The Best Stratification to Impute Missing Values of Turnover in Economic Surveys

Takayuki Ito[1,2], Yutaka Abe[1], and Tatsuo Noro[1]
[1] National Statistics Center, Tokyo, JAPAN
[2] Corresponding author: Takayuki Ito, e-mail: titou@nstac.go.jp

Abstract

Along with the trends in the Japanese official statistics, the Economic Census for Business Activity was conducted February 2012, for the first time in Japanese history. This survey aims to cover the enterprises and establishments in all of the Japanese industrial fields, in order to clarify the actual economic conditions across the nation as well as in each region; furthermore, its goal is to obtain population information on businesses. However, missing values and errors are frequently produced in such accounting items as turnover. The Economic Census will be used as basic economic data, so that it is important to get hold of information about all enterprises and establishments; thus, missing values must be imputed in one way or another. Nevertheless, the problem is how to impute missing values. One aspect of this problem is to evaluate the accuracy of imputation based on stratified data. In each stratum, the evaluation method is to compare the difference between the true values and the imputation of the missing values. By way of comparing these differences across several strata, our goal is to find the stratum with the smallest difference between the true values and the imputation of the missing values. Therefore, we experimented with two-industry datasets and found the best stratification in these industries. If the method is useful for others, it will be used for the Economic Census.

Key Words: Economic Census, evaluation method

1. Introduction

Missing values and errors are frequently produced in the Economic surveys. In order to impute missing values and errors, it is important to study the imputation method. One of the aspects of this problem is to evaluate the accuracy of imputation based on stratified data. Here, the focus is how to stratify data. The goal is to find the most accurate stratum in the imputation of the missing values. The point of the evaluation method is to compare the difference between the true values and the imputation of the missing values. The most accurate stratum is defined as the smallest difference across several strata.

There are four strata for imputation of the missing values. These strata are ‘Major group of industries’, ‘Intermediate group of industries’, ‘Number of worker as well as Major group of industries’ and ‘Number of worker as well as Intermediate group of industries’.

First, the largest unit of four strata is ‘Major group of industries’, which is regarded as ‘Basic stratum’. Second, ‘Intermediate group of industries’ is a part of the major group of the same industries, so this stratum is below only one stratum compared with ‘Basic stratum’. Third, the worker is an important variable to impute the missing values. Actually the variable used to impute the missing values is the worker. The way this variable will be stratified is ‘Number of worker as well as Basic stratum’. This stratum is also below one stratum compared with ‘Basic stratum’. Last, ‘Number of worker as well as Intermediate group of industries’ is the smallest stratum of four strata. This stratum is below two strata compared with ‘Basic stratum’.

By way of comparing these differences across four strata, we will find the stratum of the smallest difference between the true values and the imputation of the missing

values. In this imputation, the model we used is 'ratio imputation', and our data is 2004 Survey on Service Industries for Japan instead of the Economic Census data.

2. Model for Imputation method

The Model to find the best stratum of four strata is ratio imputation. This model is equation (1).

$$\tilde{y}_i = \hat{\mathrm{R}}\mathrm{x}_\mathrm{i} \qquad (1)$$

$$\hat{\mathrm{R}} = \frac{\sum_{\kappa=1}^{n} obsY_\kappa}{\sum_{\kappa=1}^{n} obsX_\kappa}$$

Here, i is all data, k is observed data, $obsX_k$ is observed worker units, $obsY_k$ is observed Turnover units, $\hat{\mathrm{R}}$ is ratio. $\tilde{y}_i$ is complete values after imputation. x_i is complete values of worker. This model shows that the total of the observed values is completely the same value as the total of fitted ones in observed data. This is the special feature of this model.

3. Experiment

In order to consider several missing rates, we will make five missing rates; 10%, 20%, 30%, 40% and 50%. In one industry, the total number of the experiment's times is 17 (5 times with each of 10% and 20%, 3 times with 30%, and 2 times with each of 40% and 50%), and the method of the missing data generation is systematic sampling. For example the sample in case of 50% is presented Table 3.1 and Missing pattern[1] is 'First Time Missing' and 'Second Time Missing'. (Ascending ordered only 10 records) The unit of Turnover is ten thousand yen in Japan.

