Imputing zeros in business survey items using a binary classification method

Ichiro Murata
National Statistics Center, Japan

Abstract

Missing values are often problematic in official statistics, especially in terms of data processing. National Statistics Center of Japan has treated them in many cases and as regards to Unincorporated Enterprise Survey which is conducted annually, some imputation methods for missing values have been taken before calculating statistics. Through the analysis of recent data of the survey, the characteristic that many zero values had occurred at specific survey items such as purchases, salaries, and inventories was observed and it seems to be difficult to impute adequate values by the current data processing of the survey. This leads to the idea of imputing zero values by any other methods before the current data processing goes, which can be regarded as a two-step imputation. Imputing a zero value of certain data could be done if the value is predicted to be zero at a high possibility, so binary classification method that determines whether it is zero or not is appropriate for the imputation. We took logistic regression model for it and built the model by using several explanatory variables that seem to be good for reasoning zero values in accounting items of enterprises. To examine the effectiveness of imputing zero values as a first step, we compared the accuracy of the two-step imputation to the previous one by conducting a simulation using the microdata of Unincorporated Enterprise Survey 2019-2021. The result showed that imputing zeros in advance decreased the gap between a collected value and an imputed value, especially in the imputation of salary and inventory. Therefore, it could be a good option for imputation under the situation of the data having many zero values.

Keywords: Official statistics, Logistic regression, Data editing, Imputation

1 Introduction

Survey non-response is always problematic in official statistics. It causes missing values in survey data, and missing values can also come from the editing of data when the value is regarded as invalid. Those missing values hinder calculating statistics, and a major possible solution is to impute the values beforehand.

In Unincorporated Enterprise Survey in Japan which collects accounting items such as purchases, salaries, inventories, some imputation methods have been taken since 2019 and National Statistics Center (NSTAC) has reviewed the results to assure proper data processing and for further improvement. Three methods are being applied to the survey: Last Observation Carried Forward (LOCF), classified group mean value, and hot deck imputation. LOCF is used for imputing sales, which seems working well. Classified group mean value used for inventories and hot deck imputation used for both purchases and salaries are also working but they have a drawback with imputing specific values such as zero.

Through the observation of the survey data, we found there are many unincorporated enterprises mainly related to the service industry that don't have a goods stock or employees thus the corresponding accounting values, namely purchases, inventories, and salaries, are zero. Considering this situation, specifying whether a missing value is zero or not in advance would be helpful for appropriate imputation, hereafter regarded as binary classification method that an accounting value is zero or not.

As regards to hot deck imputation in official statistics, CANCEIS (Statistics Canada, 2020) provides the implementation of donor imputation with various distance calculations, and Farnell and Darby (2020) proposed utilizing administrative data to improve donor imputation. However, the application of zero value specification combined with hot deck imputation has not been seen before. This paper aims to share the case of this exploratory work.

2 Methodology

2.1 Zero-specifying model

Some accounting items collected in Unincorporated Enterprise Survey tend to have a lot of zero values. Table 1 shows the proportions of zero values in each item. Purchases, inventories, and salaries often have zero values, while sales seldom have zero. So our interest is on the former items and if the missing values could be inferred better based on this information.

Table 1: Proportions of zero values in Unincorporated Enterprise Survey 2019

Sales of the past survey	Industrial Classification	Sales	Purchases	Initial Inventories	Final Inventories	Salaries
Under 90 percentile	Construction	0.3%	8.9%	55.6%	56.6%	66.7%
	Manufacturing	0.1%	21.0%	48.1%	48.8%	66.0%
	Wholesale and retail trade	0.0%	2.0%	13.4%	13.7%	70.6%
	Accomodations and eating & drinking servide	0.0%	0.9%	31.5%	32.6%	52.4%
	Living-related & personal services and amusement services	0.0%	11.9%	36.5%	37.4%	79.3%
	Services except above	0.1%	62.5%	80.2%	79.5%	74.8%
Equal or over 90 percentile	Construction	0.0%	3.7%	43.8%	44.8%	37.3%
	Manufacturing	0.0%	9.6%	30.4%	31.4%	20.2%
	Wholesale and retail trade	0.0%	0.3%	6.8%	6.9%	12.8%
	Accomodations and eating & drinking servide	0.0%	0.4%	10.4%	11.3%	4.9%
	Living-related & personal services and amusement services	0.0%	6.7%	16.9%	18.1%	18.9%
	Services except above	0.0%	61.5%	73.0%	73.4%	22.9%

To specify whether the value is zero or not, we considered logistic regression, which is used to obtain outputs between 0 and 1, that are usually thought of as a probability. It is essential for practical data processing to select good explanatory variables, that are mainly taken from the information collected in the same survey. While some survey items related to accounting are potential candidates, they are not applicable because we observed the accounting items are almost missing in most cases of the data which need imputation. This means if one item, purchases for example, is missing then the other items such as inventories, salaries, and several expenses are also missing at high possibility. So we didn't take those accounting items into account as an explanatory variable except for sales, which has a high respondent rate and is ensured availability by imputing another way beforehand. The other survey items are considered to be included in a model specifying zero. We examined the results of forward stepwise variable selection, as well as the direct comparison of coefficients estimated by logistic regression. Provided most of the survey items are categorical, relativity evaluation based on Exploratory Data Analysis (Mutoh and Shirakawa, 2023) was also useful. Through the consideration, we selected the items below as explanatory variables. Note that the items of continuous values were changed into the levels of discrete groups because of the difficulty of fitting a linear model. (In fact, those groups are the same as those used in disseminating the statistics tables of Unincorporated Enterprise Survey.)

- Industrial Classification
- Range of sales (17 groups)
- Range of persons engaged (3 groups)
- Prefectures
- Rate of commissions received in sales (4 groups)

These variables are selected to specify zero values in purchases, inventories, and salaries. While the best variable combination would vary with each objective item, we used the same variables for simple implementation. Among the 5 variables, the "Rate of commissions received in sales" may not be directly comprehensible. The data of Unincorporated Enterprise Survey shows that the enterprises making all sales from commissioned work tend to have no inventories, especially in the manufacturing industry. That is the reason for the variable being suggested at the variable consideration.

Another part of making a model is to control a threshold. The outputs of logistic regression are between 0 and 1 so a threshold is necessary to enable binary classification in a deterministic way. We set a fixed value (0.3) as a threshold for this research because, at that value, precisions of predicting zeros for all objective items (purchases, inventories, and salaries) were around 80%, which is the practical requirement level we assume. All the conditions above are used in the rest of this paper.

2.2 Combining to imputation process

In the current data processing of Unincorporated Enterprise Survey, hot deck imputation is used for purchases and salaries while classified group mean value is used for imputing inventories. These are not adequate when a zero value should be filled in. Thus it's natural to treat zero values before it, as described below.

1. Specify whether it is zero or not for all the missing values. If specified as zero then zero will be filled in.
2. Impute the rest of the missing values using hot deck method or mean value.

This sequence can be regarded as a two-step imputation, where zero imputation is the first step and imputation for the rest is the second step.

3 Evaluation and Discussion

We conducted a simulation using the microdata of Unincorporated Enterprise Survey 2019-2021. Only the data which consists of collected values for all relevant items (i.e. complete data) are used in the simulation. Each item of individual data was assigned a value through the imputation process. The distances between collected values and corresponding imputed values are evaluated as the accuracy of imputation by two kinds of indicator, Normalized Root Mean Squared Error (NRMSE) and Mean Absolute Error (MAE) defined as,

$$NRMSE = \frac{1}{\sigma_{clct}} \sqrt{\frac{\sum (x_{clct} - x_{imp})^2}{n}}, MAE = \frac{\sum |x_{clct} - x_{imp}|}{n}$$

where x_{clct}, x_{imp} are a collected value and an imputed value, σ_{clct} is a standard deviation of collected values, and n is the number of data. Note that these indicators represent a distance so the smaller value is the better.

Table 2: Evaluation with NRMSE for the data 2019

Sales of the past survey	Industrial Classification	Purchases		Initial Inventories		Final Inventories		Salaries	
		base	z-tr	base	z-tr	base	z-tr	base	z-tr
Under 90 percentile	Construction	0.635	0.634	0.998	0.998	0.998	0.998	1.115	1.038
	Manufacturing	0.783	0.781	0.982	0.979	0.983	0.980	0.927	0.883
	Wholesale and retail trade	0.572	0.571	0.986	0.986	0.986	0.986	0.839	0.808
	Accomodations and eating & drinking services	0.619	0.614	1.000	0.999	1.000	0.999	0.823	0.803
	Living-related & personal services and amusement services	1.125	1.126	1.000	1.000	1.000	1.000	0.762	0.716
	Services except above	0.880	0.824	0.998	1.000	0.998	1.000	0.849	0.804
Equal or over 90 percentile	Construction	0.413	0.417	0.993	0.993	0.994	0.994	1.187	1.146
	Manufacturing	0.553	0.552	0.988	0.988	0.987	0.987	0.941	0.938
	Wholesale and retail trade	0.227	0.226	0.999	0.999	1.000	1.000	0.565	0.560
	Accomodations and eating & drinking services	0.629	0.624	0.999	0.999	0.999	0.999	0.747	0.752
	Living-related & personal services and amusement services	0.573	0.569	0.999	0.999	1.000	0.999	0.789	0.790
	Services except above	0.711	0.731	0.992	1.002	0.995	1.001	0.794	0.791

Table 3: Evaluation with MAE for the data 2019

Sales of the past survey	Industrial Classification	Purchases		Initial Inventories		Final Inventories		Salaries	
		base	z-tr	base	z-tr	base	z-tr	base	z-tr
Under 90 percentile	Construction	2066.4	2073.2	503.5	441.3	464.0	407.5	1113.5	951.6
	Manufacturing	1317.5	1316.5	556.3	521.7	550.7	514.8	772.9	697.5
	Wholesale and retail trade	2077.2	2079.5	1855.7	1855.7	1827.7	1827.7	639.1	570.3
	Accomodations and eating & drinking services	1144.6	1135.6	169.7	169.0	173.3	172.8	670.6	629.0
	Living-related & personal services and amusement services	449.1	444.5	149.4	147.1	147.1	145.1	368.6	301.3
	Services except above	733.0	617.0	557.8	337.5	564.9	340.7	525.8	435.8
Equal or over 90 percentile	Construction	6345.8	6400.2	1922.4	1894.1	1934.3	1906.4	4425.3	4152.5
	Manufacturing	6413.7	6469.5	2566.0	2547.9	2651.7	2632.9	4145.2	4079.9
	Wholesale and retail trade	9830.2	9773.6	5470.5	5470.5	5619.5	5619.5	4045.3	4003.2
	Accomodations and eating & drinking services	3614.6	3615.3	538.2	538.2	535.6	535.6	2616.5	2624.6
	Living-related & personal services and amusement services	1942.3	1936.3	401.7	398.8	408.1	404.9	2364.3	2324.0
	Services except above	3092.0	2606.7	936.1	631.4	1191.6	778.2	4620.6	4557.0

The hot deck donor selection and mean value calculation were conducted within each imputation class of the survey, which is the combination of Medium groups of Industrial Classification and 2-groups divided by the 90 percentile of sales of the past survey. To simulate specifying zeros, all the data was divided into 10 groups and each data of one of the groups was predicted whether zero or not using the model fit to the rest of the groups (it is associated with 10-fold cross-validation), while the hot deck imputation was taken as if each data was missing uniquely in the imputation class (it is associated with the leave-one-out method) to simulate the best performance because its evaluation is not the interest here. Table 2 to 5 shows the result of the evaluation, which rolls Medium groups of Industrial Classification up to Major groups of Industrial Classification for a practical view. These tables compare the current imputation process ('base') and that combined with the zero-treating method ('z-tr'). For the values of NRMSE, we hardly observe the difference between the two methods. However, the two-step imputation shows the smaller values for MAE, especially for those of under 90 percentile of past sales. It seems to be improved the accuracy of the imputation by specifying and imputing zeros. In addition, the improvement is significant for the imputation class which contains many zero values (e.g. Industry of 'Services except above'). The improvement that appears on inventories can be explained as inventories are originally imputed by a mean value, which produces a persistent gap between the collected zero value and the imputed value while imputing zeros doesn't produce a gap. On the other hand, through this simulation, the number of the data which is imputed zero at the first step was relatively small, which means the affected data by this additional step was limited. Meanwhile, NRMSE tends to exaggerate the distance of large values because it includes a squared term. It might be the reason why the values of NRMSE don't show considerable change after the zero-treating method. We think a further examination is needed for this matter.

Table 4: Evaluation with NRMSE for the data 2020

Sales of the past survey	Industrial Classification	Purchases		Initial Inventories		Final Inventories		Salaries	
		base	z-tr	base	z-tr	base	z-tr	base	z-tr
Under 90 percentile	Construction	0.661	0.654	0.999	1.000	0.999	0.999	1.167	1.075
	Manufacturing	0.828	0.825	0.988	0.984	0.986	0.981	0.894	0.855
	Wholesale and retail trade	0.611	0.612	0.986	0.986	0.986	0.986	0.683	0.656
	Accomodations and eating & drinking services	0.593	0.578	0.999	0.998	0.999	0.999	0.810	0.789
	Living-related & personal services and amusement services	1.052	1.047	1.000	0.999	1.000	0.999	0.633	0.578
	Services except above	1.012	0.979	0.997	1.000	0.998	1.001	0.849	0.811
Equal or over 90 percentile	Construction	0.412	0.408	0.997	0.997	0.997	0.997	1.095	1.099
	Manufacturing	0.513	0.508	0.989	0.987	0.990	0.989	0.965	0.955
	Wholesale and retail trade	0.219	0.212	0.999	0.999	0.999	0.999	0.538	0.536
	Accomodations and eating & drinking services	0.603	0.603	0.999	0.999	0.999	0.999	0.640	0.639
	Living-related & personal services and amusement services	0.510	0.508	1.000	0.999	1.000	0.999	0.627	0.624
	Services except above	0.726	0.740	0.991	1.000	0.990	0.999	0.811	0.804

4 Conclusion

Under the situation of being found a lot of zero values, specifying zero values using binary classification method seems to work well, and utilizing it for imputation looks preferable. We considered the zero-specifying model and tried combining it with the

Table 5: Evaluation with MAE for the data 2020

Sales of the past survey	Industrial Classification	Purchases		Initial Inventories		Final Inventories		Salaries	
		base	z-tr	base	z-tr	base	z-tr	base	z-tr
Under 90 percentile	Construction	1977.3	1979.7	381.0	329.3	379.4	328.2	1179.7	997.9
	Manufacturing	1277.4	1248.6	548.6	490.8	537.8	482.4	689.4	615.6
	Wholesale and retail trade	1960.7	1971.5	1691.9	1690.1	1685.4	1683.7	560.2	498.7
	Accomodations and eating & drinking services	1091.1	1081.3	147.7	146.9	157.7	157.2	607.2	563.0
	Living-related & personal services and amusement services	431.7	430.9	122.4	120.5	126.8	125.1	292.9	228.8
	Services except above	692.5	579.9	323.9	204.7	397.2	244.1	547.1	451.7
Equal or over 90 percentile	Construction	6675.0	6654.7	1929.7	1901.3	2049.9	2018.7	4342.3	4129.7
	Manufacturing	6033.2	6023.0	2828.7	2727.1	2859.0	2758.7	4049.9	3962.3
	Wholesale and retail trade	10179.7	10056.1	5903.8	5901.4	6039.4	6037.0	3890.5	3861.5
	Accomodations and eating & drinking services	3549.5	3542.9	460.5	460.5	458.3	458.3	2599.8	2595.6
	Living-related & personal services and amusement services	1917.5	1914.5	402.3	399.7	401.1	398.8	2432.7	2373.0
	Services except above	2886.1	2410.7	684.7	492.3	641.3	463.0	4656.8	4523.2

imputation process of Unincorporated Enterprise Survey. The results of the simulation showed the advantage of specifying zeros preceding the current imputation process, while the effectiveness of that still needs to be evaluated carefully from a more detailed point of view. It could be worth considering adjusting some parameters of the model such as the number of variables or thresholds. We will take further deliberation on this method and consider its application for ongoing data processing.

About the Data

In this paper, the microdata of Unincorporated Enterprise Survey conducted by Statistics Bureau, Ministry of Internal Affairs and Communications (MIC), was used for statistical research under the Statistics Act of Japan. All statistical results herein are produced by the author while MIC doesn't necessarily disseminate these results.

The views and opinions expressed in this paper are those of the author, and not necessarily those of National Statistics Center, an Incorporated Administrative Agency of MIC to perform data processing for official statistics.

References

[1] CANCEIS (2020), Users Guide V5.4. CANCEIS Development Team. Statistical Integration Methods Division, Statistics Canada, 2020.

[2] Farnell, J. and Darby, P. (2020), Administrative Data Informed Donor Imputation in the Australian Census of Population and Housing. Statistical Journal of the IAOS, vol. 36, no. 1, pp. 117-124.

[3] Mutoh, A. and Shirakawa, K. (2023), 探索的データ解析のための尺度水準および尺度水準に基づく要約統計量. データサイエンス研究, vol. 2.