A draft version

An application of a complex measure to model–based imputation in business statistics

Andrzej Młodak
Statistical Office in Poznań, Small Area Statistics Centre, e–mail: a.mlodak@stat.gov.pl

Abstract

Faced with missing data in a statistical survey or administrative sources it is necessary to apply imputation. The technique ensures the best possible way to fill the gaps and reduce the major part of bias that can affect aggregated estimates as a consequence of these gaps. This paper presents research on the efficiency of model–based imputation in business statistics, where the explanatory variable is a complex measure constructed by taxonomic methods. The proposed approach involves selecting from a set of possible explanatory variables for imputed information the best ones in terms of variation and correlation and then replacing them with a simple meta–feature exploiting their whole informational potential. In this way the resulting imputation models are simpler: they are less computationally demanding and easier to interpret, while being sufficiently efficient. The following five types of such imputation techniques are presented: ratio imputation, multiple imputation, multiple imputation with iteration, predictive mean matching and the propensity score method. A simulation study and empirical study were used to verify the efficiency of the proposed approach. The second study involved a simulation of missing data using IT business data from the California State University in Los Angeles, USA.

Keywords: *model–based imputation, taxonomic measure of development, ratio imputation, multiple imputation, predictive mean matching, propensity score method.*

Aim and scope

The main purpose of this paper is to study the efficiency of the use of a complex measure as an auxiliary variable in some methods of model–based imputation, especially for business statistics. A complex measure reflects the diversification of entities in terms of a complex social or economic phenomenon, described by many variables. A measure of this kind is constructed in such a way as to ensure that information contained in the variables and mutual relationships between them are maximally exploited, which traditional models of dependency (i.e. regression function) can overlook. The utility of such an approach will be verified using the following methods of model–based imputation: ratio imputation, multiple regression imputation, multiple regression imputation with iterative extension, predictive mean matching, propensity score method.

This paper is a development of some issues which were not included in the final version of some sections of „*Handbook on Methodology for Modern Business Statistics*", edited by L. Willenborg, S. Scholtus and R. van de Laar (Collaboration in Research and Methodology for Official Statistics), created in 2014 r. within the ESSnet project MeMoBuSt, which were investigated during that project.

Construction of a complex measure

The construction of a complex measure consists of the following steps.

Step 1. Choice of variables and data collection: one should use information which properly describes the subject of research. The collected variables containing such information should be measurable, complete and comparable. To improve data comparability, they should have the form of indices (i.e., need to be calculated per capita, per 1 km^2, per 1000 inhabitants, per enterprise, etc.).

Step 2. Verification of variables: firstly, the elimination of variables that are not effective in discriminating entities, i.e. dropping variables for which the absolute value of the coefficient

of variation (CV) is smaller than an arbitrarily established threshold (usually 10). Next, variables are verified in terms of correlation – we eliminate variables that are too correlated with others (and, hence, carry similar information). Here the inverse correlation matrix method was used: its diagonal entries belong usually to $[1, \infty)$; if they are >10 or < 1 then there exist 'bad' variables; they should be carefully eliminated).

Step 3. Identification of the character of diagnostic features (variables after verification): considering the impact of variables on the situation of an entity with respect to a phenomenon of interest, we can we distinguish three types of variables:

- *stimulants* – the higher the value, the better the situation of an entity in a certain sense
- *destimulants* – higher values indicate a deterioration of the entity's situation
- *nominants* – variables which behave like stimulants below a certain critical point and may switch to being destimulants after crossing it.

Destimulants and nominants are converted into stimulants by taking their values with opposite signs (in the case of nominants this is done only to the part with destimulative properties).

Step 4. Normalization of features, aimed at obtaining a comparable form of diagnostic variables. To exploit all connections between them it is good to use the Weber median, i.e. the vector $\Theta = (\theta_1, \theta_2, \dots, \theta_m) \in \mathbb{R}^m$ minimizing the sum of Euclidean distance from points reflecting given entities (cf. A. Młodak (2006)):

$$z_{ij} = \frac{x_{ij} - \theta_j}{1{,}4826 \cdot \underset{i=1,2,\dots,n}{\text{med}} \left| x_{ij} - \theta_j \right|}, \quad i = 1,2, \dots, n,\ j = 1,2, \dots, m.$$

Step 5. Definition and determination of the taxonomic benchmark of development – an artificial, ideal entity is defined, which others are compared with, defined endogenously:

$$\psi_j = \max_{i=1,2,\dots,n} z_{ij}, \qquad j = 1,2, \dots, m$$

Step 6. Computation of distances of entities from the benchmark

$$d_i = \underset{j=1,2,\dots,m}{\text{med}} \left| z_{ij} - \psi_j \right|, \quad i = 1,2, \dots, n.$$

Step 7. Determination of a synthetic measure

$$\mu_i = 1 - \frac{d_i}{\text{med}(\boldsymbol{d}) + 2{,}5 \cdot \text{mad}(\boldsymbol{d})},$$

$i = 1,2, \dots, n$, $\boldsymbol{d} = (d_1, d_2, \dots, d_n)$, $\text{mad}(\boldsymbol{d}) = \underset{i=1,2,\dots,n}{\text{med}} |d_i - \text{med}(\boldsymbol{d})|$.

