Estimation methods for business statistics variables that cannot be obtained from administrative data sources

Lewis, Daniel
UK Office for National Statistics, Methodology Directorate
Government Buildings, Cardiff Road
Newport. NP10 8XG. UK.
E-mail: daniel.lewis@ons.gov.uk

Brodie, Pete
UK Office for National Statistics, Methodology Directorate
Government Buildings, Cardiff Road
Newport. NP10 8XG. UK.
E-mail: pete.brodie@ons.gov.uk

1. Introduction

Statistics offices are increasingly in the position where they wish to, or are required to, replace surveys with administrative data. For multivariate surveys, it is often the case that some of the variables required are available directly from administrative sources, whereas others are not. This paper presents results on work to assess the applicability of estimation methods for a number of variables that are not generally available without using a dedicated survey. The work is part of the European Statistical System Network (ESSnet) Project on the uses of Administrative and Accounts Data for Business Statistics. The aim of this part of the project, known as Work Package 3 (WP3), is to provide statistical offices with the tools to be able to replace surveys by a combination of administrative data and estimated variables. WP3 is a collaboration between members of the statistical offices of the Netherlands, Italy, the UK, Lithuania and Germany.

This paper presents the results of investigating estimation methods for the first five variables considered by the project – Growth of new manufacturing goods, Change in stocks of goods and services, Purchases of goods and services for resale in the same condition as received, Payments for agency workers, and Number of employees in full time equivalent units. There will also be discussion of the remaining variables due to be analysed and of the anticipated final outcome of the project.

2. Estimating Growth of new manufacturing goods

Methods for estimating the variable Growth of new manufacturing goods (New orders) were investigated by the Netherlands. In the Netherlands, New orders data are currently collected by a monthly survey for large businesses, defined as those with more than 50 full time equivalent workers. Small and medium enterprises (SMEs) were previously also included in the survey, but have since been excluded in order to reduce burden on those businesses. In the Netherlands SMEs are defined as businesses with more than 20 and less than 50 full time equivalent workers.

The main source of related administrative data is the Turnover variable available from Value Added Tax (VAT) data, a goods and sales tax. Four methods were tested to estimate year-on-year growth rates of New orders for SMEs. The year-on-year growth is defined as the growth between the New orders for the current month and the New orders from the same month in the previous year. The methods were evaluated using survey and VAT data from 2004 to 2007, during which period the SMEs were still part of the monthly survey.

The first method is to estimate the year-on-year growth rate of New orders for SMEs using the growth rate of large businesses, which are still surveyed. The suitability of the method was tested by calculating correlation coefficients and mean absolute differences in year-on-year growth rates between large businesses and SMEs. All calculations were carried out within 2 digit industrial groupings (using the European NACE industry classification). The correlations were reasonably high for a lot of industries, but in most cases the mean absolute differences were too large to be able to use the method. The method was also tested using the growth rate of the smaller of the large businesses, but this did not improve results.

The second method is to estimate the year-on-year growth rate of New orders using the year-on-year growth rate of VAT Turnover. VAT Turnover is expected to be well correlated with New orders. However, because the order comes first and Turnover follows it is possible that there will be a delay between growth of New orders and growth of Turnover. In most industries the mean absolute differences between growth of New orders and growth of VAT Turnover are unacceptably high. However, there are some industries where the method appears to work well. Analysis showed that these are the industries that have a high contribution to total New orders.

The third method uses output from a qualitative business cycle survey. The survey includes a question on whether the value of new manufacturing orders has grown, diminished or stayed at the same level compared to the previous month. Attempts to use this qualitative information to estimate quantitative growths in New orders turned out to be unsuccessful, with very low correlations.

