Selective editing approach: an application

Rodrigo S. Von Doellinger[1,2], Maysa S. De Magalhães[3,4], and Pedro N. Silva[5]

[1]Coordination of Methods and Quality, Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil

[3,5]National School of Statistical Sciences, Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil

[2,4]Corrresponding authors: Rodrigo S. Von Doellinger, e-mail: vdoellinger@gmail.com and M.S. De Magalhães, e-mail: maysademagalhaes@gmail.com

Abstract

The Household Budget Survey (HBS) 2008/2009 realized by the Brazilian Institute of Geography and Statistics (IBGE), investigated households in urban and rural areas throughout Brazil, from May 2008 to May 2009. One of the main objectives of the survey is to measure the pattern of consumption, expenditure, income and partial assets variation of families. In addition to information regarding the budget structure, other characteristics of households and of their inhabitants are collected, making the potential use of the survey results quite broad. When we deal with variables related to income and expenditure, it is not an uncommon occurrence of cases where the declared income is not commensurate with the expense, or, with the characteristics of the household. Due to the fact that Household Budget Survey is a sample survey, to make inferences with good precision, it is necessary a preliminary analysis of data collected in order to detect the possible presence of outliers, so that they undergo treatment later. In this paper, a method of selective editing was applied to identify outliers and influential observations in the Household Budget Survey (HBS 2008/2009) through the use of the following variables, the monthly household income and the annual household expenditure.

Keywords: consumption, expenditure, household budget survey.

1. Introduction

One of the most important stages of the planning process and execution of a survey is data editing, in which are identified and eliminated erroneous values, called outliers, i e, observations that deviate from a data model (Barnett and Lewis, 1994; Lee, 1995) and do not reflect the reality of the phenomenon been studied. In certain cases, for good quality statistical information be provided, it may not be necessary to identify all the errors presented in the data. It is just sufficient to detect influential observations, that is, those which when included or excluded from the analysis, significantly impact on the estimate of the parameter of interest.

The approach generally used to identify influential observations is called selective editing (Latouche and Berthelot, 1992; Lawrence and McKenzie, 2000). In the methods of selective editing, potentially influential observations are ranked based on values of a score function which expresses the impact of the error in the estimate of parameter of interest. The observations with scores above a pre-set threshold are considered critical and should be revised. The definition of the score function implies in determining the probability of the observation to present error (risk component), as well as the magnitude of the error (component of influence). Risk and influence components are used by score functions presented in the literature (Jader and Norberg, 2005). According to Di Zio et al. (2008) the methods commonly employed to obtain the risk and influence components are based on comparison of the observed values of a given variable and the predicted values for a particular model. The differences between observed and predicted values are used in the calculation of scores for identifying observations that generate greater impact on the estimated of parameter of interest.

Di Zio et al. (2008) proposed a multivariate model to estimate the probability of error and as well as the error magnitude. The method is based on contaminated normal models (Little, 2008). The data observed are described by a mixture of two multivariate normal distributions that represent the erroneous or contamined data and the data without errors. It is assumed that the distribution of the contaminated data can be obtained by the distribution of the data without errors with an increase in the variance (Ghosh-Dastidar and Schafer, 2006).

In this paper, the method of selective editing proposed by Di Zio et. al. (2008) was applied to identify outliers and influential observations in the Household Budget Survey (HBS 2008/2009) of the Brazilian Institute of Geography and Statistics (IBGE) through the use of the following variables, the monthly household income and the annual household expenditure.

2. Household Budget Survey

The Household Budget Survey (HBS 2008/2009), performed by the Brazilian Institute of Geography and Statistics (IBGE), is a household survey carried out by sampling, in every five years, which investigates patterns of consumption, expenditure, income and partial assets variation of families. In addition, several characteristics of households and families are investigated. Through the HBS data, it is possible to draw a profile of the living conditions of the Brazilian population from the analysis of household budgets, to know the goods consumed and services used by households living in urban and rural areas throughout the Brazilian territory, as well as which represents each of these goods and services in the overall expenditure of these families.