Investigated methods of model–based imputation

Now we describe briefly the methods of model–based imputation analysed in the paper.

Ratio imputation consists in replacing missing values with the value of a known auxiliary variable multiplied by the ratio of some descriptive summary statistics of the variable with the missing value (e.g. mean, median or sum) and the relevant statistics for the auxiliary variable. It is here tacitly assumed that the ratio of the values of these variables for a given unit is the same as the ratio of some 'total' values of these two variables.

In ***multiple regression imputation*** missing values are replaced with predicted values established using a specific regression equation constructed on the basis of the available data for the variable with gaps (as the value of the dependent variable resulting from the regression

model) and some fully available auxiliary variables treated as explanatory variables. The basic model is given by:

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j X_j,$$

where $\mathrm{Y} = (\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_\mathrm{n})$ is the target variable with gaps, $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{m}$ $(\mathrm{m} \in \mathbb{N})$ – auxiliary variables. OLS estimator of coefficients has the form $\hat{\beta} = (\mathbf{X}_\mathrm{r}^\mathrm{T}\mathbf{X}_\mathrm{r})^{-1}\mathbf{X}_\mathrm{r}^\mathrm{T}\mathrm{Y}_\mathrm{r}$, where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_\mathrm{m})$, $\mathbf{X}_\mathrm{r}$ and Y_r are matrix $\mathbf{X} = [\mathrm{x}_{\mathrm{ij}}], \mathrm{x}_{\mathrm{i0}} = 1,\ \mathrm{i} = 1,2, \dots, \mathrm{n},\ \mathrm{j} = 0,1,2, \dots, \mathrm{m}$ and vector Y restricted only to those units for which data on Y are available, respectively.

Multiple regression imputation with iterative extension can be used if there are many variables $Y_1, Y_2, \dots, Y_k,\ k \in \mathbb{N}$ to be imputed using the same set of covariates as in the classic case. The missing values are then replaced by predictors obtained from the equation

$$Y_{r_l l} = \beta_{*0} + \sum_{j=1}^{m} \beta_{*m} X_{r_l j} + z_{r_l}\sigma_{*l},$$

where $X_{\mathrm{r}_l j}$ are the values of covariates for such units for which data on Y_l are unavailable and z_i is a simulated normal deviate, $r_l = 1,2, \dots, n,\ l = 1,2, \dots, k$. This operation can then be repeated starting from the above formula and so on. The number of iterations depends on the assumptions of the quality control (cf. Y. C. Yuan (2010)). The synthetic measure can be applied instead of the set of covariates.

Predictive mean matching is method similar to the regression method with iterative extension, except that, instead of the main predictive equation for each missing value, it imputes an observed value which is closest to the predicted value from the simulated regression model (cf. Y. C. Yuan (2010), N. J. Horton and S. R. Lipsitz (2001)).

Propensity score method is based on the propensity score understood as the conditional probability of assignment to a particular treatment, given a vector of observed covariates. In this method, the propensity score is generated for each variable with missing values to indicate the probability of that observation being missing (Y. C. Yuan (2010)). The observations are then grouped on the basis of these propensity scores and an approximate Bayesian bootstrap imputation (stepwise algorithm of imputation under monotone missing pattern assumption).

MSE and its decomposition based on imputed data

Let $\hat{\theta}_A$ be an estimator of parameter θ computed using all sample data about the target variable. C. E. Sårndal (1992) showed that the total variance or – in terms of the theory of estimation – Mean Squared Error (MSE), $\hat{V} = E\left(\hat{\theta} - \theta\right)^2$, can be decomposed into sampling, imputation and mixed effect components: $\hat{V} = \hat{V}_{\mathrm{SAM}} + \hat{V}_{\mathrm{IMP}} + 2\hat{V}_{\mathrm{MIX}}$, where $\hat{V}_{\mathrm{SAM}} = E\left(\hat{\theta}_A - \theta\right)^2$, $\hat{V}_{\mathrm{IMP}} = E\left(\hat{\theta} - \hat{\theta}_A\right)^2$, $\hat{V}_{\mathrm{MIX}} = E\left(\left(\hat{\theta}_A - \theta\right)\left(\hat{\theta} - \hat{\theta}_A\right)\right)$. MSE and its components can be estimated as

