Methodological News
A Quarterly Information Bulletin
Australian Bureau of Statistics
ABS Methodology and Data Management Division
March 2013

Articles

An Alternative Approach to Measure Record Level Disclosure Risk in Micro-data

The ABS is required by legislation to ensure that no statistical outputs are released in a manner that is likely to enable the identification of a particular person or organisation. A key component of the ABS dissemination strategy is the release of micro-data files in the form of licensed Confidentialised Unit Record Files (CURFs), therefore each micro-data file produced must be assessed to ensure that the likelihood of identification is minimised.

The current assessment method of risk of identification in the ABS is through a number of manual tabulation procedures as well as the use of the Special Uniques Detection Algorithm (SUDA) program. This can sometimes be manually intensive.

The ABS is looking for a way to improve the assessment method by identifying a set of unit record risk measures that are:

- statistically valid and objective so that risk can be measured reliably
- consistent across CURFs so that we can properly assess relative risks
- in accordance with practical experience
- fast to calculate
- easy to interpret and apply.

One approach suggested by Elamir and Skinner (2006) is to use a log-linear model to estimate the probability that a unique record in the CURF is a match to a person known in the population. They assume that for each combination of key characteristics, the number of people in the population that have that combination can be modelled by a Poisson distribution. Using this assumption and log-linear modelling they provide an estimate for the probability described. A drawback of this method is that it can lead to biased estimates due to the presence of many zero counts. To overcome the supposed shortfalls of the log-linear model, Manrique-Vallier and Reiter (2012) suggest using a grade of membership model to calculate the same probability.

Data Access and Confidentiality Methodology Unit (DACMU) will be looking into the use of the log-linear modelling approach, grade of membership modelling approach and other relevant approaches (including SUDA output), analysing the accuracy and precision of these approaches for future CURF assessments. The results of current CURF assessment procedures will be used to validate the risk measures obtained from these new methods.

References

Elamir, E & Skinner, C. (2006) 'Record level measures of disclosure risk for survey microdata', *Journal of Official Statistics*, vol. 22, no. 3, pp. 525-539.

Manrique-Vallier, D & Reiter, J. (2012) 'Estimating identification disclosure risk using mixed membership models', *Journal of the American Statistical Association*, vol. 107, pp. 1385-1394.

Further Information

For more information on this work-in-progress, please contact Gareth Biggs (02 6252 6504, gareth.biggs@abs.gov.au)

Disclosure-Protected Inference with Linked Micro-data using a Remote Analysis Server

Large amounts of micro-data are collected by data custodians in the form of Censuses and administrative sources. Often, data custodians will collect different information on the same individual. Many important questions can be answered by linking micro-data collected by different data custodians. For this reason, there is a very strong demand from analysts, within government, business and universities, for linked micro-data. However, many data custodians are legally obliged to ensure the risk of disclosing information about a person or organisation is acceptably low. Different authors have considered the problem of how to facilitate reliable statistical inference from analysis of linked micro-data while ensuring that the risk of disclosure is acceptably low. The methodology area of ABS wrote a paper for its Methodology Advisory Committee (MAC) meeting in November 2012 that considered this problem from the perspective of an *Integrating Authority*.

An Integrating Authority is trusted to link the micro-data and to facilitate analysts' access to the linked micro-data via a remote server. A remote server allows analysts to fit models and view the statistical outputs without being able to observe the underlying linked micro-data itself. One disclosure risk that must be managed by an Integrating Authority is that one data custodian may use the micro-data it supplied to the Integrating Authority and statistical outputs released from the remote server to disclose information about a person or an organisation. The MAC paper measures the utility and disclosure risk of a proposed method in simulation and with a real example. The evaluations show that the protections prevent disclosure in a high risk scenario and have only a small impact on inferences for analysis involving moderate sample sizes.

The paper is available from James Chipperfield and may be made available on the Australian Bureau of Statistics website.

Further Information

For more information, please contact James Chipperfield (02 6252 7301, james.chipperfield@abs.gov.au)

Panel Data Modelling of Innovation and Flexible Working Arrangements

The Analytical Services Unit (ASU) has been undertaking an analysis of the relationship between innovation and flexible working arrangements, as part of a bigger project that looks into the various panel data analyses that can be undertaken using the Business Longitudinal Database (BLD). In particular, the analysis makes use of firm-level data from three waves of the BLD, namely the 2007-08, 2008-09, and 2009-10 waves, and focuses on small- and medium-sized enterprises. It examines the effects of flexible working arrangements on innovation, while controlling for the effects of other factors including competition, ICT intensity, collaboration, and skill shortages.

ASU is testing a range of models to assess the relationship between innovation and flexible working arrangements. The models include:

- a pooled model with robust standard errors
- a standard random effects model, which assumes that the firm-specific effects are orthogonal to the other covariates in the model
- a random effects model with allowance for correlation between unobserved firm heterogeneity and covariates following the approaches suggested in Mundlak (1978) and Chamberlain (1984)
- a dynamic random effects probit model that follows Wooldridge (2005) to deal with the initial conditions problem.

