Methodological News
A Quarterly Information Bulletin

ABS Methodology Division

Articles

Introduction of Standardised Imputation Methodology for the Building Activity Survey

The Building Activity Survey (BACS) currently uses a custom-built method to impute values for non-respondents. This method was developed in around 2005 as the previous method was found to introduce systematic bias into preliminary estimates. However, in an ideal world, BACS would be able to use a standard, corporately supported imputation tool as long as estimate quality is maintained. The benefits of this include less reliance on in-house expertise, reducing cost and risk of maintaining a non-standard system and reduced complexity of future transitions to new infrastructure.

Investigations have shown that relatively simple methods perform to the same standard as more involved methods - including the current method, which uses a complex regression model not able to be performed in standard corporate infrastructure. In particular, the Respondent Mean method was recommended. This method calculates imputes based on the mean values of respondents with similar characteristics to the non-respondents, initially at a very fine level, and falling back to broader classes if required. Only two data items (probability of commencement, and proportion of work done) are imputed, and all other data items are able to be derived from these together with other information already known about the non-respondent. This method results in revisions between preliminary and final estimation that are distributed very similarly to the current method. Parallel estimates produced over a number of quarters show that the resulting estimates closely aligned (within two standard errors) with no evidence of any systemic bias or change in bias.

Recently, Business Statistics Methodology, the Construction Business Statistics Centre, and Technology Applications Delivery have been working together to refine, specify, test and implement the new methodology, which has resulted in some minor changes, such as introducing an additional fallback stage to make the imputes more robust. The new method is expected to go live for the September 2015 cycle providing that final testing is passed. The transition to the new method will be closely monitored to ensure it is robust to any shift in respondent behaviour and that published estimates continue to be of the same high quality.

Further Information

For more information, please contact Melanie Black (methodology@abs.gov.au)

A New Analytical Platform to Explore Linked Data

The Emerging Data and Methods (EDM) section has recently completed research work on identifying true and spurious firm deaths using a combination of traditional and new techniques. In particular, a prototype Graphically Linked Information Discovery Environment (GLIDE) was created from tax data sources and ABS Business Register data using a network-oriented Semantic Web approach. This enabled the application of multilevel modelling and Bayesian Network

methods to the analysis of the network of employer-employee interactions among firms and people. The results of this work have been released as an ABS research paper, *A New Analytical Platform to Explore Linked Data*, available from the website.

The Semantic Web framework provides an alternative approach to data representation, linking and retrieval that can unlock the full potential of interconnected and multi-dimensional datasets. Instead of organising datasets in a structured row-column tabular form, the Semantic Web approach models information in the form of a network of entities and relationships. The relationships are given strong computable semantics by precisely specifying their logical properties in a machine-interpretable format.

The Semantic Web approach opens up new avenues of data exploration, visualisation and network analysis. One example of this has been demonstrated in the prototype GLIDE by using it to derive network statistics and create models to distinguish true firm deaths from spurious ones. The ABS has an established process for identifying firm exits, but is not able to distinguish the type of exit – whether it is due to restructuring, merger/takeover or a genuine death.

Both multilevel logistic regression and Bayesian Network (BN) models were used to distinguish true and spurious firm deaths. Multilevel models were developed both with and without network statistics, with the data partitioned into modelling (training) and prediction (test) subsets to assess the quality of out-of-sample predictions from the models. It was found that the model with network statistics performed substantially

better (95% accuracy vs 74% accuracy). Significant variables were then incorporated in a BN model. This approach took account of the relationships between all the variables, achieving similar prediction accuracy with a subset of variables, and also handling observations with missing variables in the test data. The intention was not to compare both methods on the prediction outcomes but to build on the multilevel modelling results to provide a statistical framework for the BN model.

The analytical results have shown that it is important to account for spurious firm deaths for statistical production. This is because failure to account for spurious firm deaths can result in continuing enterprises being incorrectly classified as deaths, and as a result it can affect the statistical quality from the perspectives of survey frame and accuracy of the statistics. The conclusion is that the Semantic Web is a useful approach for statistical purposes, and that network analysis can be used to effectively distinguish true and spurious firm deaths.

Further Information
For more information, please contact Andreas Mayer or Joseph Chien (methodology@abs.gov.au)

A Standard Estimation Approach that Adjusts for Non-Response

The estimation systems used by the ABS for its suite of household and business surveys are based on the theory of the generalised regression (GREG) estimator. While these

generalised systems guarantee a high level of standardisation in the production of statistics, there is still great diversity in the estimation methods used for the treatment of non-response in sample surveys. There can be clear benefits in terms of quality and efficiency from adopting consistent estimation methods across all sample surveys. A number of alternative calibration estimators which could be used for the treatment of non-response in sample surveys were evaluated, and it was recommend that the response propensity calibration estimator, an estimator that is relatively simple to implement and which performed well across various simulations, should be adopted.

The rules for the choice of auxiliary variables to be included in the response modelling, along with how this choice interacts with strategies for making the predictions from the response modelling more robust, were examined. Firstly, the over-specification of the response propensity model can potentially cause an increase in the standard errors of the regression parameters, which can lead to inefficient estimates. One possible solution to overcome this problem is to use model selection methods to choose a parsimonious statistical model from a set of candidate statistical models (i.e. eliminate irrelevant explanatory variables in the response propensity model). Secondly, the mis-specification of the response propensity model can be a potential cause for extreme response propensity weight adjustments (i.e. inverse of estimated response propensity), which can lead to inefficient estimates, as well as potentially unreasonable estimates particularly for domains. One possible solution to overcome this problem is to trim the extreme response propensity weight adjustments.

The proposed modified boxplot method (based on Tukey's boxplot method for the identification of outliers) appeared to provide a suitable treatment for trimming extreme estimated response probability weight adjustments. The AIC selection method (an information criteria method based on the likelihood function of the model plus a penalty term) also appeared to provide a suitable semi-automated method for choosing auxiliary variables in the logistic regression. The proposed modified boxplot method and the AIC selection method performed well across various simulations and can be fully automated without the need for any user interaction.

The value of models that specifically incorporate the impact of follow-up strategies on the probability of unit response was also considered. Where different follow-up approaches are assigned randomly, an indicator of the follow-up strategy should not be included in estimation. However, where the follow-up is assigned deterministically, consideration should be given for inclusion in estimation, particularly where it identifies a particular type of unit that may have distinctive values for the study variables, so that that type of unit is appropriately represented in the estimates.

Further Information
For more information, please contact John Preston or Philip Bell (methodology@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:

Peter M Byron
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.