$$\tilde{V} = \frac{1}{|A|^2}\sum_{i \in A}\left(y_i^* - \hat{\tilde{\theta}}_A\right)^2 = \tilde{V}_{\mathrm{SAM}} + \tilde{V}_{\mathrm{IMP}} + 2\tilde{V}_{\mathrm{MIX}}$$

with sampling effects $\tilde{V}_{\mathrm{SAM}} = \frac{1}{|A|^2}\sum_{i \in A}\left(\tilde{y}_i - \hat{\tilde{\theta}}_A\right)^2$, imputation effects $\tilde{V}_{\mathrm{IMP}} = \frac{1}{|A|^2}\sum_{i \in A}(y_i^* - \tilde{y}_i)^2$ and mixed effects $\tilde{V}_{\mathrm{MIX}} = \frac{1}{|A|^2}\sum_{i \in A}(y_i^* - \tilde{y}_i)\left(\tilde{y}_i - \hat{\tilde{\theta}}_A\right)$

where $y_i^* = py_i + (1-p)\hat{y}_i$ and $\tilde{y}_i = py_i + (1-p)\left(\left(n\hat{\bar{\theta}}_A - |R|\hat{\bar{\theta}}_R\right)/|R|\right)$, $\hat{\bar{\theta}}_A = \sum_{i\in A} y_i^*/|A|$ with $p = 1$ if the value of Y for i-th unit on Y is available and $p = 0$ otherwise, $\hat{\bar{\theta}}_R = \sum_{i\in R} \hat{y}_i/|R|$ and $|\cdot|$ denotes the cardinality of a given set.

Simulation study

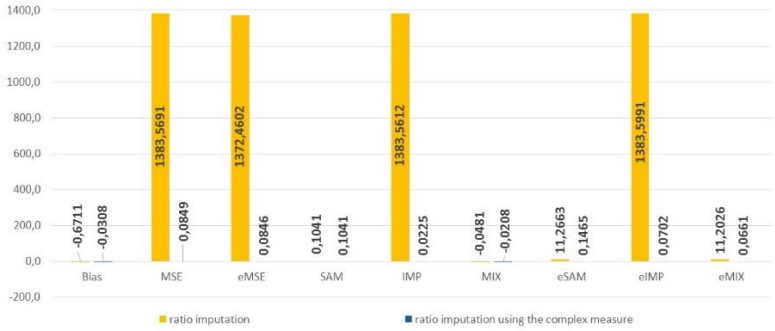
To verify the efficiency of our method a simulation experiment was conducted. A sample consisting of 200 units was constructed to account for circumstances observed in business statistics. The values of target variable Y and first auxiliary variable X_1 were drawn from the two–variate normal distribution with $\boldsymbol{\mu} = (2,10)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(11,2)$. Next, four auxiliary variables were defined as $X_2 = X_1 - 10r$, $X_3 = X_1 - 333r$, $X_4 = X_1 + 15r$ and $X_5 = X_1 + 255r$, where r is drawn from the standardized normal distribution with expected value 0 and variance 1 as the disturbance factor, separately for each of these variables.

This choice ensures that auxiliary variables are well diversified and are not or weakly correlated with the target variable and the disturbance factor is also taken into account. We have assumed that 20% of observations of Y are unavailable. This is the average rate of missing data observed in various statistical surveys. Thus, we have removed values of Y for 40 units selected randomly according to the uniform distribution on [0,1]. The experiment was repeated 1000 times and average quality indicators were computed.

Results of the simulation

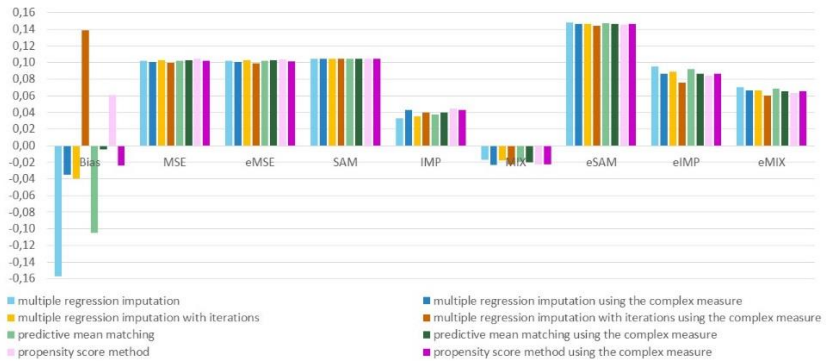
The results of the experiment in terms of quality indicators of imputation are presented in Fig 1 and Fig. 2. Given the much larger values for MSE, eMSE, IMP and eIMP for ratio imputation the results for this method are shown in separate figures.

Fig. 1. Results of the simulation for ratio imputation



Source: Results obtained using an algorithm written in SAS Enterprise Guide 4.3. (with IML environment).

Fig. 2. Results of the simulation for other methods



Source: Results obtained using an algorithm written in SAS Enterprise Guide 4.3. (with IML environment).

