

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

ANDRZEJ MŁODAK



Statistical Office in Poznań  
Small Area Statistics Centre

# Aim and scope

---

- Main goal: use a complex measure as an auxiliary variable in some methods of model-based imputation, especially for business statistics
- Complex measure reflects diversification of entities in terms of some composite social or economic phenomenon, described by many variables
- The utility of this approach will be verified using the following methods of imputation:
  - Ratio imputation
  - Multiple regression imputation
  - Multiple regression imputation with iterative extension
  - Predictive mean matching
  - Propensity score method

# Construction of a complex measure

---

## ■ Step 1. Choice of variables and data collection

- we use information which describes properly the subject of research.
- collected data should be measurable, complete and comparable.
- to improve the data comparability, they should have a form of indices (i.e., are required to be calculated per capita, per 1 km<sup>2</sup>, per 1000 inhabitants, per enterprise, etc.).

## ■ Step 2. Verification of variables

- elimination of variables having too small power of diversification of objects, i.e. such for that  $|CV|$  is smaller than the arbitrarily established threshold, usually 10,
- correlation verification – elimination of variables too much correlated with others (and, by the same way, being carriers of similar information). Two alternative approaches:
  - parametric method (based on maximal row sums of entries of correlation matrix)
  - inverse correlation matrix (its diagonal entries belong usually to  $[1, \infty)$ ; if they are  $>10$  or  $< 1$  then there exist 'bad' variables; their elimination should be carefully done).

# Construction of complex measure

- Step 3. Identification of character of diagnostic features (variables after verification)
  - *stimulants* – the higher is the value the better is the situation of object in a given sense
  - *destimulants* – higher values inform about worse situation of an object
  - *nominants* – having an imbuement point, below which it has a character of stimulant and above it – stimulant or on the contrary

Destimulants and nominants are converted into stimulants by taking their values with opposite signs (in the case of nominants it concerns only destimulative 'part' of variable).

- Step 4. Normalization of features using the Weber median

$$\Theta = (\theta_1, \theta_2, \dots, \theta_m) \in \mathbb{R}^m$$

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

# Construction of complex measure

- Step 5. Definition and determination of taxonomic benchmark of development

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

- Step 6. Computation of distances of objects from the benchmark

$$d_i = \operatorname{med}_{j=1,2,\dots,m} |z_{ij} - \psi_j|, \quad i = 1, 2, \dots, n.$$

- Step 7. Determination of synthetic measure

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

$$i = 1, 2, \dots, n, \mathbf{d} = (d_1, d_2, \dots, d_n), \operatorname{mad}(\mathbf{d}) = \operatorname{med}_{i=1,2,\dots,n} |d_i - \operatorname{med}(\mathbf{d})|.$$

# Model-based methods of imputation

## ■ Ratio imputation

- 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.
  - for example, if data about the value of sales for an enterprise is missing, but its total expenditure amounts to €20,000, mean sales for the whole analyzed group of enterprises, which the given one belongs to is €30,000 and the mean expenditure is €21,000, then the predicted sales are computed as  $20,000 \times (21,000 / 30,000) = 20,000 \times (7 / 10) = €14,000$ .
- one can, of course, make a choice of the best auxiliary variable from several variables which are strictly connected with the imputed one, e.g. by analyzing the distribution of the known values of the imputed variable and appropriate values of the possible auxiliary variable (e.g. using the Wilcoxon signed rank test or Pearson's correlation coefficient). The synthetic measure allows, however, for avoiding loss of important information contained in the 'rejected' variables.

# Model-based methods of imputation

- 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

- basic model

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

$Y = (y_1, y_2, \dots, y_n)$  – variable with gaps,  $X_1, X_2, \dots, X_m$  ( $m \in \mathbb{N}$ ) – auxiliary variables; OLS estimator of coefficients:  $\hat{\beta} = (\mathbf{X}_r^T \mathbf{X}_r)^{-1} \mathbf{X}_r^T Y_r$ , where  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)$ ,  $\mathbf{X}_r$  and  $Y_r$  are matrix  $\mathbf{X} = [x_{ij}]$ ,  $x_{i0} = 1, i = 1, 2, \dots, n, j = 0, 1, 2, \dots, m$  and vector  $Y$  restricted only to those units for which data on  $Y$  are available, respectively.

