Investigating Outliers Detection Methods for the Iranian Manufacturing Establishment Survey Data

Zahra Rezaei Ghahroodi [1,3], Taban Baghfalaki[2] and Mojtaba Ganjali[2]
[1] Statistical Research and Training Center, Tehran, Iran.
[2] Department of Statistics, Shahid Beheshti University, Tehran, Iran.
[3] Corresponding author: Zahra Rezaei Ghahroodi, e-mail: z_rezaee@sbu.ac.ir

Abstracts

The role and importance of the industrial sector in the economic development specify the necessity of having accurate and timely data for exact planning. As outliers data in establishment surveys are common due to the structure of the economy, the evaluation of survey data by identifying and investigating outliers prior to the release of data is necessary. In this paper the practical application of different robust multivariate outlier detection methods based on the Mahalanobis distance with BACON algorithm, minimum volume ellipsoid (MVE) estimator, minimum covariance determinant (MCD) estimator, Stahel-Donoho estimator is presented. Also some univariate outlier detection methods such as Hadi and Simonoff (1993) method, using some regression models, are presented. These methods are illustrated using a real data set on Iranian Manufacturing Establishment Survey (IMES). These data are collected each year by the Statistical Center of Iran using sampling weights. In this paper it is demonstrated that the use of different robust outlier detection methods (multivariate and univariate), in a number of manufacturing industries, leads to the same results.

Keywords: robust multivariate outlier detection; sampling weight; Winsorization; Mahalanobis distance

1. Introduction

Manufacturing sector is one of the principle components in the Economic Development Plans. To assess and realize the goals determined for the manufacturing sector, availability of updated and accurate statistics is very essential. Like all statistical surveys, manufacturing establishment survey is subject to measurement errors, including sample and non-sample errors. These measurement errors affect the accuracy of the published statistics.

As outliers are a common phenomenon which is present in every data set in any application domain such as establishment surveys, identification and correction of outliers are an important objective of survey processing which carried out by statistical centers. Many researchers in work with establishment sample surveys often encounter observations that differ substantially from most of the observations in the sample. This increases the possibility of anomalous data and makes their detection more difficult. Outliers are so unlike or divergent values from the rest of data that ignoring them can lead to inaccurate survey estimates. Outliers can be recorded due to errors in the data capturing process or they may be valid. The former data, outliers which are identified as errors, are non-representative (one which can be regarded to be unique in the population) and the latter, valid values, refer to representative outliers (one which can not be regarded to be unique in the population) (Chambers, 1986).

A common class of such errors is errors in writing out the response, misunderstanding of type of unit (e.g. thousands of ponds instead of single pounds) or misunderstanding of the question, which results in an erroneous response. The standard approach for solving these kinds of problems is to use a large number of edits during survey processing. However, sometimes these data couldn't be identified. Sometimes a correct response can be an outlier. The causes of outliers in this situation can be related to the method of choosing samples or because of large change in reported values due to a time lag between the time when samples drawn and the time when these samples are used in a further investigation.

There are different methods for outlier detection. One of the classifications of outlier detection methods is the division of methods to univariate approach or multivariate approach. Another fundamental classification of outlier detection is to have parametric or nonparametric methods. One of the non-parametric methods is distance-based methods. A classical way of identifying multivariate outlier base on distance method is Mahalanobis distance. In order to avoid the masking effect, robust estimates of location and scatter estimates of the data set are considered. Many methods suppose that the data follow some elliptical distribution and try to estimate robustly the center and the covariance matrix. Then, they use a corresponding Mahalanobis distance to detect outliers. There is a large literature on outlier detection. Many methods for the detection of multiple outliers use very robust methods to split the data into a clean part and the potential outliers. For example in multivariate data, Rousseeuw and van Zomeren (1990) proposed to find the subset of observations within a minimum volume ellipsoid (MVE) as non-outliers data. In 1999, Rousseeuw and van Driessen proposed finding the subset of observations with the minimum covariance determinant (MCD). Another option is the forward search method introduced by Hadi (1992a) and Hadi and Simonoff (1993). The basic idea of this method is to identify a clean subset of the data, defined from a robust method, and then includes clean observations until only the outlying units remain out. This method rapidly leads to the detection of multiple outliers. All multiple outlier detection methods suffer from a computational cost that escalated rapidly with the sample size. Billor et. al. (2000) proposed a new general approach titled as BACON (Blocked Adaptive Computationally efficient Outlier Nominators) algorithm, based on Hadi (1992b) and Hadi and Simonoff (1993), which can be computed quickly regardless of the sample size. Beguin and Hulliger (2008) proposed the BACON-EEM algorithm for multivariate outlier detection in incomplete survey data.

