WP. 34
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24–26 September 2012)

Topic (v): Software & tools for data editing and imputation

APPLICATION OF THE DEVELOPED SAS MACRO FOR EDITING AND IMPUTATION AT STATISTICS LITHUANIA

Prepared by Vilma Nekrašaitė-Liegė and Jurga Rukšėnaitė, Statistics Lithuania, Lithuania

I. Introduction

1. The development of statistical process is rather difficult. Therefore, in practice, there are almost no surveys proving that there were no random errors or outliers or missing values in the data. In such cases data editing is used. Statistical data are checked using editing rules. Different logical, arithmetical or statistical rules are constructed for separate variables or groups of variables. These rules are based on a questionnaire, the analysis of statistical data or expert opinion. Using suitable editing rules, a specialist may find the missing values, outliers, and other errors.
2. In this paper, we will refer to the recommendations for editing and imputation of Luzi et al. (2007). All errors may be grouped into missing, systematic, influential errors and outliers, and random errors. Missing values (nonresponse) may appear due to several reasons. The respondent may not know the answer, may not be willing to respond or may have simply missed a question. A systematic error is an error that is reported consistently over time by responding units. This phenomenon is caused either by the consistent misunderstanding of a question or by the consistent misinterpretation of certain answers. Influential errors are errors in the values of variables that have a significant influence on publication target statistics for those variables. An outlier is an observation which is not fitted well by the model. The definition of an outlier is strictly related to the concept of an influential observation. Random errors are not caused by a systematic reason, but by accident. They may arise due to inattention of respondents, interviewers and data processing staff during various phases of the survey cycle.
3. The aim of editing/imputation is to have a data set without missing values. All values that do not meet certain requirements should also be edited. The main requirement for the methods is that after the editing/imputation the distribution of the parameter's estimator should not be changed. Therefore, it is useful to investigate the features of editing/imputation methods and to find out how these methods will affect statistical data.
4. Statistical data editing/imputation may be grouped into three stages: primary editing, micro editing, and macro editing. At Statistics Lithuania, micro editing is done by the specialists of regional statistical offices or, in some cases, by the central staff of Statistics Lithuania. Macro editing is done by the specialists of separate divisions. Usually, changes compared to the previous period are checked. If the values of macro data do not meet the editing rules, micro editing is performed. Micro editing is the second stage of the editing/imputation process that takes much time and requires many human resources. In this paper, we consider only micro editing.
5. Currently, micro editing at Statistics Lithuania is done by separate divisions independently. Most of editing/imputation rules for checking the variables from different surveys are programed in statistical software SAS. It was found out that some specialists who are responsible for surveys use quite sophisticated methods, but some specialists use too few editing/imputation rules and make things manually which could be done by programming tools.
6. As it is known, the manual detection of errors (values that do not meet certain requirements) takes a lot of time. Thus, it was decided to convert data editing/imputation into a concrete autonomous

process and to standardize all editing/imputation operations. In other words, the **essential objectives** were as follows:

(a) To define general editing/imputation principles;

(b) To define specific data editing/imputation rules, which can be applied to a separate statistical survey according to demand;

(c) To describe the editing/imputation process and to create suitable algorithms.

7. At Statistics Lithuania, most specialists use SAS for data editing/imputation. Moreover, we have adapted the experience of other countries. Consequently, it was decided that the most suitable programing and calculation tool is statistical program SAS. To improve the work of statisticians at Statistics Lithuania, SAS Macro program was developed for editing and imputation. It consists of the following parts: detection of errors, detection of outliers, imputation using the nearest neighbor method, imputation using models, imputation using distributions.

8. The structure of the paper is as follows: the theoretical issues of editing/imputation methods used are described in sections II and III; some practical applications are demonstrated in Section IV; concluding remarks are presented in Section V.

II. Detection of errors and outliers

A. Detection of errors

9. For the detection of errors, one type of checking rules has been programed. A checking rule is a logical condition or a restriction to the value of a data item or a data group which must be met if the data are to be considered correct. To detect errors using this specific SAS Macro, first of all, the user must have two data sets. One data set is for the data where at least three types of variables must be present: identification variable, study variable (variable of interest), and auxiliary variables. The other data set is used for a checking rule specification. It should be said that, if auxiliary variables have specific values, the study variable value must fall within a specific interval. Having these data sets, the user of SAS Macro just needs to enter the names of data sets and variables, and the program automatically creates a new data set, in which all units with errors are saved.

B. Detection of outliers using a general method

10. For the detection of outliers, several methods have been programed:

(a) Universal method;

(b) Interval method;

(c) Standard deviation rule;

(d) Testing of hypothesis.