- replacing several (sometimes many) explanatory variables with synthetic measure constructed using them exploits all necessary information and saves time of processing.

# Model-based methods of imputation

## ■ Multiple regression imputation with iterative extension

- this method 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 classical case
- let  $\hat{\sigma}_l^2 \mathbf{V}_l$  (where  $\mathbf{V}_l = (\mathbf{X}_{r_l}^T \mathbf{X}_{r_l})^{-1}$  and  $\hat{\sigma}_l^2$  be the estimated variance of  $Y_l$ ) be a covariance matrix for the model with  $Y_l, l = 1, 2, \dots, k$  being the explained variable in classical formula. We start from the classical model and next new parameters  $\beta_* = (\beta_{*0}, \beta_{*1}, \dots, \beta_{*m})$  and  $\hat{\sigma}_{*l}^2$  are drawn from the posterior predictive distribution of the parameters.
- 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  $X_{r_l j}$  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 be next repeated starting from the above formula and so on. The number of iterations depends on the assumptions of the quality control. The synthetic measure can be here applied instead of the set of covariates.



# Model-based methods of imputation

---

## ■ Predictive mean matching

- this method is 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. The predictive mean matching method ensures that imputed values are plausible and may be more appropriate than those obtained using the standard regression method, even if the normality assumption is omitted
- the classical multiple regression (with iterative extension) and predictive mean matching algorithms can be practically realized using the `mi` procedure of the SAS 9.2 software with the monotone options: `reg( )` (for classical multiple regression) or `regpmm( )` (for predictive mean matching) command with appropriate arguments.

# Model-based methods of imputation

## ■ Propensity score method

- the propensity score is 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. The observations are then grouped on the basis on these propensity scores and an approximate Bayesian bootstrap imputation,
- the stepwise algorithm of imputation under monotone missing pattern assumption
  - 1. Create an indicator variable  $\Lambda_l$  with the value 0 for observations with missing  $Y_l$  and 1 otherwise.
  - 2. Fit a logistic regression model

$$\text{logit}(p_l) = \beta_0 + \sum_{j=1}^m \beta_j X_j ,$$

where  $p_l = \Pr(\Lambda_l = 0 | X_1, X_2, \dots, X_m)$  and  $\text{logit}(p_l) = \log(p/(1-p))$ .

- 3. Create a propensity score for each observation to estimate the probability that it is missing.

# Model-based methods of imputation

## ■ Propensity score method (cont.)

- the stepwise algorithm of imputation under monotone missing pattern assumption (cont.)
  - 4. Divide the observations into a fixed number of groups (typically assumed to be five) based on these propensity scores. This can be done by arbitrarily establishing some structure of intervals of propensity values and indicating observations whose propensity values belong to such particular intervals.
  - 5. Apply approximate Bayesian bootstrap imputation to each group. That is, for a given group, suppose that  $Y_{obs}$  denotes the  $n_1$  observations with nonmissing  $Y_l$  values and  $Y_{mis}$  denotes the  $n_0$  observations with missing  $Y_l$  ( $n_0 < n_1$ ). Approximate Bayesian bootstrap imputation first draws  $n_1$  observations randomly with replacement from  $Y_{obs}$  to create a new data set  $Y_{obs}$ . This is a nonparametric analogy of drawing parameters from the posterior predictive distribution of the parameters. The process then draws the  $n_0$  values for  $Y_{mis}$  randomly with replacement from  $Y_{obs}$ . These values are implants.
  - Steps 1 through 5 are repeated sequentially for each variable with missing values.
- to implement this procedure one can use also the `mi` procedure of the SAS 9.2 software but with the `monotone propensity()` command with appropriate arguments