The preliminary results from the above models indicate that there is persistence in innovation and that flexible working arrangements have a positive and significant impact on innovation.

The tests conducted on the results of the different models indicate that the firm specific effects play an important role in the analysis, there is evidence of correlation between firm heterogeneity and covariates, and the lag effects are positive and significant.

References

Chamberlain, G. (1984) 'Panel data', in Z Griliches & M Intriligator (eds), *Handbook of Econometrics*, vol. 2, North-Holland, Amsterdam.

Mundlak, Y. (1978) 'On the pooling of time series and cross section data', *Econometrica*, vol. 46, no.1, pp. 69-85.

Wooldridge, J.M. (2005) 'Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', *Journal of Applied Econometrics*, vol. 20, pp. 39-54.

Further Information

For more information on this work-in-progress, please contact Cristian Rotaru (02 6252 5098, cristian.rotaru@abs.gov.au)

Measuring the Impact of Webforms in the Survey of Employee Earnings and Hours

The ABS is currently in the process of introducing an internet-based mode of collection (web forms) for its business and household surveys. Web forms were introduced for the May 2012 Employee Earnings and Hours (EEH) survey. EEH is conducted biennially and provides statistics on the composition and distribution of earnings of employees, the hours they are paid for and the methods used to set their pay (ie awards, collective agreements and individual arrangements).

An "opt-out" approach was taken for the introduction of the EEH web form, whereby businesses selected in the survey were initially given a link to an online web form. A paper or spreadsheet form was only sent out if a selected business subsequently requested one. This approach proved very successful, with about 90% of respondents providing data using the web form. The relatively small amount of data obtained directly from paper and spreadsheet forms limited our ability to detect a systematic impact on EEH responses due to the introduction of web forms. Nevertheless, an analysis was undertaken to provide reassurance that an obvious web form effect did not impact on EEH estimates.

The analysis comprised four parts:

1. Exploratory data analysis - Using the 2012 EEH data, a scatter plot matrix of the continuous variables of interest was produced. For each scatter plot,

the distributions of the web and non-web responses were compared.

2. Comparison of EEH responses with other data - This involved comparing EEH web and non-web responses at the employer level with corresponding data provided by the same employer in other data sources. The values of common variables of interest for these units were compared using scatter plots, to examine if the distributions for the web and non-web responses differed significantly. A more formal analysis was then conducted using linear regression analysis.
3. Modelling earnings and number of employees - Data from the May 2010 and May 2012 Average Weekly Earnings (AWE) surveys (both paper form surveys) was used to estimate how units common to both the 2010 EEH and 2012 EEH surveys would have responded if these units were provided with a paper form in 2012. The relationship between the modelled and actual 2012 EEH values for the web and non-web businesses was then compared.
4. Propensity score sub-classification - A logistic regression model was created to estimate the probability that each business in the EEH sample would respond via a web form. The sample was then grouped into five categories based on these estimated probabilities. A web form impact was estimated within each category separately, and these were then combined to form an overall estimated impact.

Our analysis provided reassuring evidence that there were no systematic web form impacts for the 2012 EEH. The opt-out approach will continue to be the main method adopted by the ABS for migrating the rest of its business surveys to web forms. For surveys moving to web forms, conducting these analyses at the same time as the editing stage should help processing staff to identify particular responses that may have been impacted by the introduction of a web form. Further editing effort may then be directed towards those responses to determine whether a change is required.

Further Information

For more information, please contact Melanie Black (02 6252 7241, melanie.black@abs.gov.au) or Lyndon Ang (02 6252 5279, lyndon.ang@abs.gov.au)

Methodology Architecture Unit

To stay a strong and relevant central statistical agency into the 21st century, the Australian Bureau of Statistics (ABS) is transforming the way it acquires, collates, uses, reuses and disseminates statistical information. To support this transformation, innovative, industrialised and contemporised statistical methods and tools will be required. Methodology Architecture (MA) provides a roadmap for systematically assessing and developing these 21st century methods and tools covering the full spectrum of the statistical production cycle. It comprises an envisaged future inventory of statistical methods and tools, covering the full spectrum of the Generic Statistical Business Process Model (GSBPM). Contrasting with the current

inventory, the MA will also provide a transition plan for migrating from the current methodological state to the future state.

The Methodology Architecture Unit has been set up within MDMD to develop the MA. As well as making a considerable contribution to corporate transformation works such as the Enterprise Architecture, assessment of Enterprise Data Warehouse proposals and the Business Assessment Team, the unit has produced a set of principles for MA and is working on a draft MA, with a high level first cut in March and a detailed version in June. Briefly, the principles are that the suite of methods should be:

- soundly based
- flexible
- efficient
- standardised
- modular & transparent in interaction
- shareable
- tested in application, outcome measured & have good holistic outcomes
- applicable by business areas
- transformable.

Further Information

For more information, please contact Bill Gross (02 6252 6302, bill.gross@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 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 added to or removed from 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