11. Using the universal method, the outlier is the value of the study variable which is smaller than $q_1 - 3d$ or bigger than $q_3 + 3d$. A conditional outlier is such a value which belongs to the interval

$$[q_1 - 3d;\ q_1 - 1{,}5d) \text{ or } (q_3 + 1{,}5d;\ q_3 + 3d]\,, \qquad (1)$$

where q_1, q_3 are the first and the third quartiles, and $d = q_3 - q_1$.

12. Using the interval method, the program first finds out if the study variable has the normal distribution or not. If the hypothesis about normality is not rejected, then the outlier is the value of the study variable which does not belong to the interval

$$(\bar{y} - c_a\hat{s};\ \bar{y} + c_a\hat{s}), \qquad (2)$$

where $\bar{y}$ is a sample mean of the study variable, $\hat{s}$ is an estimator of standard deviation, and $c_a = 1.96$ is a quintile of standard normal distribution. If the hypothesis about normality is rejected, then the outlier is the value of the study variable which does not belong to the interval

$$(q_2 - c_a d_a;\ q_2 + c_v d_v), \qquad (3)$$

where q_1, q_2, q_3 are quartiles of the study variable, $d_a = q_2 - q_1$ and $d_v = q_3 - q_2$ are auxiliary values, c_a and c_v are constants.

13. Following the standard deviation rule, the outlier is the value of the study variable for which an inequality

$$y_k - y_{k-1} \geq \alpha \hat{s}, k = 2, \dots, n \quad (4)$$

is true. Here y_k is the study variable value in the variation line, $\hat{s}$ is an estimator of standard deviation, and $\alpha$is a constant.

14. Using the testing of hypothesis method, we find out whether the minimum (maximum) value of the study variable is an outlier or not. Let $x_n = \max_i x_i$ and $x_1 = \min_i x_i$. The minimum value is an outlier if T_1 is bigger than the critical value from Student's distribution. The maximum value is an outlier if T_n is bigger than the critical value from Student's distribution (5). The minimum and maximum values are

$$T_1 = \frac{\bar{x} - x_1}{\hat{s}} \text{ and } T_n = \frac{x_n - \bar{x}}{\hat{s}}, \quad (5)$$

where $\bar{x}$ is the estimator of the sample mean, $\hat{s}$ is an estimator of standard deviation.

15. All these four methods are used in the developed SAS Macro for quantitative variables. The user of this program just needs to write from what data set the data must be taken and identify which variable is the identification variable and which is the study variable. As a result, the program gives a data set in which there are the identification variable, the study variable and four indicators, which shows which method identifies the value as an outlier.

III. Imputation

A. Imputation using distributions

16. All variables can be divided into two groups: quantitative and qualitative. For the qualitative variable, the model of discrete random variable can be used, and for the quantitative variable – the model of discrete random or continuous random variable (Figure 1).

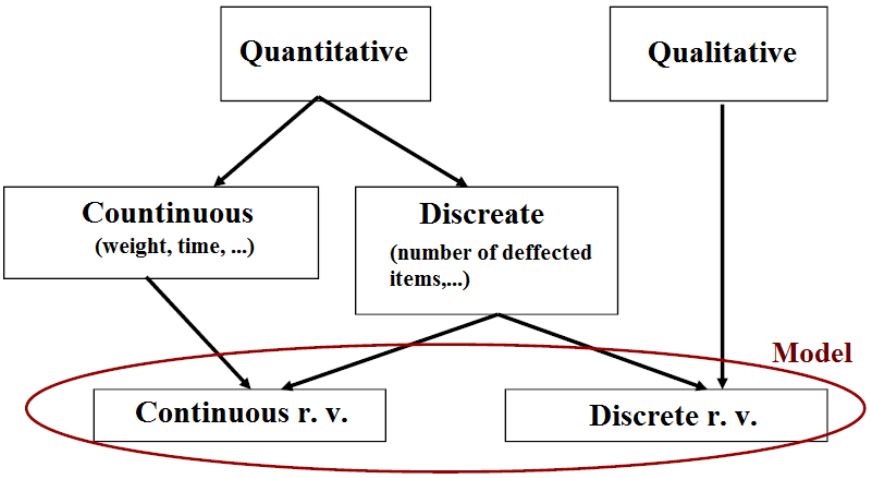

Figure 1. Statistical models

17. There are two models of a discrete random variable and four models of a continuous random variable used in this program. The program chooses the model according to the number of different values of the study variable:

(a) If the study variable has just two different values, the program chooses Bernoulli distribution;

(b) If the study variable has from three to eight different values, the program chooses the distribution of discrete random variable;

(c) If the study variable has more than eight different values, the program chooses one of the continuous distributions.