# Measurement of the quality of imputation

## ■ MSE and its decomposition based on imputed data

- Let  $\hat{\theta}_A$  be an estimator of parameter  $\theta$  computed using all sample data about the target variable. C. E. Särndal (1992) showed that the total variance or – in terms of the theory of estimation – MSE,  $\hat{V} = E(\hat{\theta} - \theta)^2$ , can be decomposed in the sampling, imputation and mixed effect components:

$$\hat{V} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}},$$

where  $\hat{V}_{\text{SAM}} = E(\hat{\theta}_A - \theta)^2$ ,  $\hat{V}_{\text{IMP}} = E(\hat{\theta} - \hat{\theta}_A)^2$ ,  $\hat{V}_{\text{MIX}} = E((\hat{\theta}_A - \theta)(\hat{\theta} - \hat{\theta}_A))$ .

## ■ Approximation of the MSE and its components

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

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

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

# Simulation study

## ■ An experiment

- a sample consisting of 200 units with simulation taking into account circumstances observed in business statistics of was constructed:
  - the values of target variable  $Y$  and first auxiliary variable  $X_1$  were drawn from the two-variate normal distribution with  $\mu = (2,10)$  and  $\Sigma = \text{diag}(11,2)$
  - 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  $Y$  is unavailable. This rate is, on average, observed in practice 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.

