Australian Bureau of Statistics

METHODOLOGICAL NEWS

A QUARTERLY INFORMATION BULLETIN FROM THE METHODOLOGY AND DATA MANAGEMENT DIVISION

December 2010

Benchmarking Small Area Estimates

The demand from users for regional estimates has been growing rapidly in Australia and throughout the world over the last 20 years. This demand is driven by the increasing requirement on policy makers to formulate evidence based policy and deliver programs that are cost effective, responsive to a changing world and targeted to relevant areas.

Small area estimation (SAE) is concerned with developing methods for calculating reliable estimates for geographic areas or domains that are sample deprived. In most small regions or domains, sample sizes are often so small that estimates calculated using the conventional design based estimation methods (such as survey weighted totals) are subject to very high sampling errors, making these estimates statistically meaningless. SAE can often overcome this problem by using a statistical model that relates survey data to available auxiliary data. In other words, SAE overcomes the small sample problem by borrowing strength from auxiliary information and similar units in other areas. The ABS has been involved in producing small area estimates (SAEs) for many years and completed projects have included: disability, health, Indigenous health and water use and SAE feasibility studies of labour force status and household net wealth.

Users expect the small area estimates provided to them to be both consistent and coherent, with respect to published official statistics. Coherence is the seventh dimension of the ABS Data Quality Framework and is defined as "the internal consistency of a statistical collection, product or release, as well as its comparability with other sources of information, within a broad analytical framework and over time". Small area estimates that are not consistent and coherent will struggle to gain credibility with users. Other advantages in benchmarking SAEs are that it can reduce the impacts of model misspecification (including poor quality of auxiliary data) as well as ameliorate the effects of influential or outlier data points.

Recently, the Analytical Services Unit (ASU) developed a methodology for producing SAEs that are guaranteed to sum to higher level published estimates. Measures of accuracy for these "benchmarked" estimates were also produced. The work was carried out on Labour Force Survey data using Centrelink and Census data as auxiliary variables in a logistic binomial model with random effects. The method involved adding a constraint to the standard log likelihood function and then using the Lagrange multiplier method to derive a maximum penalised quasi likelihood algorithm to estimate model parameters. Four different constraint levels were trialed, these being the Australian, state, state by capital city / non-capital city and dissemination region levels. Relative root mean square errors (RRMSEs) were calculated using the parametric bootstrap.

The results showed that when benchmark constraints were set at the Australian or state levels, SAEs and their RRMSEs were indistinguishable from the corresponding unconstrained estimates. This occurred because the unconstrained model produced SAEs that came close to summing to these published estimates and the sampling errors on these broader level benchmarks were quite low. However, when the constraint level was set at either the sub state or dissemination region level, benchmarked SAEs varied considerably and RRMSEs were much higher. This suggests that it is not advisable to constrain SAEs to benchmark levels that are themselves subject to high sampling error.

Further information can be obtained from Daniel Elazar on (02) 6252 6962 or daniel.elazar@abs.gov.au.

Modelling Business Provider Response Behaviour: A Survival Analysis Approach

Understanding business survey data provider's response behaviour is an important consideration for selecting an efficient data collection procedure. If providers are contacted too infrequently, they may not respond; however if they are contacted too frequently, the data collection procedure may be inefficient. In more extreme cases, contacting a provider too many times (or via inappropriate channels) leads to excessive respondent burden and potentially, provider complaints and non-response. Therefore, in order to allocate resources effectively, we need to understand a provider's reaction to our attempts to obtain their cooperation.

ABS business surveys are typically conducted as mail-out, mail-back collections, which are supplemented by both written reminder letters and telephone follow-up calls. While every provider who has not yet responded will typically receive a reminder letter at the same time, telephone contact is more costly and therefore is prioritised on the basis of the provider's significance to the estimate and on their expected level of cooperation. For example, a very large business with a poor response history is much more likely to be contacted than a smaller business with a good response history.

