Imputation Classes by Size Measure for the Annual Survey of Manufactures of Statistics Canada

Yi Li
Statistics Canada, yi.li@statcan.ca, 150 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6, Canada

Abstract

Ratio type imputation methods are used extensively for the imputation of Statistics Canada's Annual Survey of Manufactures (ASM). For example, one method consists of deriving the ratio between the variable requiring imputation and an auxiliary variable using a group of eligible records within an imputation class. This ratio is then applied to the record requiring imputation to generate an imputed value. In order to improve the quality of imputed data, in addition to industry classification and geographical location, a size measure variable was introduced in constructing the imputation classes so that units of different sizes can be properly represented. In this paper, we discuss the criteria used to define the size measure and the strategy used to determine the size groups. We also discuss how the imputation classes can be redefined when a larger pool of units is necessary for the calculation of the ratios.

KEY WORDS: Imputation Class, Size Measure, Ratio Imputation

1. Introduction

Imputation is often used to fill in the missing values in case of item non-response. Ratio imputation, a special case of regression imputation, is widely applied in business surveys since the economic variables are often strongly correlated to each other. However, as business data are typically highly skewed, the quality of ratio imputation is vulnerable to misspecification of the underlying model. In order to deal with this situation, units are often divided into homogenous groups called imputation classes such that different models can be built in different classes independently. In this paper, using the Annual Survey of Manufactures (ASM) of Statistics Canada as an example, we will discuss how a size measure variable can be used in constructing imputation classes such that the units of different sizes can be properly represented.

We will first have an overview of the ASM in section 2, which includes an introduction of the automated edit and imputation system. In section 3, we will discuss how the size variable was chosen. And then in section 4, we will briefly describe the methodology that we used to determine the size groups and boundaries. Finally in section 5 we will present the conclusion.

2. Overview of the Annual Survey of Manufactures

2.1 A Brief Description of the Sampling Design

The Annual Survey of Manufactures, as part of the Unified Enterprise Survey (UES) program of Statistics Canada, annually collects information about the manufacturing sector and the logging industry in Canada. The collected information includes principal industrial statistics such as revenue, employment, cost of materials and supplies used, cost of energy and water utility, inventories, and destination of shipments as well as data about the commodities produced and consumed. The ASM consists of 23 surveys. The target population of the ASM comprises all establishments primarily engaged in manufacturing and logging activities. Under the North American Industry Classification System (NAICS), logging establishments are classified to NAICS 1133 and manufacturing establishments to NAICS sectors 31, 32 and 33.

To reduce respondent burden and to save on survey costs, the ASM target population is divided into two principal parts using predetermined exclusion thresholds: the survey population portion, from which units are sampled and data are collected via questionnaires; and the take-none portion, from which no units are selected to receive questionnaires. The survey population is stratified into four strata: the Must-Take, the Take-All, the Take-Some1 and the Take-Some2. The units in the Must-Take stratum are pre-specified by the subject matter analysts based on the complexity and size of the units. The boundaries for the Take-All, the Take-Some1 and the Take-Some2 strata are derived by using a variable that is the maximum between the Gross Business Income (GBI) and the Shipment. All the units in the Must-Take and the Take-All strata are selected for the survey, while only a fraction of the units in the two Take-Some strata are selected. The sampling fraction for the Take-Some2 stratum is higher than that for the Take-Some1. The reader may consult Lebrasseur and Turmelle (2007) for more details on the UES sampling.

2.2 The Imputation Strategy and the Automated Imputation System

All selected establishments receive questionnaires. After data collection and some preliminary editing, the ASM data are processed by BANFF, an automated edit and imputation (E&I) system developed at Statistics Canada.

The ASM has 90 variables for processing during edit and imputation. The automated edit and imputation system is driven by metadata. It uses a number of different imputation methods, including direct replacement with administrative data (tax data) or annualized Monthly Survey of Manufactures (MSM) data, historical imputation, current year ratio imputation and donor imputation. The 7 key total variables are C2098 (total revenue), C4699 (total costs/expenses), C2302 (total operational revenue), C2303 (total other revenue), C4019 (total purchase), C5550 (total opening inventories) and C5555 (total closing inventories). They may be mapped directly to the tax data file, hence imputation by direct tax replacement is possible. Tax data are not available for the 83 detailed variables. The general imputation strategy is to first impute the 7 key variables, and then the detailed variables, using the totals as anchors. Reader may consult Provençal, Chepita, Li, Yeung (2007) for additional information on the impact of using tax data for imputation.