18. There are four different models for a continuous random variable programed in SAS Marco: uniform, normal, lognormal, and exponential. The user may choose a model himself or let it be chosen by the program. The program chooses the best distribution according to the goodness-of-fit test

$$\chi^2 = \sum_{i=1}^{g} \frac{(obs_i - \exp_i)^2}{\exp_i}, \quad (6)$$

where $i = 1, \dots, g$ is the index of the group, obs_i is the observed value, exp_i is the expected value.

B. Imputation using donors

19. Donor methods are used when the missing value is replaced by another real value from the donor. The donor value may be used from the previous survey or another data source, e.g. administrative data. According to the choice of the donor, there are several different methods:

(a) Historical, or cold-deck, imputation;

(b) Hot-deck imputation;

(c) Nearest neighbor imputation.

20. Historical, or cold-deck, imputation is a statistical procedure that replaces the missing value of an item with a constant value from an external source such as the value from a previous survey. Thus, if the user of the developed SAS Macro wants to use cold-deck imputation, he must have the previous survey data in the same data set as the current one and identify that the donor variable is the same variable, just from the previous survey.

21. Hot-deck imputation replaces missing data with comparable data from the same data set. Thus, if the user of the developed SAS Macro wants to use hot-deck imputation, he must identify that the donor variable is the same variable as the study variable.

22. Nearest neighbor imputation replaces missing data with the donor value. The right donor is found by calculating the distance function from a set of auxiliary information. Thus, if the user of the developed SAS Macro wants to use nearest neighbor imputation, he must identify the study variable, the donor variable, and auxiliary variables from the same data set.

23. At Statistics Lithuania, imputation using donors is commonly used because most of the surveys are repeated from year to year, and in some surveys rotation is used. When this scheme is used, from 25 to 75 per cent of units are the same as in the previous survey. Such type of sample selection gives a good opportunity to use historical, or cold-deck, imputation.

C. Imputation using models

24. The model-based imputation is an approach when the function of auxiliary variables is used. According to the type of study variable (quantitative or qualitative), two models are used:

(a) Regression model for quantitative variables

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad (6)$$

where $i = 1,2, \dots, n$ is the index of the quantitative variable, $k = 1,2, \dots, K$ is the index of the explanatory variable, α is the intercept parameter, $\beta_1, \dots, \beta_k$ are the parameters.

(b) Logistic model for qualitative variables (McFadden, 1974)

$$log\left(\frac{\boldsymbol{P}(Y=j|\boldsymbol{x})}{\boldsymbol{P}(Y=k+1|\boldsymbol{x})}\right) = \alpha_j + \beta_j'\boldsymbol{x}, j = 1,2, \dots, k, \quad (7)$$

where $\alpha_1, \dots, \alpha_k$ are *k* intercept parameters, $\beta_1, \dots, \beta_k$ are *k* vectors of parameters.

25. The parameters of the model are calculated on the basis of available statistical data and can be influenced by the sampling plan. Thus, in this case, the user must identify not only the identification, study and auxiliary variables but also the weight variable in which sample weights are stored.

IV. The use of the developed SAS Macro at Statistics Lithuania

26. In this section, we will shortly present the results of data analysis from different surveys. Some SAS outputs are demonstrated as well.

27. This SAS Macro program was tested not only on simulated data but also on the real data from different surveys. Of course, not all parts of this program are needed for a separate survey. The results of SAS Marco are stored in separate data sets. Moreover, the user can find the main information at Output page (Figure 2). For example, for an *inbound tourism survey*, the detection of outliers and imputation using distributions are used. For a quarterly statistical *survey on short-term statistics on service enterprises*, the detection of outliers and imputation using donors are used.

Example 1. Detection of outliers

28. As an example, the *quarterly statistical survey on short-term statistics on service enterprises* is taken. The results are shown just for one economic activity – *freight transport by road*. The study variable is income in each quarter (PAJ3), and the auxiliary variable is the number of employees.

Study variable PAJ3 has 852 different values:

different_values
852

Four methods are used for finding outliers.
In the table below, it is shown how many methods found the same values as outliers.

Obs	number_of_methods	number_of_values
1	1	5
2	2	40
3	3	1

There are 46 outliers in study variable PAJ3.
They are saved in ISSKIRTYS2 dataset.
In the table below, the first three lines of this data set are shown.

Obs	id	nace	paj3	universal	s_rule	hypothesis	interval
1	17	4941	10681057	1	.	.	1
2	29	4941	10549159	1	.	.	1
3	33	4941	22376552	1	.	.	1

Figure 2. Fragment of results

29. Figure 2 shows a fragment of the results when the user is using a program for the detection of outliers. The first table shows how many different values the study variable has. The second table shows how many methods indicate the same values as an outlier. The third table shows that the first three lines of a data set are outliers. The last four columns show with what method the outliers were detected (indicated as '1').