The survey response process can be regarded as a survival process to attempt to answer: what is the fraction of the whole sample of businesses who will response before a certain time? We commence by posting out forms, which the provider is able to return at any time. After a certain period of time, if they are still not responding, we begin intervention - telephone and written reminder contact. The probability of response, and time to response, are dependent on both these interventions (which are time-dependent covariates, as they change value through the collection period) and on demographic information such as business size (which are generally fixed). We cannot simply ignore the impact of time, since we are interested in selecting appropriate timing for the follow-up, and also because we have censored data. In our case, censored data arises because although a response would eventually occur, we do not observe the response time, either because we have ceased follow-up, or because the covariate values have changed. Therefore, we use a survival analysis model where we are modelling the time to response.

Some key results of this analysis (for a selection of ABS business surveys) include:

- Compared with the week of a written reminder, businesses are more likely to respond the following week, but less likely to respond in any week after that;
- A business that receives a call in a given week is much more likely to respond in that week than a business that did not receive a call;
- If a business does not respond in the week in which it was called, that call continues to boost their probability of response in subsequent weeks, but at a reduced rate;
- Businesses with a previously good response history are much more likely to respond than those which are new or have a poor response history; and
- Businesses in South Australia, Tasmania and the ACT have higher probabilities of response than those in other states.

Further research is underway to refine the results and apply these to selecting the most appropriate forms of follow-up procedures.

For further information, please contact Melanie Black on (02) 62527241 or melanie.black@abs.gov.au.

Methods for Imputing Age in the Census

Imputing age for partial non-respondents in the Census is a difficult problem. Age is related to almost every item in the Census, including education, employment, disability and income. It is important to impute an age that is consistent with other responses the respondent may have given. It is also important that the age of a respondent be consistent with their position in their household and with the ages of other members of the household (so that parents have a realistic age with respect to their children and so on).

The current method used for Census imputation has been used for many Censuses and has some unfortunate properties. It depends on age distributions from the previous Census which are five years out of date, and it has a tendency to produce too many imputed values around 'threshold' ages associated with particular characteristics (for instance at age five when most children start school). This method was scheduled for review and replacement in the 2011 Census, but budget cuts delayed this review. The Methodology Development Unit (MDU) is now looking at possible methods for replacing it in the 2016 Census.

Donor based methods (hotdecking) are in use for imputing other items in the Census. These work well for imputing age in non-contact dwellings, as ages drawn from a responding Census dwelling will automatically be consistent within the dwelling. Hotdecking is not suitable for the case of partial non-response because it is very difficult to find donors that are consistent with all the additional information available for respondents.

Because of the complexities associated with age imputation we are looking instead to emerging techniques from the field of data mining or predictive analytics. These methods allow the construction of complex models with minimal input from the analyst. Already we have conducted some experiments with a method called Bayesian Additive Regression Trees (BART). This is a recent addition to the regression tree family of models in which many small trees are generated and then added together to give an overall model. The use of many small trees allows for the approximation of complex relationships including interactions and additive effects. There is a package that implements BART in R, which is very convenient. Unfortunately R does not handle large volumes of data well, so the testing that could be done on Census data was very limited. Also unfortunately the performance of BART in creating suitable age imputes was less than satisfactory. In particular it imputed a large number of ages that were inconsistent with the reported Year of Arrival.

While there are options for tailoring BART to give more suitable imputes the nature of the model underpinning BART means it is likely we will never be able to completely avoid these sorts of edit failures. We are currently looking into other methods.

Further information can be obtained from Claire Clarke on (08) 8237 7468 or claire.clarke@abs.gov.au

Progress with the AIC Review

The Annual Integrated Collection (AIC) is a key data source for measures of Australia's economic performance for each financial year. It replaced the previous ABS annual collection (the Australian Economy Wide Survey) in 2006-07.

In 2009, senior management of PLIES and MIG decided to review the performance of the AIC after the first two years' worth of estimates were produced (in respect of 2006-07 and 2007-08). Charles Aspden, a former Assistant Statistician of the National Accounts Branch, was brought in as a consultant to review the performance of AIC in mid-2009. His review was completed in a whirlwind three weeks; the opening paragraph of his Review report stated that "NAB found substantial differences in the economic picture portrayed by the AIC core and other data sources, most notably the ABS's quarterly surveys, both in terms of coherence within each of the two years and growth between them".