In this study, as placed on the introduction the variables of the Household Budget Survey used are the monthly household income and the annual household expenditure. The variable annual household expenditure was conceived as the sum of consumption expenditure related to groups of food, housing, clothing, transportation, hygiene and special care, health care, education, recreation and culture, personal services and miscellaneous expenses. The variable monthly household income comprised the sum of the monthly income of the residents of the housing unit, excluding those of persons under ten years of age and those whose condition in the household is a pensioner, domestic employee or relative of the domestic household employee.

3. The model

In this section, we describe briefly the model proposed by Di Zio et. al. (2008) This model determines the probability for an observation to be in error, that is, to be considered an outlier, and the magnitude of the error. In describing the model proposed by Di Zio et al. (2008), we tried to keep the same notation presented in their article.

A model with contaminated multivariate normal distribution was used to describe the observed data. Observations without errors come from a distribution with smaller covariance and contaminated observations from a distribution with larger covariance. The observed data are represented by the mixture of these two distributions.

Let Z be a n × k matrix of *n* independent realizations of a random vector of dimension *k* that follow a log-normal distribution. Let Y = ln (Z), and let Y * denote the observations without error, where $y^* \sim N(\mu, \Sigma)$. In the absence of error, the data observed are considered to be correct, that is, Y = Y *. Given that the errors have normal distribution, the error mechanism can be described by Y * + ε , where $\varepsilon \sim N\left(0, \Sigma_{\varepsilon}\right)$.

The probability density function of Y / Y* is given by:

$$f\left(y / y^{*}\right) = (1-\pi)\delta\left(y - y^{*}\right) + \pi N\left(y^{*}, \Sigma\right) \tag{1}$$

where π represents the priori probability of contamination and δ (t'-t) is the Dirac delta function with mass in t.

The distribution of the observed data is described by

$$f(y) = (1-\Pi)N(\mu, \Sigma) + \Pi \, \mathrm{N}(\mu, \Sigma_c) \tag{2}$$

where $\Sigma_c = \Sigma + \Sigma_{\varepsilon}$ is the variance of contaminated data.

4. Selective editing

Again, according to Di Zio et al. (2008), for the application of models of contamination in the context of selective editing, it is necessary to determine, from Eq (1), the distribution of Y * given Y or, in the original scale, Z * given Z, which is described by:

$$f\left(z^{*} / z\right) = \tau_1(\ln(z))\delta\left(z^{*} - z\right) + \tau_2(\ln(z))\ln\left(\mu^{*}, \Sigma_{*}\right) \tag{3}$$

where $\Sigma_{*} = \left(\Sigma^{-1} + \Sigma_{\varepsilon}^{-1}\right)^{-1}$, $\boldsymbol{\mu}^{*} = \Sigma^{*}\left(\Sigma_{\varepsilon}^{-1}\mathbf{x} + \Sigma^{-1}\boldsymbol{\mu}\right)$, τ_1 and τ_2 are the posteriori probability of a given observation to belong, respectively, to the correct and erroneous/contaminated data set.

The prediction of the correct values conditioned to the observed values can be obtained for all observations.

$$\hat{z}_{\mathrm{i}} = E\left(z_{\mathrm{i}}^{*} | z_{\mathrm{i}}\right) = \int z_{\mathrm{i}}^{*} f\left(z_{\mathrm{i}}^{*} | z_{\mathrm{i}}\right) d\, z_{\mathrm{i}}^{*} \qquad \mathrm{i} = 1,\ldots,\mathrm{n}$$

Suppose that the estimate of interest is given by the total T_z of the variable Z, that is, $T_z = \sum z_i$. An estimator of T_z is given by $\hat{T}_Z = \sum z_i w_i$, where w_i are the sampling weights tied to each sample unit. The substitution of the observed values by their preditions leads to $\hat{T}_Z^{*} = \sum \hat{z}_i w_i$. Thus, one can calculate the individual relative error (r_i), the predicted error ratio and the estimate of interest, for all *k* variables in the analysis. In our study, *k* = 2. The absolute value of r_i is called local score. The global score for each observation is defined as the higher local score.