Table 3.1: Image of missing data generation , systematic sampling

Data No.	Major group of industries	Intermediate group of industries	Number of worker	worker	Turnover	Missing Pattern[1]
1	Eating and drinking places, accommodations	Eating and drinking places	1-4	1	2000	First Time Missing
2			1-4	3	4000	Second Time Missing
3			5-9	7	7500	First Time Missing
4			5-9	9	12000	Second Time Missing
5			10 and over	12	24000	First Time Missing
6		Accommodations	1-4	2	3000	Second Time Missing
7			5-9	8	20000	First Time Missing
8			5-9	9	30000	Second Time Missing
9			10 and over	15	55000	First Time Missing
10			10 and over	31	75000	Second Time Missing

[1] About Missing Pattern, 'First Time Missing' means Data No.1, 3, 5, 7 and 9 are missing data, then Data No.2, 4, 6, 8 and 10 are observed data. While 'Second Time Missing' means Data No.2, 4, 6, 8 and 10 are missing data, then Data No.1, 3, 5, 7 and 9 are observed data.

About 'Major group of industries', we experiment with two major groups of industries. One is 'Eating and drinking places, accommodations' and the other is 'Medical, health care and welfare'. Each major group of industries consists of two intermediate groups of industries which is consisted of the number of worker as Table 3.2. In the actual economic survey, the missing data surely exist, but we removed these missing data for the purpose of experiment. Then, we will use the rest of the complete data, which we call the truth (or the true value) .

Table 3.2: Data for experiment

Major group of industries	Intermediate group of industries	Number of worker	Number of Obs.
Eating and drinking places, accommodations	Eating and drinking places	1-4 persons	18956
		5-9 persons	3911
		10 and over	725
	Accommodations	1-4 persons	4183
		5-9 persons	1079
		10 and over	239
Medical, health care and welfare	Medical and other health services	1-4 persons	1416
		5 and over	127
	Social insurance and social welfare	1-4 persons	66
		5 and over	119

Table 3.3 presents the summary statistics of the two Major groups of industries.

Table 3.3: Summary Statistics (raw data)

Major group of industries	Number of Obs.	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
Eating and drinking places, accommodations	29093	1	400	807	1227	1500	12600
Medical, health care and welfare	1728	3	200	500	932	1044	28830

4. Evaluation method

The evaluation method is the difference between the true values and the imputation of the missing values; specifically, an average rate of deviation is one indicator of the difference. So in this experiment, the method we used as an indicator is average rate of deviation, which is equation (2).

$$\text{Average rate of deviation} = \sum_{i=1}^{n} \frac{\left|Truth - Imputation_i\right|}{Truth \times n} \times 100[\%] \qquad (2)$$

The value of *n* depends on the missing rate, for example if that rate is 10% or 20%, *n* is 5, if that rate is 30%, *n* is 3, if that rate is 40% or 50%, *n* is 2. *Truth* is the true values and *Imputation* is complete values after imputation. The smaller value is desirable about this average. One way of the expressing outputs in this experiment is to count the best and worst strata of four compared strata in each missing rate.

5. Reference Stratum for Evaluation

There is one problem to find the best strata. Which strata should we compare? There are four choices about the reference stratum for the evaluation; 'Major group of industries', 'Intermediate group of industries', 'Number of worker as well as Basic stratum' and 'Number of worker as well as Intermediate group of industries'. Table 5.1 presents the sample values (Complete values after imputation) described in both compared stratum and reference stratum for the evaluation.