Among the different imputation methods, the current year ratio imputation is the one used most extensively in the ASM, in particular for the detailed variables since tax data and annualized MSM data are not available for these variables and many units do not have reported historical data. The current year ratio method consists of deriving the ratio between the variable requiring imputation (*Y)* and an auxiliary variable (*X)* using a group of eligible records within an imputation class. This ratio is then applied to the record requiring imputation to generate an imputed value. This method is called CURRATIO (CURrent RATIO) in BANFF. To avoid estimated ratios that are too unstable, the ASM requires at least 5 eligible units in an imputation class. The units with missing values for one or both variables are excluded from the calculation of ratios. Outliers are also excluded. The classes with insufficient eligible units are collapsed in order to expand the pool of eligible units.

In survey practice, imputation classes are often defined by the stratification variables and/or the variables for publication domains. In the ASM, the original imputation classes that were used for CURRATIO method were defined by industrial classification and geographic location. It has been observed in ASM and other business surveys that large units have different characteristics from the smaller ones. Therefore the use of the ratios based on units within imputation classes that contain a mixture of large and small units for imputation may result in "unsuitable" estimates. In order to improve the quality of imputed data, in addition to the industrial classification and geographical location, a size measure variable was introduced in constructing the imputation classes so that units of different sizes can be properly represented.

3. Choice of size variable

The size measure variable to be used must be available for every unit including both the respondents and the non-respondents. This variable must also be reliable in distinguishing between large and small units. On the sampling frame we investigated three variables for possible consideration. These variables are the maximum between the Gross Business Income (GBI) and the Shipment, the sampling stratification variable, STRATE and the sampling design (adjusted) weights, WUES_C. On the auxiliary file, the variable Total Revenue (C2098AUX) was considered.

The variable STRATE was created primarily for sampling design purposes. It has four values corresponding to the 4 strata described in section 2.1. As a size variable, STRATE was found to be unsuitable since it could not reliably distinguish between large and small-size businesses in certain cases. For example, small-sized must-take units may be considered as take-all (usually large-sized businesses). The sampling design (adjusted) weights, WUES_C, is essentially similar to the sampling stratification variable, STRATE, in the sense that small-sized must-take units may be combined with large-sized take-all units, both having WUES_C=1. Hence, WUES_C was considered not to be a reliable measure of size.

Unlike the maximum between the Gross Business Income (GBI) and the Shipment, which is used to derive the boundaries of the sampling strata, the total revenue from the auxiliary file (C2098AUX) is created after data collection is completed and therefore is considered to be more up-to-date information. This variable is mainly based on tax and historical data. After a series of analyses, it was determined that the C2098AUX is the most appropriate size-measure variable.

4. The Methodology for Determining Size Groups and Boundaries

We employed the following five steps to determine the appropriate size boundaries.

Step 1: SAS Procedure FASTCLUS:
The FASTCLUS procedure performs a disjoint cluster analysis on the basis of distances computed from one or more quantitative variables (here C2098AUX). The observations are divided into clusters such that every observation belongs to one and only one homogeneous cluster in the sense that the distance between any given observation and its cluster center is less than the distance between this given observation and the center of any other

cluster. By default, the FASTCLUS procedure uses Euclidean distances and the cluster centers are the means of the observations assigned to each cluster.

In our analysis, we first used the FASTCLUS procedure at the NAICS6 level to obtain the initial clusters, i.e. size groups and their boundaries. In cases where one or more NAICS6 had too few units, they were merged into a new NAICS6 group. The FASTCLUS procedure was then applied to the new group.

Step 2: Distributions of C2098AUX
The second step was to study the distributions of C2098AUX. Additional information from the distributions was used to modify the clusters suggested by FASTCLUS. We merged some groups where it was deemed appropriate, especially when groups contained too few units; we also modified some boundaries suggested by FASTCLUS procedure to make grouping easier to implement.