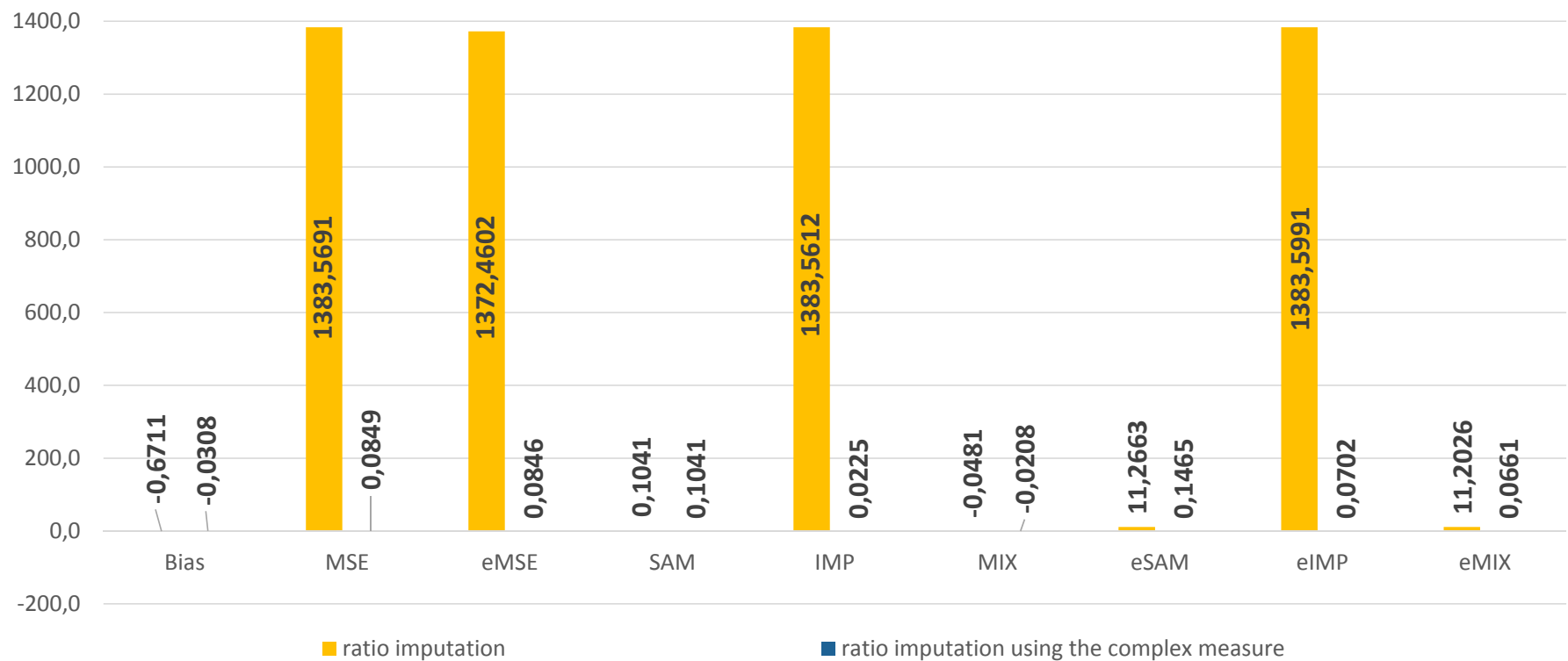
# Simulation study

---

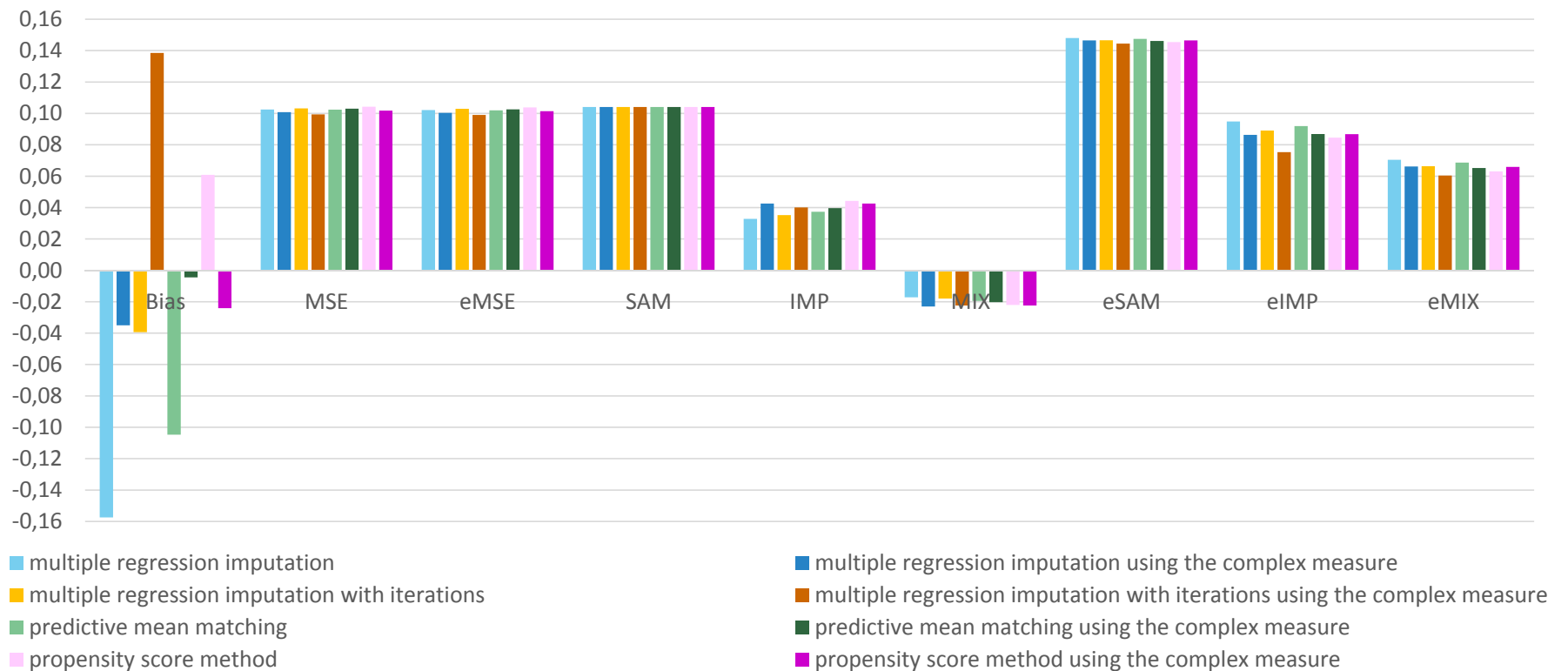
## ■ Simulation experiment (cont.)