Since Iranian Manufacturing Establishment Survey (IMES) data set like all statistical surveys is subject to measurement errors and these measurement errors affect the accuracy of the published statistics, in this paper we present an application of outlier

detection methods, mentioned above, to IMES data collected by the Statistical Center of Iran (SCI). The paper is organized as follows. The description of the IEMS data is given in Section 2. Section 3 gives some description of outlier detection methods and presents results, using these methods, and compares their performance. In the end, some concluding remarks and recommendations are given.

2. Description of IEMS

In order to identify the industrial structure of the country, to provide information needed for planning on industrial development, to assess the results of these plans, and to formulate the proper economic policies; the Statistical Centre of Iran has implemented the survey on Manufacturing Establishments from 1972. It is obvious that annual data collection is done after finalizing the financial accounts of manufacturing establishments; thus, in this survey, the data of preceding year is collected.

The country's first manufacturing survey was launched in 1963 by the former General Department of Public Statistics, for which the Ministry of Economics was an immediate replacement in 1964 to 1972. It kept on performing the job up to 1973 when the Ministry of Industries and Mines took the duty over. The Statistical Centre of Iran (SCI) launched the first survey of Large Manufacturing Establishments (with over 10 or more workers) in 1972, which has annually repeated. In 1997 and 2002, the SCI conducted the General Census of Manufacturing and Mines (GCMM) and the General Census of Establishments, respectively, to collect a frame data set for the nation's economic activities and household's activities.

The target population for this survey is all manufacturing businesses operating in Iran. The objective of this survey is to collect economic data required for compiling National Accounts and in details is to estimate input value, output value and value added. In this survey which is undertaken by the SCI, the information is collected directly based on face-to-face interview with officer or director (statistical units) of manufacturing establishment. Sampling frame is list of manufacturing establishments obtained from General Census of Establishments in 2002 and is annually updated. Survey method is complete enumeration of large manufacturing establishments with 50 or more workers and sample survey for other establishments with 10-49 workers. It should be mentioned that in the survey, the data on manufacturing establishments with 10-49 workers for some provinces were collected by sampling method and the data related to the remaining provinces as well as manufacturing establishments with 50 and more workers was collected through a census. The survey population is split into ISIC (Iranian Standard Industry Classification) industries and stratified according

to size. The sampling method is stratified random sampling in which the stratification variables are number of workers and economic activity based on ISIC 4-digit Codes. The industry classification utilized in the survey is the ISIC, Rev. 3 with some changes. In each stratum, sample establishments are selected using systematic method. In IEMS, questionnaire is sent out to approximately 12500 manufacturing establishments during Jun and December every year.

3. Methodology and Results

In our study, first of all, we apply the presented methods on one of the ISIC 4-digit Codes (2710) related to the primary production of iron and steel stick. The number of manufacturing establishments in this code for the data collected in 2010 is 194 units. Some of the interested variables for identifying outliers are input, output, total number of employees, salaries and wages, non-industrial payment, energy and water costs or expenses and Input-output (I/O).