The desire to improve the accuracy of AIC estimates, plus enhance the AIC's coherence with other data sources such as the Quarterly Business Indicators Survey (QBIS), has been a driving force behind the efforts of the AIC Review Team in the last year or so. The role that MDMD plays is in providing methodological advice to the Review, in particular pursuing the key work identified by the AIC Review Team which wound up in July 2010.

Statistical Services Branch (SSB) within MDMD has been involved in a number of investigations in assisting the Review. In trying to improve the efficiency of the AIC sample (which currently stands at about 20,000 units selected from the Australian Business Register (ABR)), John Preston (SSB) is experimenting with using a modelling approach which uses Business Activity Statement (BAS) data from the Taxation Office (combined with other information from the ABR) to produce unit-level values for each unit on the AIC frame. This method allows for a flexibility of models for the estimation of key measures of financial and economic performance that are output by the AIC, thereby permitting a powerful use of the available auxiliary information.

Edward Szoldra (SSB) has also completed some work confronting wages estimates between AIC and QBIS. This work attempted to shed some light on the discrepancies in the estimates between these two collections. After adjusting for a number of known differences (such as scope and differences in estimation methodology), comparative estimates of annual movement between the AIC and the sum of 4 quarters of QBIS data showed that some industries had discrepancies beyond what is explicable by sampling error. Further investigations indicated some promising (though not definitive) potential reasons for the discrepancies (including issues around differences in annual and quarterly survey reporting).

In addition, more recent work within SSB has been in addressing issues around coherence as indicated in the September 2010 AIC Review Steering Committee meeting. These projects will include SSB working with NAB to ascertain more precisely their quarterly and annual data requirements in the construction of the National Accounts and some work around the effectiveness of the generalised regression (GREG) estimation methodology used in AIC (and whether it is a viable candidate for use in the QBIS). Lastly, SSB will undertake a project which aims at documenting NAB's Supply-Use Table transformations and how AIC and QBIS data flows through the accounts.

For further information regarding these projects, please contact Edward Szoldra on (02) 9268 4214 or edward.szoldra@abs.gov.au.

Dealing with a Break in Series - Job Vacancies

Seasonal adjustment is designed to remove predictable calendar (or seasonal) patterns from time series data; where exactly the same thing is measured at regular time intervals. The time series of the ABS Job Vacancies Survey (JVS) was broken in the five quarters from August 2008 to August 2009. During this period, ABS temporarily ceased the collection of the Job Vacancies data and, in November 2009, reinstated the survey. The break in series coincided with the Global Financial Crisis which made it difficult to predict the missing data using historical data.

Availability of long and unbroken series is key to undertaking time series analyses, including seasonal adjustments. In order to continue the production of seasonally adjusted and trend estimates for the Job Vacancies series, ABS decided to fill up the data gap using an econometric model. Conceptually, job advertisement was considered closely related to the job vacancies because they measure a similar concept. Initially, job advertisement time series from non-ABS sources were evaluated in the modelling exercise. However, there is no co-integration relationship between these series and the ABS Job Vacancies series, meaning that the job advertisement time series do not contain sufficient information about the Job Vacancies trend direction.

The Time Series Analysis (TSA) section found that there was a six-month lag between the business cycles of the Job Vacancy rate and the unemployment rate. The scale of changes in the Job Vacancy rate was consistent with the scale of changes in the unemployment rate. So TSA also evaluated autoregressive (two-quarter lag) models relating the Job Vacancy estimates to Employment estimates, the Full Time Equivalent estimates, and the Hours Worked estimates. After consultation with Treasury -- the main client who uses the job vacancy series in the TRYM for macroeconometric modelling -- TSA chose the last option.

TSA reintroduced seasonality and used these modelled estimates in the gap between measured JVS estimates. Thus series continuity is mostly restored.

The modelled data may be obtained from the TRYM Modeller's database, on the ABS website under cat no. 1364.0.15.003.

For more information, please contact Rachel Barker on (02) 6252 6183 or rachel.barker@abs.gov.au.