Example 2. Verification of imputation for quantitative data

31. For different types of imputation, different verification tables are calculated. For example, if the user wants to use imputation using models for a quantitative variable, the verification table shows the percentage difference between the predicted and the real value (Figure 3). According to this information,

predicted values are grouped into four groups: more than 70, between 30 and 70, between 10 and 30, and less than 10 per cent.

VERIFICATION_PAJ3

Obs	veikla	pakl_paj3	COUNT	PERCENT	CUM_FREQ	CUM_PCT
1	4941	1. dideles lab.	10	1.8450	10	1.845
2	4941	2. dideles.	23	4.2435	33	6.089
3	4941	3. vidutines	70	12.9151	103	19.004
4	4941	4. mazos	439	80.9963	542	100.000

In data set PATIKRA_PAJ3, it is shown how good the predicted values are. It shows the percentage difference between the predicted and the real value: more than 70 per cent (very big - dideles lab.), between 30 and 70 per cent (big - dideles), between 10 and 30 per cent (average - vidutines), less than 10 per cent (small - mazos).

Figure 3. Example of a verification table for quantitative variables

32. Figure 3 shows that for the survey from the example above imputation using models is a reasonable choice. The difference between the predicted and the real value is more than 30 per cent (first two groups) in just 6 per cent of data. For the biggest proportion of the data, the difference between the predicted and the real value is less than 10 per cent.

Example 3. Verification of imputation for qualitative data

33. For the last example, simulated data are used. Here the study variable *y4* has just two possible values: 1 and 2. Its value depends on three auxiliary variables: *x1*, *x2*, and *x3*. The results showed that the difference between the real (*real*) and the predicted (*y4*) value exists just for three units (Figure 4). This information is calculated just for those units for which real values are known. Whether to use such a model or not depends on the user's opinion and skills.

VERIFICATION_Y4

Obs	y4	real	predicted	difference_percent
1	1	14	11	-21.4286
2	2	20	23	15.0000

In two groups, the difference between the real and the predicted number of values is bigger than 3 per cent.

Figure 4. Example of a verification table for qualitative variables

IV. Conclusions and future work

34. In this paper, the SAS Macro program for statistical data editing and imputation was presented. SAS Macro program consists of five parts: detection of errors, detection of outliers, imputation using the nearest neighbor method, imputation using models, and imputation using distributions.

35. After this program had been developed, several trainings were organized for the employees of Statistics Lithuania. Interest in this program was sufficient: 37 employees attended the training of this

program. Half of them is using or going to use the SAS Macro in their work. The program was tested using real data. The results showed that time spent for data editing/imputation was reduced.

36. This program might also be used for developing the best way of imputation for a separate survey. It not only gives a new data set with imputed values but also calculates several statistics (sample mean before and after imputation, standard deviation before and after imputation), which can be used to assess the quality of imputation.

37. The latest improvement to this program enables the identification of strata variable. This improvement allows finding errors or outliers and imputing missing values separately in each stratum, group or domain. That is very useful because in many cases the strata are very different and a program without the identification of strata might not find all the errors and outliers or the imputation might be wrong at the domain level.

38. The methods programed now are the simplest one; therefore, later, more complicated methods for the imputation and detection of outliers will be added to the program.

References

Chen J. and Shao J. Nearest neighbour imputation for survey data. Journal of Official Statistics, 16: 113–131, 2000.

Čekanavičius V., Murauskas G. Statistika ir jos taikymai // 1 dalis. TEV, Vilnius, 2000.

Čekanavičius V., Murauskas G. Statistika ir jos taikymai // 2 dalis. TEV, Vilnius, 2002.

Granquist L. Macro-editing. A review of some methods for rationalizing the editing of survey data. http://www.unece.org/stats/publications/editing/SDE1chB.pdf

McFadden, D. Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics, ed. P. Zarembka, New York: Academic Press: 105-42. 1974.

Krapavickaitė, D., Plikusas, A. Imčių teorijos pagrindai. Vilnius: Technika, 2005.

Little R.J.A. and Rubin D. B. Statistical analysis with missing data. Wiley, 1987.

Luzi O., et al. Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. EDIMBUS-RPM, 2007.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf

Nordholt E. S. Imputation: Methods, Simulation Experiments and Practical Examples. International Statistical Review, 66: 157–180, 1998.

Statistical data editing. Methods and techniques. Vol. 1, United Nations, 1994.

Statistical data editing. Impact on data quality. Vol. 3, United Nations, 2006.