The prefix "e" denotes estimate of a given index. It is clear that the application of the complex measure instead of several separate auxiliary variables reduces either bias or MSE.

Empirical study

Our alternative study was based on data about 36 firms representing the IT sector in Bermuda, Canada, China, Denmark, Finland, Germany, Greece, Indonesia, Japan, Mexico, Russia, South Korea, Sweden, UK and USA, placed on Instructional Web Server of the California State University in Los Angeles, USA (http://instructional1.calstatela.edu/mfinney/Courses/491/hand/sas_exercise/tech3.xls). Five following variables are here recorded: Return on Equity (ROE, %), Revenues (in millions $)m Revenue Growth (%), Total Shareholder Return (%) and Profits (in millions $).

The set originally contained 39 firms, but due to missing data for ROE three had to be dropped. For the purposes of the study, the revenue was chosen as the variable to be imputed. 6 randomly (according to the uniform distribution) selected observations of revenue were removed. The same methods of imputation were applied as in the case of the simulation experiment, but with some adjustments. That is, the complex measure was composed of three indicator variables which were strongly diversified and weakly correlated with the revenue, i.e.: Return on Equity (ROE, %), Revenue Growth (%) and Total Shareholder Return (%). In classical ratio imputation the Total Shareholder Return was used as a reference variable, because it is the one most correlated with the target variable. As a result, basic descriptive statistics for revenue in complete and imputed sets were computed and compared. – see Table 1. We can see that in most cases the variations of methods which use the complex measure provide less biased results or better reflect the true distribution of the data.

Conclusions

The main conclusions which can be formulated on the basis of our studies are as follows. Efficient construction of a complex measure ensures a more efficient exploitation of mutual connections between possible auxiliary variables and therefore more informative imputation. Moreover, in most cases using a complex measure instead of the classic approaches reduces the bias of estimation or improves its precision. The complex measure provides more stable results, i.e. with a significantly lower risk of excessive outliers. However, one should remember that the conditions for the efficient use of a complex measure are: proper choice of auxiliary variables on the basis of which it is constructed and methods of its construction.

References

Horton, N. J. and Lipsitz, S. R. (2001), *Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables*, Journal of the American Statistical Association, vol. 55, pp. 244 – 254.

Młodak, A. (2014), *On the construction of an aggregated measure of the development of interval data*, Computational Statistics, 2014, vol. 29, str. 895 – 929.

Młodak A. (2006), *Multilateral normalisations of diagnostic features*, Statistics in Transition, vol 7., pp. 1125 – 1139.

Sårndal C. E. (1992), *Methods for estimating the precision of survey estimates when imputation has been used*, Survey Methodology, vol. 18., pp. 241 – 252.

Yuan Y. C. (2010), *Multiple Imputation for Missing Data: Concepts and New Development* (Version 9.0), SAS Institute Inc, Rockville, MD, U.S.A.

Table 1. Results of the empirical study – basic descriptive statistics.

Method	Mean	Variance	Minimum	Maximum	Lower Quartile	Median	Upper Quartile	Skew-ness	Kurtosis
Actual	**16571.47**	**467248738**	**623.10**	**83221.00**	**2479.85**	**7928.40**	**22353.50**	**1.7949**	**2.5318**
Classical ratio imputation	16845.50	524349964	-6083.35	83221.00	2401.74	7919.25	21937.58	1.7678	2.1689
Ratio imputation using the complex measure	16671.37	428181031	623.10	83221.00	3425.40	9276.10	21204.72	1.9779	3.4620
Multiple regression imputation	15446.17	545338613	-35193.42	83221.00	1961.50	8595.65	23026.99	1.1249	2.1510
Multiple regression imputation using the complex measure	16882.19	492366622	-23957.95	83221.00	2709.45	9276.10	22353.50	1.4566	2.2231
Multiple regression imputation with iterative extension	15711.63	543091000	-30116.28	83221.00	1445.50	7983.90	19301.32	1.3608	1.5474
Multiple regression imputation with iterative extension using the complex measure	16448.00	467858762	-13438.99	83221.00	2477.50	8020.10	25702.00	1.6518	2.0885
Predictive mean matching	16042.26	452294165	623.10	83221.00	1445.50	7983.90	19005.00	1.8423	2.5280
Predictive mean matching using the complex measure	15232.29	429598243	623.10	83221.00	2477.50	7983.90	15526.70	2.0377	3.3398
Propensity score method	15301.66	463397401	623.10	83221.00	1961.50	6889.15	14172.45	1.9436	2.8201
Propensity score method using the complex measure	17203.43	564862904	623.10	83221.00	1287.05	7919.25	17265.85	1.7607	1.9665

Source: Own elaboration using the algorithm written in SAS Enterprise Guide 4.3. (with IML environment).