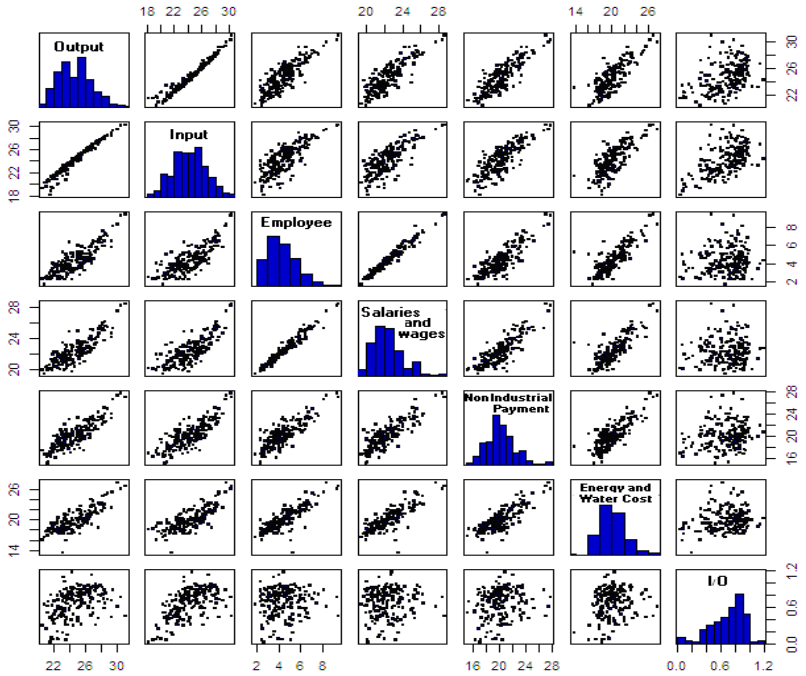

Figure 1: Scatter plot of the logarithm of the variables and variable I / O with their marginal histogram

Since the variation of the response variables are high then we should consider a transformation on it. One of the methods of choosing suitable transformation is using the Box-Cox method (1964). After drawing the Box-Cox plot of interested response variables, the logarithm was chosen for all above-introduced variables except I/O. Scatter plot and histogram for the logarithm of the variables and for I / O are shown in Figure 1.

Given the linear structure evident in Figures 1 between output and other interested variables, we first applied the univariate forward search algorithm described in Section

3 to these data. In this method an appropriate model will be fitted to the basic subset. We consider the following linear model

$$Log(output) = \beta_0 \log(RM) + \beta_1 \log(SW) + \beta_2 \log(EW) + \varepsilon$$

Where ε has normal distribution, $\beta = (\beta_0, \beta_1, \beta_2)$ is the vector of regression coefficients for the logarithm of output as response, RM is the amount of raw material, SW is salaries and wage and EW is energy and water cost.

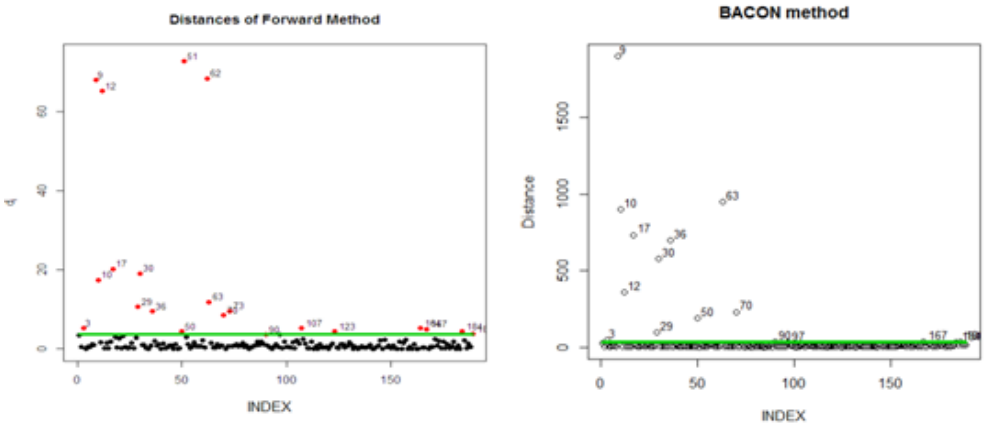

Figure 2: The IMES Data: The index plot of distances of forward method and BACON method

Figure 2 includes index plots of distances of forward method and BACON method. These plots detect values higher than the cut-off point (green lines) as outliers.