Quality Gates for the Mitigation of Statistical Risk

The Australian Bureau of Statistics (ABS) leads Australia's national statistical services, running hundreds of surveys and publishing thousands of pages of output every year. As with any large and complex organisation, problems with processes do arise and the ABS has suffered errors in our data in the past with varying degrees of impact on the public domain. Most errors are detected in-house before publication; however this has at times resulted in intense last-minute work to correct the problems leading to delays in the release of data. Other errors have only been discovered after release, resulting in re-issue of statistical output. As a result of these errors the ABS has endeavoured to instigate better quality management practices through the development and use of the risk mitigation strategy known as quality gates.

Quality gates are an organisational risk mitigation strategy the ABS has adapted to improve the early detection of errors or flaws in any part of statistical processes, be it collecting, processing, analysing or disseminating statistics. Quality gates are a powerful tool for improving an statistical organisation's ability to manage statistical risk by:

- providing explicit evidence relating to the statistical process at strategic places in the cycle to determine fitness for purpose of the process (and data) at that point in time; and
- improving knowledge management and information sharing of data relating to specific stages of a statistical process.

Each quality gate is a checkpoint at which an assessment of the quality of the process is made either qualitatively or quantitatively, to determine whether to proceed to the next stage of the process. This is achieved through the six components of a quality gate:

- Placement - placement refers to the points in statistical processes at which a quality gate should be implemented based on the risk associated with that given point;
- Quality Measures - quality measures are indicators which provide information about potential problems to allow for their early detection in a statistical process, e.g. response rates or data availability;
- Roles - roles involves assigning tasks and accountability to areas or people connected to quality gates, including an operational person (gate keeper), stakeholders and a sign-off person;
- Tolerance - tolerance or threshold refers to an acceptable level of quality for each quality measure, agreed in advance;
- Actions - actions are a set of predetermined responses if a tolerance level or threshold is met or not met which allow faster responses to arising problems; and
- Evaluation - evaluation is an examination of where improvements may be made to the quality gates in future cycles based on problems identified throughout the overall process.

Although the ABS is in the early stages of implementing quality gates, the impact has been very positive and we are therefore keen to promote their use to external organisations. To achieve this, the ABS will release an information paper that outlines the concept of quality gates and their six components in more detail, discusses the benefits of quality gates and provides examples to assist organisations to implement quality gates in their own statistical processes.

The information paper, 'Quality Management of Statistical Processes Using Quality Gates (cat. no. 1540)', will be released on Thursday, 23 December 2010. A home page icon will promote the paper on the ABS website from early December, and will later be linked to the paper upon its release.

Further information on the implementation of quality gates in a statistical process can be obtained from Andrew Doherty on (03) 9615 7038 or andrew.doherty@abs.gov.au, or from Narrisa Gilbert on (08) 9360 5283 or narissa.gilbert@abs.gov.au.

Harvard Professor Alan Zaslavsky Visits the ABS

Professor Alan Zaslavsky, an expert in health care policy and statistics, recently visited the ABS in conjunction with his trip to deliver the E.K Foreman Lecture at the Australian Statistics Conference (ASC) held in Perth last December 6-10.

Professor Zaslavsky is Professor of Health Care Policy (Statistics) in the Department of Health Care Policy, Harvard Medical School. His statistical research interests include surveys, census methodology, small-area estimation, official statistics, missing data, hierarchical modelling, and applied Bayesian methodology. He is a member of the Committee on National Statistics (CNSTAT) of the National Academy of Sciences and has served on CNSTAT panels on census methodology, small area estimation and race/ethnicity measurement, as well as several Institute of Medicine committees on measurement and reporting of health and of health care quality. He is a Fellow of the American Statistical Association and a National Associate of the National Academy of Sciences.

At the ASC, Alan was the keynote speaker, giving the biennial E. K. Foreman lecture on "Using Hierarchical Models to Attribute Sources of Variation in Consumer Assessments of Health Care"

When he visited the ABS, Alan gave a lecture on the above topic and also presented several well attended sessions, including on missing data; modelling cost and expenditure data; recent research in the analysis of inexactly linked data; and diagnosing imputation models by applying target analyses under posterior replications of observed data.