After the determination of the global scores, it is initiated the procedure for detecting the influential observations. It must be established a maximum admissible limit (Δ) for the residual error remaining in the data. Then, according to the global scores, the observations are put in a decreasing order. Finally, for each of the *k* variables, it is selected the first *p* observations (which are the observations classified as influential) such that from the observation *p+1* and on, the summation of r_i be always less than Δ.

5. Application to the Household Budget Survey

In this section, the method proposed by Di Zio et al. (2008) was applied to identify outliers and influential observations in the Household Budget Survey (HBS 2008/2009) of the Brazilian Institute of Geography and Statistics (IBGE). For this, we used the library "Selemix" R (R statistical software). As already mentioned, the variables considered in the process of identifying outliers and influential observations were monthly household income and annual household expenditure of 55867 households spread throughout the Brazilian territory.

To ensure the normality of the data, the variables, mensal household income and annual household expenditure were transformed to the logarithmic scale, and were denoted by REND and DESP. Figure 1 shows the scatter plot involving variables REND and DESP. The chart analysis reveals the existence of households whose relationship between income and expenditure is not usual. Households with such behavior are possibly candidates to be labeled as outliers and/or influential observations.

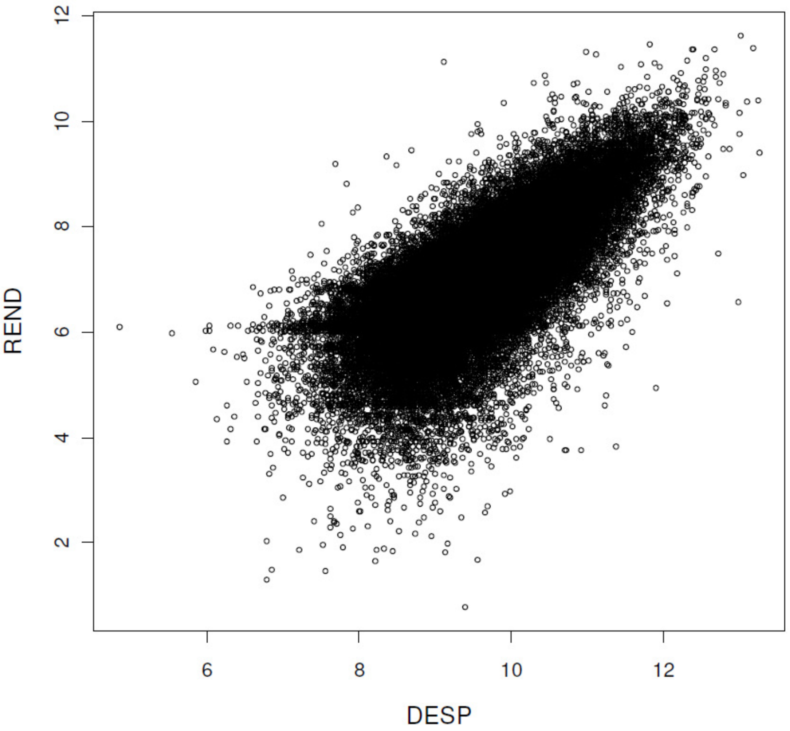

Figure 1- Scatter plot of the variables DESP and REND.

The selective editing procedure requires the definition of a maximum admissible limit (Δ) for the residual error remaining in the data. The amount of influential observations varies depending on Δ. As Δ increases, the amount of observations considered influential diminishes. In this work, the following values for Δ were considered: 0.01, 0.02, 0.03, 0.05 and 0.1

The highest value of Δ (0.1) caused a greatly reduced amount of influential observations, only 495 addresses, corresponding to 0.88% of the total of households. In contrast, the application of the lower limit admissible (0.01) generated 25901 influential observations (46.36%). With respect to the limits having intermediate values, the amount of influential observations varied between 1188 (2.12%) and 7976 (14.27%).

The number of households considered outliers, i.e. households with probability greater than 0.5 of belonging to the contaminated data set was 5965, corresponding to 10.67% of the total of households. We observed that the number of outliers exceeded the amount of influential observations in situations where Δ took values equal to 0.03, 0.05 and 0.1. This highlights the main feature of the process of selective editing, that is, prioritize the most important observations to be revised.