The final method estimates growth in New orders using modelled VAT Turnover data. The method involves estimating Manufacturing turnover by multiplying VAT Turnover by the ratio between Manufacturing turnover and Total turnover calculated from an annual business survey. The ratios are calculated within each 2 digit industry and at a delay of 1.5 years, since this is the stage at which the annual survey data are available for most businesses. The growth in this estimated Manufacturing turnover is then used to estimate growth in New orders. This was the most successful method tested. The growths from the two sources are well correlated and for all but four 2 digit industries the mean absolute differences are below 0.4%.

The conclusion of research into estimating New orders from administrative data is that the most promising method is to model VAT Turnover data using annual ratios between Total turnover and Manufacturing turnover.

3. Estimating Change in stocks of goods and services

Italy investigated methods for estimating the variable Change in stocks of goods and services (CS) and two component variables Change in stocks of finished products and work in progress (CSFP), and Change in stocks of raw materials and for resale (CSRM). The following relationship holds between these variables: CS = CSFP – CSRM.

In Italy, data on changes in stocks for SMEs (in this case, enterprises with less than 100 employees) are collected from an annual survey. Data for large businesses come from a separate census of accounts data. The focus of this study was to find methods for replacing the SME survey with administrative data.

In Italy, information on Change in stocks is available from two administrative sources. The financial statements of corporate enterprises include all three variables: CS, CSFP and CSRM. However, these financial statements only cover around 20% of businesses, accounting for around 57% of employees. The Sector studies sample survey carried out by the Italian fiscal authority includes only one variable on Change in stocks: CS. This Sector studies survey covers around 67% of businesses.

Estimation methods were considered for two scenarios, one where administrative data are available for CS but not CSFP or CSRM, and one where data are not available for any of the Change in stocks variables.

The first scenario, where administrative data are only available for CS, is treated as a missing data problem. It is assumed that the variables CSFP and CSRM are missing at random. Four methods were tested to estimate the missing CSFP and CSRM. The methods are nearest neighbour donor imputation, robust regression, mean imputation and median imputation. The mean and median imputation methods are both calculated within cells.

For the nearest neighbour imputation, the proportion of CSFP to CS is imputed from the selected donor. This proportion is multiplied by the available CS value. CSRM can then be calculated using the relationship between the variables. Imputation classes are defined by Industry, Legal form (corporate, non-corporate or sole-proprietorship) and the sign of CS (positive or negative). Number of employees (from the business register), Turnover, CS and Purchases (from the administrative source) are all used as matching variables in the nearest neighbour imputation.

The robust regression method predicts CSFP and CSRM using the relationship with available auxiliary variables. Least Trimmed Squares are used to estimate regression coefficients. Auxiliary variables tested in the model were Industry, Number of employees, CS, Turnover and Purchases.

For the second scenario, where no Change in stocks variables are available from administrative data, two methods were tested. The first method is nearest neighbour imputation, using the same donor to impute all three Change in stocks variables. The matching variables used were Industry, Legal form, Number of employees, Turnover and Purchases. Imputation classes were based on Industry, Legal form and Employment size-band. The second method tested in this scenario involves predicting CS using unit level robust regression modelling. Following prediction of CS, the components CSFP and CSRM are derived using methods from scenario 1. The regression modelling is within cells defined by Industry and Legal form, and uses Number of employees and Turnover as auxiliary variables.

The methods for both scenarios were evaluated by randomly generating non-responses in data which had all three variables present and then comparing estimates from the original data with those based on estimating missing data.

For scenario 1, the robust methods (robust regression and median imputation) worked best and both would be recommended for use when CS is available but CSFP and CSRM are missing. For scenario 2 the nearest neighbour imputation method performs well for most industries, but the regression method does not provide satisfactory results. The scenario 2 methods are less well developed than those for scenario 1 and it is recommended that they are viewed as a useful starting point for estimation methods when all Change in stocks variables are missing, rather than a final solution.