During the week-long visit to ABS, Professor Zaslavsky and a good number of ABS methodologists and analysts spent time discussing methodological issues. The ABS also invited analysts and modellers from the Australian Institute of Health and Welfare (AIHW) and the National Centre for Social and Economic Modelling (NATSEM) to discuss with Professor Zaslavsky, particularly on health data analysis, modelling and microsimulation studies.

For more information about Professor Zaslavsky's visit to the ABS, the E. K. Foreman lecture or any of the sessions Alan presented at the ABS, please contact Shaun McNaughton on (02) 6252 5125 or shaun.mcnaughton@abs.gov.au.

Professor Rubin to Visit ABS and Present Two Short Courses

The Australian Bureau of Statistics, CSIRO and the Statistical Society of Australia, Inc. (SSAI) are sponsoring two one-day short courses to be presented by Donald B. Rubin, Professor of Statistics at Harvard University.

Professor Rubin is one of the best known and most widely cited statisticians of the past forty years, with more than 350 articles published in more than thirty journals. He is also the author or co-author of several books which remain seminal works in their field. Prof. Rubin has lectured extensively throughout the Americas, Europe and Asia.

Prof. Rubin will be visiting the ABS and CSIRO in Canberra during the week of 10–14 January 2011 and CSIRO in Sydney during the week of 17–21 January 2011.

In Canberra, Prof. Rubin will present two one-day short courses at the ABS:

- A short course on Causal Inference in Observational Studies on Wednesday 12 January, and
- A short course on Imputation for Missing Data in Official Statistics on Thursday 13 January.

Ms Elizabeth Zell will co-present both courses with Prof. Rubin. Ms Zell is a distinguished mathematical statistician working for the Centres for Disease Control in Atlanta. Her current areas of research are missing data, including multiple imputation and a proper imputation evaluation; propensity score methods for causal inference and for believable conditional association; and vaccine preventable bacterial diseases.

Registrations are now being accepted for the short courses.

To register, please complete and forward the appropriate form from the SSAI website < http://www.statsoc.org.au/CPD18 >.

The registration fee will be waived for ABS and CSIRO employees, but registration will still be required as a guide to catering, and to ensure that the events are not over-subscribed.

More information on Prof. Rubin's visit may be obtained by following the links from the NSS home page < http://www.nss.gov.au >.

Data Confidentiality Symposium

The Australian Bureau of Statistics, CSIRO and the Statistical Society of Australia, Inc. (NSW) are sponsoring a one-day symposium on data confidentiality in Sydney on 18 January 2011.

National Statistical Agencies and other data custodian agencies and organisations currently face a challenge in balancing requests for access to data for research and policy development with privacy and confidentiality protection.

This symposium aims to provide an overview of data confidentiality issues and an introduction to current and emerging approaches. Topics to be covered will include:

- an overview of current, large-scale successful initiatives in Australia which make health and other personal data available for research;
- a review of techniques designed to confidentialise data before release to researchers;
- a review of the role of synthetic data methods; and
- a discussion of the role and design of remote access in future systems for balancing data access with confidentiality protection.

Speakers at the Symposium will include Professor Donald Rubin (Harvard University), Dr Christine O'Keefe (CSIRO), Mr Tim Hawkes (Statistics New Zealand) and researchers from the ABS Data Access and Confidentiality Methodology Unit.

The preliminary program may be accessed at
< http://www.cmis.csiro.au/conferences-seminars/DataConfSympProg.htm >.

There will be no charge for attendance at the Symposium. However prospective participants are asked to register in advance to facilitate planning. Please use the registration form at <http://www.cmis.csiro.au/conferences-seminars/DataConfSympReg.htm >.

More information on the Symposium may be obtained by following the links from the NSS home page < http://www.nss.gov.au >.

How to Contact Us and Subscriber Emailing List

The Methodological Newsletter features articles and developments in relation to methodology work done within the ABS Methodology and Data Management 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 placed on our electronic mailing list, please contact:

Valentin M. Valdez

Methodology & Data Management Division
Australian Bureau of Statistics
Locked Bag No. 10
BELCONNEN ACT 2617

Tel: (02) 6252 7037
Email: methodology@abs.gov.au