing list, please contact:

Nick Husek
Methodology Division
Australian Bureau of Statistics
Locked Bag No. 10
BELCONNEN ACT 2617

Email: methodology@abs.gov.au

The ABS Privacy Policy outlines how the ABS will handle any personal information that you provide to us.