4. Estimating Purchases of goods and services for resale in the same condition as received

The UK investigated methods for estimating Purchases of goods and services for resale in the same condition as received (Purchases). Purchases are currently collected in the UK using an Annual Business Survey. The work on Purchases tested two scenarios. In the first scenario, it is assumed that the survey is stopped completely so that Purchases has to be estimated for the whole population. In the second scenario, the survey is only stopped for a subset of the population, and Purchases is estimated using the relationship between administrative data and Purchases data in the remaining part of the survey.

In the first scenario, Purchases was estimated by modelling the relationship between Purchases and VAT Turnover and Expenditure variables using linear regression. Unit level regression coefficients were estimated using historic data and then applied to current VAT data to estimate Purchases. This method only gives predictions for those businesses that were previously in the Annual Business Survey. Estimates for other businesses were calculated using trimmed mean or median imputation, within cells defined by industry and employment size-band. The results of this method were not accurate enough to be used in practice, suggesting that it would not be possible to stop the whole survey in the UK with the current available administrative data.

Three separate methods were tested for the second scenario. In each case the survey was stopped for the lowest employment size-band (band 1) and data from the next size-band up (band 2) are used to help estimate Purchases for the lowest size-band.

The first method is a simple ratio adjustment that multiplies the estimate of Purchases from band 2 by the ratio of total VAT Turnover in band 1 to a survey estimate of VAT Turnover from band 2. The second method is unit level linear regression modelling, fitting a model to predict Purchases using multiple auxiliary variables. The model is fitted using data from band 2 and then applied to businesses in band 1. VAT Turnover and various business register variables (Region, Industry, Employment and Turnover) were tested in the model. The third method is generalized calibration estimation, using a method described in Haziza et al. (2010). The method effectively performs a bias correction to the ratio estimator to take account of the fact that a subset of the population is not surveyed.

The methods in the second scenario were evaluated by comparing survey estimates of Purchases with estimates calculated using the three methods described above. The simple ratio adjustment and generalized calibration estimation methods performed best. However, the performance of the generalized calibration estimation varied considerably from year to year. The simple ratio adjustment is therefore recommended for estimating Purchases in this scenario.

5. Estimating Payments for agency workers

Lithuania investigated methods for estimating Payments for agency workers, which is currently part of the Lithuanian Structural Business Statistics survey. This variable is particularly challenging to estimate as there are no administrative sources obviously correlated with Payments for agency workers. The idea for estimating Payments for agency workers was to use information from the providers of the agency work – temporary employment agencies. The income of the temporary employment agencies could be viewed as a proxy for Payments for agency workers.

Two methods were initially considered for estimating the income of temporary employment agencies. The first method is to estimate the income from profit and loss accounts. For this to give an accurate estimate of Payments for agency workers, it would be necessary to remove the non-domestic part of the income. In Lithuania the profit and loss accounts do not provide a split between domestic and non-domestic income, so this method was not pursued.

The second method is to use information from employment agencies' answers to the annual services survey, which does include separate questions on domestic and non-domestic income. Whilst this method involves using survey data, it could still lead to efficiency savings if the services survey was retained and the structural business survey stopped. This method was tested by comparing estimates of domestic income for temporary employment agencies from the services survey, with estimates of Payments for agency workers from the Structural Business Statistics survey. Data from 2008 and 2009 were used.

The results showed that estimates of Payments for agency workers from the Structural Business Statistics survey are on average around 30% higher than those derived from the domestic income of temporary employment agencies. This difference can partly be explained by sampling variation and measurement error. However, the main difference in estimates may be due the fact that businesses can employ agency workers from foreign employment agencies that are not included in the services survey.

Payments for agency workers is a difficult variable to estimate without using a dedicated survey, as no administrative data are correlated. The method tested to estimate Payments for agency workers from the domestic income of temporary employment agencies did not prove entirely successful. However, if a method could be found for estimating payments to foreign employment agencies, this approach may be useful.