Step 3: Generalized Linear Models (GLM)
In this step, we used GLM to validate the proposed size groups derived from steps 1 and 2 above. The basic model was as follows:

Variable-for-which-imputation-may-be-required vs.
Size, NAICS, GEO and their interactions

where the variable-for-which-imputation-may-be-required included the 7 key totals and some selected detailed variables. The independent variable that explained most of the total sum of squares variation (say SIZE) was considered to be the most important classification variable, followed by the next most important and so on. This also provided an indication of which of the three variables could be collapsed when necessary. For most of the analyses done, GEO (provinces) was found to be the first variable that could be collapsed into regions and then into Canada, followed by NAICS. In most cases, the size group did not need to be collapsed.

Step 4: Regression model on new imputation classes
As an additional check on the importance of the imputation classification variables, we built regression models for the new classes with the size measure variable, i.e. NAICS*GEO*SIZE, and compared the results to those from the models based on the original classes, i.e. NAICS*GEO. For example, for a given industry, we compared the models based on following imputation classes:
NAICS6*REGION*SIZE vs NAICS6*REGION
NAICS6*CANADA*SIZE vs NAICS6*CANADA
NAICS3*CANADA*SIZE vs NAICS3*CANADA

We say that size matters if:

- At least two of the models for the classes with or without size have the form of $y = b*x$ (i.e. a=0 and $b\neq 0$); and
- Their slopes (b) are significantly different.

We say size does not matter if:

- None of the models for the classes with size have the desired model; or
- The slopes (b) are not significantly different for the classes with the desired models.

We say we cannot draw any conclusions regarding the importance of size if:

- All the classes with size have less than five eligible units; or
- All the units are in one size group.

The results showed that among the cases where size is applicable, size matters for 80% and 62% of the cases for the first and second runs of CURRATIO at levels of NAICS6*REGION and NAICS6*CANADA respectively.

Step 5: Test the proposed imputation classes in the automated E&I System
This involved the following steps (The ASM data for reference year 2004 were used):

- Integrate the new imputation classes into the automated E&I system by updating corresponding metadata.
- Run E&I process from Importing Metadata files to Exporting completed E&I processed data based on the proposed imputation classes.
- Results from the test runs were compared to those from production. Significant differences were observed between the two sets of data based on the original and the new imputation classes.
- The two sets of data were provided to the subject matter analysts for further analyses. There was a general acknowledgement of significant improvements in the quality of the imputed data using the new imputation classes.

5. Conclusions

Size measure is an important classification variable in defining imputation classes. Although the proposed new imputation classes by size measure are far from perfect partly due to the many implementation restraints, they have led to significant improvements to the data quality. The new imputation classes were implemented in the ASM production.

Acknowledgements

The author would like to acknowledge Emmanuel Benhin, a senior methodologist in the Business Survey Methods Division, for his significant contributions to this project. The author is grateful to the ASM production team, in particular the analysts from the Manufacturing, Construction and Energy Division, and the colleagues from the Enterprise Statistics Division for their input and collaboration. The author would also like to thank Sylvie Deblois and Jean-François Beaumont for their useful comments.

References

BANFF Support Team (2005), Functional Description of the BANFF System for Edit and Imputation, Internal Document, BSMD, Statistics Canada

Kover, J.G., Whitridge, P.J. (1995), Imputation of Business Survey Data, Business Survey Methods, edited by Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M. J. and Kott, P.S., John Wiley and Sons.

Lebrasseur, D. and Turmelle, C. (2007), Toward a Better Integration of Survey and Tax Data in the Unified Enterprise Survey, Proceedings of ICES-III, Survey Methods for Businesses, Farms, and Institutions, Canada (Montréal)

Provençal, J.-S. (2005), An update of Analysis of Data Slices and Metadata to Improve Survey Processing, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Canada (Ottawa). (http://www.unece.org/stats/documents/2005.05.sde.htm).

Provençal J.-S, Chepita R., Li Y., Yeung C. W. (2007), Impact of Using Fiscal Data on the Imputation of the Unified Enterprise Survey of Statistics Canada, Proceedings of ICES-III, Survey Methods for Businesses, Farms, and Institutions, Canada (Montréal)