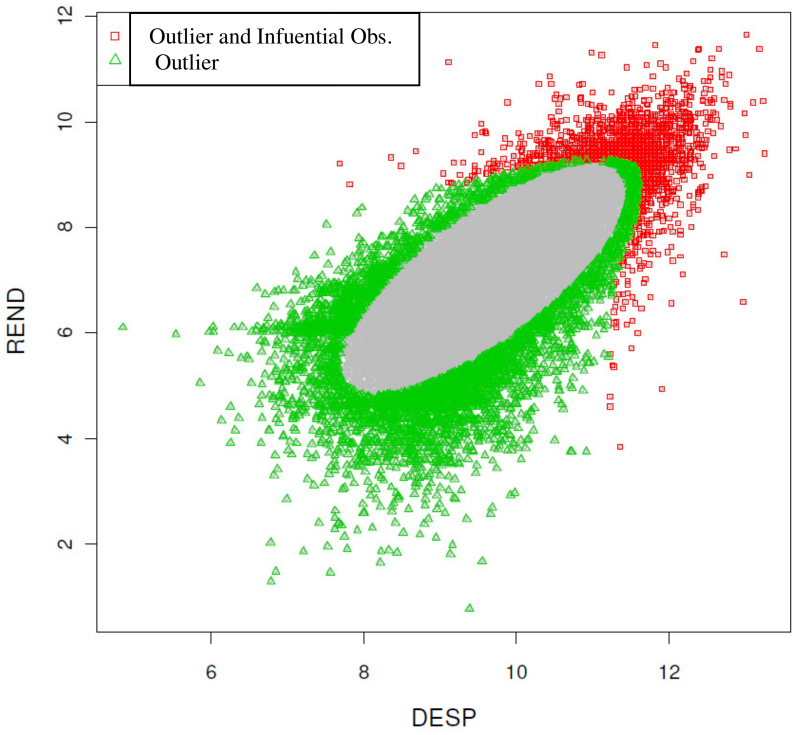

Figure 2- Representation of outliers and influential observations for the maximum admissible limit of 0.03.

The following analysis is related to the results obtained for Δ equal to 0.03, resulting in a total of 1690 influential observations. Figure 2 shows the influential observations (red squares), outliers (red squares and green triangles) and observations without errors in gray. In this situation, all 1690 influential observations were also considered outliers. It can be observed that basically the influential observations correspond to households with income and expenditure far above the standard found in most Brazilian households. These households can affect estimates of interest, such as the total household income and total household expenditure, increasing them significantly. On the other hand, households that had income and/or expenditure far from the usual Brazilian pattern were classified as outliers, independently if the values of these two variables were high or low.

References

Barnett V., Lewis T. (1994). Outliers in Statistical Data, New York: Wiley.

Buglielli M.T., Di Zio M., Guarnera U. (xxxx). Use of contamination models for selective editing. Istat, Italian National Institute of Statistic.

Di Zio M., Guarnera U. (2008). A multiple imputation method for non-Gaussian data. Metron LXVI (1):75–90.

Ghosh-Dastidar M., Schafer J.L. (2003). Multiple edit multiple imputation for multivariate continuous data. J Am Stat Assoc 98:807–817.

Latouche M., Berthelot J.M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. Journal of Official Statistics, 8, n.3, 389- 400.

Lawrence D., McKenzie R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, n. 3, 243-253.

Lee H. (1995). Outliers in Business Surveys, in: Business Survey Methods, Cox B.G., Binder D.A., Chinappa B.N., Christanson A., Colledge M.J. and Kott P.S. (Eds), John Wiley and Sons, Inc. 503-526.

Little, J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, J. R. Stat. Soc., Ser. C, Vol. 37, No. 1, 23-38.

Jäder A., Norberg A. (2005). A Selective Editing Method considering both suspicion and potential impact, developed and applied to the Swedish Foreign Trade Statistics, UN/ECE Work Session on Statistical Data Editing, Ottawa (http://www.unece.org/stats/documents/2005.05.sde.htm).