6. Estimating Number of employees in full time equivalent units

Germany investigated the estimation of Number of employees in full time equivalent units (FTE). FTE is estimated in Germany using a Structural Business Statistics survey. There is no administrative source containing the FTE variable, but a number of related variables are available. Part time and Full time employees are both available from administrative data, but only for the financial services sector. Hours worked is available from a range of sources: the Labour Cost Survey (LCS), the Labour

Force Survey (LFS) and data from the Institute for Employment Research (IAB). Hours paid is available from the Quarterly Earning Survey (QES). None of these sources has the detail required to estimate FTE on its own, but the intention is to use a combination of the sources to produce estimates.

The method tested to estimate FTE involves using Full time employees and Part time employees data from the administrative source. Each Full time employee is equivalent to 1 FTE. However, Part time employees need to be converted into FTE. This conversion can be achieved by multiplying Part time employees by the ratio of Hours worked (or Hours paid) for Part time employees to Hours worked (or Hours paid) for Full time employees. Hours worked and Hours paid can be estimated from a combination of LCS, LFS, IAB and QES data. A refinement to this method is to separate out different types of Part time employees and calculate different conversion factors for each. Using German data it improves results when ‘Mini-jobbers’ and Trainees are estimated separately to other Part time employees.

This method was evaluated by comparing the estimates of FTE from the Structural Business Statistics survey with estimates from this new method, using different combinations of data sources. The average difference between survey estimates of FTE and the best performing estimates with the new method is 1.5%. This suggests that the estimates are of sufficient quality to not need the structural business statistics survey in the financial services sector.

7. Plans for developing estimation methods for additional variables

The work described in this paper covers estimation methods for variables studied in the first 18 months of this project. There are two years remaining, during which WP3 will investigate estimation for at least six more variables that are not available from administrative sources. The first three of these will be Payments to sub-contractors, Sales of tangible investment goods, and Gross investment in tangible goods. The aim at the end of the project will be to produce a guide describing the type of estimation methods that can be used for business statistics variables that statistics offices might wish to stop surveying. As well as identifying estimation methods for those variables that are not available from administrative data, it is hoped this will include a description of variables that are available from administrative sources. In some cases it is not obvious to statisticians how variables can be derived from accounting data. A separate work package of the ESSnet will be investigating this issue and some of their results may be summarised in the final output of WP3.

8. Conclusion

This paper has presented findings from the investigation of estimation methods for five survey variables that are not usually available from administrative or accounting sources. The results of testing these methods have shown whether it is possible to estimate these variables without a dedicated survey. For all five variables, useful estimation techniques have been identified. However, in each case the availability of related data is crucial to being able to stop collecting the variable by survey. In some cases it may only be possible to stop the survey for a subset of the population, if it is desired to maintain the quality of results. WP3 will investigate estimation methods for more variables and will ultimately produce an estimation guide for business statistics variables.

REFERENCES

Haziza D., Chauvet G. and Deville J-C. (2010), “Sampling and Estimation in Cut-Off Sampling”, *Australian & New Zealand Journal of Statistics*, 52, 303-319

Kavaliauskiene D. (2011), “Development of estimation methods for business statistics variables which cannot be obtained from administrative sources. Variable: Payments for agency workers”, report for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics.

Luzi O., Seri G., De Giorgi V. and Siesto G. (2011), “Development of estimation methods for business statistics variables which cannot be obtained from administrative sources. Variable: Change in stocks of goods and services”, report for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics.

Redling B. (2011), “Development of estimation methods for business statistics variables which cannot be obtained from administrative sources. Variable: Number of employees in full time equivalent units”, report for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics.

Sanderson R., Elliott D., Lewis D. and Jones T. (2011), “Development of estimation methods for business statistics variables which cannot be obtained from administrative sources. Variable: Purchases of goods and services for resale in the same condition as received”, report for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics.

van der Holst R. (2011), “Development of estimation methods for business statistics variables which cannot be obtained from administrative sources. Variable: Growth of new manufacturing goods”, report for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics.