 ISBN: 9789073592421

CPS Paper

Imputing zeros in business survey items using a binary classification method

Author: Mr Ichiro Murata

Submission ID: 657

Reference Number: 657

Presentation File

abstracts/ottawa-2023_e5613aaaea44e9fe440dff3aafe94723.pdf

Files/Uploads

CPS0032_IchiroMurata

Brief Description

Imputation has been an important topic of research for the National Statistics Center of Japan.

Here we are interested in some characteristics of business survey items and utilizing it to improve the existing imputation method.

Abstract

Missing values, which would deteriorate the statistics results, should be treated properly. As regards to Unincorporated Enterprise Survey conducted annually in Japan, some imputation methods have been taken to fill in the accounting items such as cost, salary, and inventory. Missing values in these items are, in the current data processing, filled by a mean value or hot-deck imputation. However, 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 are zero, which seems difficult to be imputed by our approach so far.
Under the situation, specifying whether a missing value is zero or not in advance would be helpful for appropriate imputation, hereafter regarded as a binary classification problem that an accounting value is zero or not. Generally, logistic regression is a popular method to apply to such a condition, while the variable selection for the model is an important issue.
We tried to find good explanatory variables for the model using the microdata of the Unincorporated Enterprise Survey 2019-2021. Taking all the categorical items as candidates, we examined the coefficients of the logistic and lasso model, as well as the AIC calculated in a process of the stepwise variable selection. For practical reasons, instead of adopting all the items, we selected the most reasonable 5 items as variables, and their performance of classifying zeros measured by a simulation with the microdata reached almost 80% precision at a certain threshold of the model.
Subsequently, missing values classified as not zero may be imputed on the current data processing that fills in them by a mean value or hot-deck imputation. We evaluated the effect of specifying zeros through a simulation of the current imputation and that after zero specification. The result showed that imputing zeros in advance decreased the difference between an imputed value and an actual answered 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.