Most of the multivariate statistical methods are based on estimates of multivariate location and covariance; therefore these estimates play a central role in the framework. We will start with computing the robust minimum covariance determinant (MCD) estimate for the IMES data. Figure 3 shows the Distance-Distance plot introduced by Rousseeuw and van Zomeren (1990) which plots the robust distances versus the classical Mahalanobis distances and allows classifying the observations and identifying the potential outliers. This Figure (not illustrated here) is the same for different robust multivariate methods such as MVE, and Stahel-Donoho method.

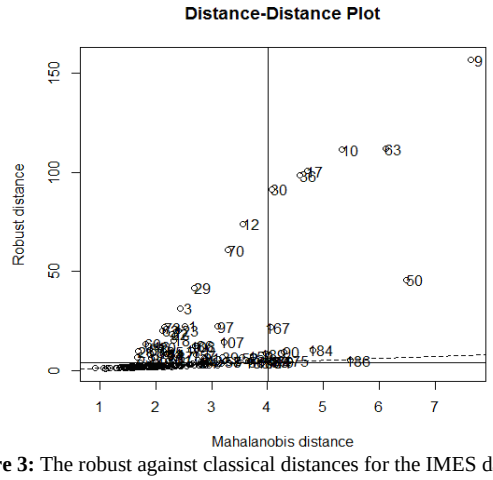

Figure 3: The robust against classical distances for the IMES data set

Table 1 compares various outlier detection methods and identified the number of detected outliers. According to surveys, and performing different methods, some of the establishments which are identified as outliers in different methods are common.

Index of establishment detected as outlier	Methods of identifying outliers
1,**3**,**9**,**10**,**12**,**17**,18,22,**29**,**30**,**36**,**50**,52,62,**63**,**70**,73,86,90,97,107,123,159,175,184,186	**MCD**
3,**9**,**10**,**12**,**17**,**29**,**30**,**36**,**50**,**63**,**70**,97,107,167,175,184,186	**Stahel-Donoho**
3,**9**,**10**,**12**,**17**,**29**,**30**,**36**,**50**,**63**,**70**,90,97,167,184,186	**BACON**
1,**3**,**9**,**10**,**12**,**17**,**29**,**30**,**36**,**50**,60,62,**63**,**70**,73,88,90,97,105,107,108,123,157,167,171,189	**MVE**
3,**9**,**10**,**12**,**17**,**29**,**30**,**36**,**50**,51,62,**63**,**70**,73,90,107,123,164,167,184,189	**Hadi's Forward Method**

Table1: Compare the index of detected outliers by various methods (bold indices are detected by all methods)

5. Conclusions

In this paper we compare different robust multivariate outlier detection methods and a univariate outlier detection method, based on Hadi and Simonoff (1993) approach, for outlier detection on the real data set of IMES. According to the results, most of the outlier detection methods have found the same most notifying observations. However, the BACON method (which is also used in Canada and can be computed quickly regardless of the sample size) may be preferred to be used as a multivariate outlier detection method.

References

Béguin C, Hulliger B (2008) The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Surv. Methodol., 34(1):91-103.

Billor, N., Hadi, A.S. and Vellemann, P.F.(2000). BACON: Blocked Adaptative Computationally -efficient Outlier Nominators. Computational Statistics and Data Analysis, 34(3), 279-298.

Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B. 26 (2): 211-252.

Chambers, R. (1986). Outlier robust finite population estimation. Journal of the American Statistical Association, 81(396), 1063-1069.

Hadi, A.S., (1992a). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society Series (B), 54 (3), 761-771.

Hadi, A.S., (1992b). A new measure of overall potential influence in linear regression. Computational Statistics and Data Analysis, 14, 1-27.

Hadi, A.S., Simonof, J.S., (1993). Procedures for the identification of multiple outliers in linear models. Journal of the American Statistical Association, 88, 1264-1272.

Rousseeuw PJ, Van Driessen K (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41, 212-223.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association, 85, 411, 633–651.