- used options of imputation
  - classical ratio imputation (based on the variable best correlated with the target one),
  - ratio imputation using the complex measure,
  - multiple regression imputation,
  - multiple regression imputation using the complex measure,
  - multiple regression imputation with iterative extension (10 iterations),
  - multiple regression imputation with iterative extension using the complex measure (10 iterations),
  - predictive mean matching,
  - predictive mean matching using the complex measure,
  - propensity score method,
  - propensity score method using the complex measure
- algorithm prepared in the SAS Enterprise Guide 4.2. (and its IML environment) was used.

# Results of the simulation – ratio imputation



# Results of the simulation – other methods





# Empirical study

---

## ■ Data source

- data on 36 firms representing IT sector from 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](http://instructional1.calstatela.edu/mfinney/Courses/491/hand/sas_exercise/tech3.xls))
- five variables:
  - Return on Equity (ROE, %)
  - Revenues (in millions \$)
  - Revenue Growth (%)
  - Total Shareholder Return (%)
  - Profits (in millions \$)
- the set primarily contained 39 firms, but due to missing data for ROE three had to be dropped.

# Empirical study

---

## ■ Rules of the study

- as the imputed variables the revenues were assumed
- 6 randomly (according to the uniform distribution) selected observations of revenues were removed
- the same methods of imputation as in the case of simulation experiment were used
- to the construction of complex measure three variables having the form of indicator, all these indicators were strongly diversified and weakly correlated with the revenues and being stimulants were used, i.e.:
  - Return on Equity (ROE, %)
  - Revenue Growth (%)
  - Total Shareholder Return (%)
- in classical ratio imputation the Total Shareholder Return was used as a reference
- basic descriptive statistics for revenues in complete and imputed sets were computed.

# Empirical study

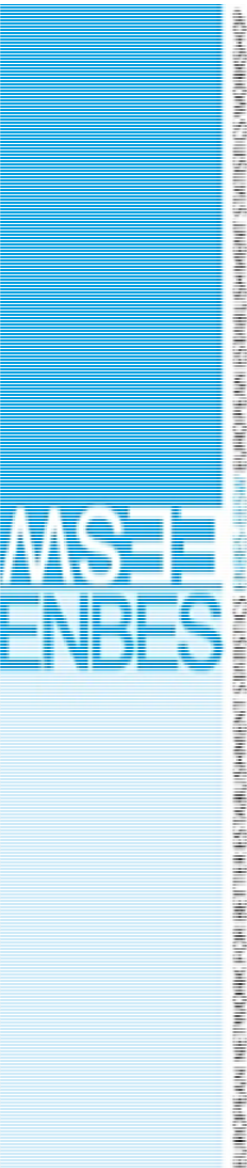
## ■ Results

Method	Mean	Variance	Minimum	Maximum	Lower Quartile	Median	Upper Quartile	Skew-ness	Kurto-sis
<b>Actual</b>	<b>16571.47</b>	<b>467248738</b>	<b>623.10</b>	<b>83221.00</b>	<b>2479.85</b>	<b>7928.40</b>	<b>22353.50</b>	<b>1.7949</b>	<b>2.5318</b>
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

# Conclusions

---

- Efficient construction of a complex measure ensures more efficient exploitation of mutual connections between possible auxiliary variables and therefore more informative imputation
- In most cases using complex measure instead of classic approaches reduces the bias of estimation or improves its precision
- Applying the complex measure provides more stable results, i.e. with a significantly smaller risk of obtaining excessive outliers
- One should remember that the conditions for efficient use of the complex measure are: proper choice of auxiliary variables on the basis of which it is constructed and methods of its construction.



---

**Thank you very much  
for your attention!**

---