Table 5.1: Image of the reference stratum for the evaluation

Compared stratum	Reference Stratum for Evaluation											
	Intermediate group of industries								Major group of industries			
	Eating and drinking places				Accommodations							
	1-4 persons	5-9 persons	10 and over	Whole	1-4 persons	5-9 persons	10 and over	Whole	1-4 persons	5-9 persons	10 and over	Whole
Basic stratum	500	1500	5000	7000	300	5000	7700	13000	800	6500	12700	20000
Intermediate group of industries	3000	3500	5500	12000	1000	3000	4300	8300	4000	6500	9800	20300
Number of worker as well as Basic stratum	1300	4000	9200	14500	400	1700	4000	6100	1700	5700	13200	20600
Number of worker as well as Intermediate group of industries	1250	3100	8200	12550	550	2700	5100	8350	1800	5800	13300	20900
True values	1000	3000	8000	12000	500	2500	5000	8000	1500	5500	13000	20000

If the reference stratum for the evaluation is the whole of Major group of industries, 'Basic stratum' (of Compared stratum) is the nearest value (20000) compared with the True value (20000).

If the reference stratum for the evaluation is the whole of Intermediate group of industries, 'Intermediate group of industries' is the nearest values (12000 in case of 'Eating and drinking places' and 8300 in case of 'Accommodations') compared with the True value (12000 in case of 'Eating and drinking places' and 8000 in case of 'Accommodations'). In this case, 'Intermediate group of industries' is the best stratum from the point of view of each intermediate of industries.

If the reference stratum for the evaluation is 'Number of worker as well as Major group of industries', that stratum is the best stratum. Because in each Number of worker (1-4 persons, 5-9 persons, 10 and over), 'Number of worker as well as Major group of industries' is the nearest values (in 1-4 persons, the value is 1700 (the True value is 1500), in 5-9 persons, the value is 5700 (the True value is 5500), in 10 and over, the value is 13200 (the True value is 13000)). Similarly, if the reference stratum for the evaluation is 'Number of worker as well as Intermediate group of industries', that stratum is the best stratum.

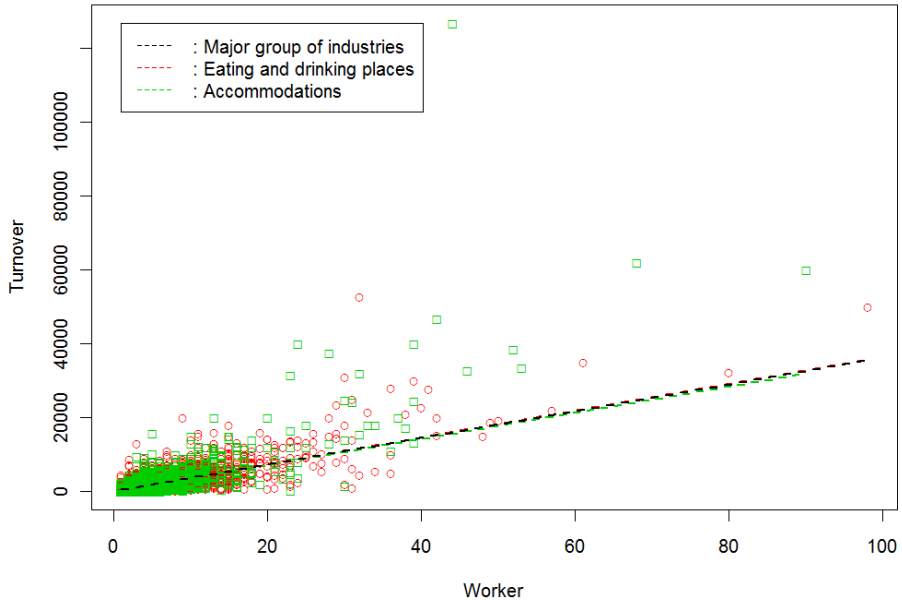
Therefore according to what is the reference stratum for the evaluation, the best will be determined. In this experiment, it is important to check the results of all strata(these four strata).

6. Results

Several results are clarified through experiment. The result is a little difference between two major groups of industries. But judging from the evaluating mixed major group of industries is leading one result among 4 reference stratum for the evaluation. Details of this result will be reported during the presentation.

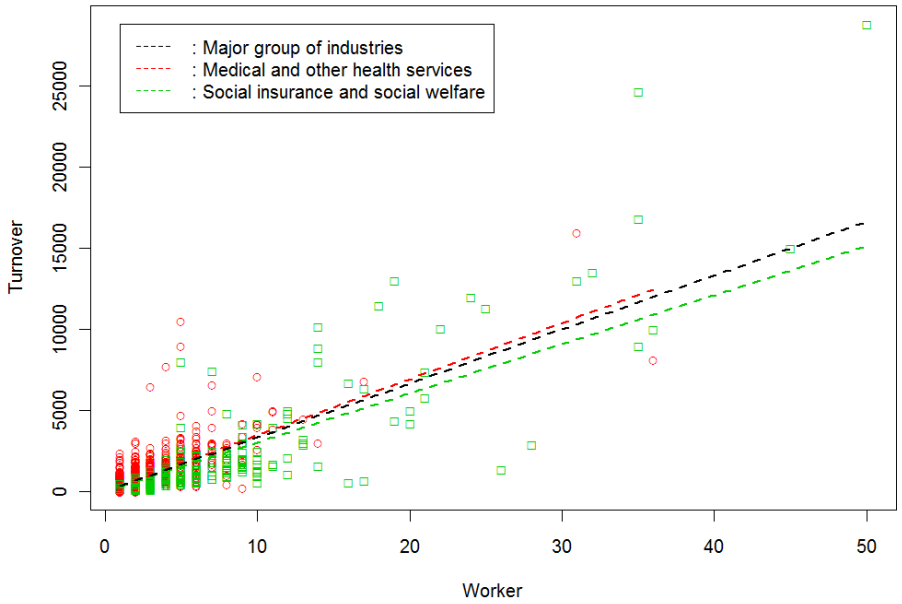
But here is the scatterplot between Turnover and Worker as a part of the results. Figures 6.1 and 6.2 present the difference between major group of industries and intermediate group of industries. (Line is the slope of ratio imputation.)

Figure 6.1: The scatterplot, Eating and drinking place, accommodations



All the slopes of the lines are close to each other in Figure 6.1.

Figure 6.2: The scatterplot, Medical, health care and welfare



Among the lines, there is a little difference in Figure 6.2.

Figures 6.1 and 6.2 present the complete data, but we will generate the missing data from this complete data. Therefore, the results after imputation of missing values can be quite another matter.

7. Conclusions

This paper described several strata for imputation of the missing values and compared their accuracy using a real dataset called 2004 Survey on Service Industries for Japan. Details of these comparisons will be reported during the presentation. The result of this experiment shows one way about the imputation of the missing values. But only two industries are used in this experiment; therefore, we must try others for the Economic Census.

But, the result of this experiment is encouraging; thus, we are looking forward to the next step for the implementation of the Economic Census.

References

1. Aoki, Shigenobu. (2009). *R niyoru Toukei Kaiseki* (*Statistical Analysis about R).* Tokyo: Ohmsha, Ltd.
2. De Waal, Ton, Jeroen Pannekoek and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation.* New Jersey: John Wiley & Sons.
3. Hashimoto, Noriko, Michiko Watanabe and Naoko Sakurai. (2009). *Excel de Hajimeru Keizai Toukei Data no Bunseki Kaiteiban (Analysis of Economic Statistical Data with Excel, A revised edition).* Tokyo: Zaidan Houjin Nihon Toukei Kyokai.
4. Mingzhe Jin. (2007). *R niyoru Data Science(Data Science about R)*. Tokyo: Morikita Publishing Co., Ltd.
5. Statistics Bureau, Ministry of Internal Affairs and Communications, Japan. (2006). *2004 Survey on Service Industries Volume 1 Results for Japan*. Tokyo: Statistics Bureau, Ministry of Internal Affairs and Communications, Japan.