American Journal of Applied Mathematics and Statistics, 2013, Vol. 1, No. 1, 6-10
Available online at http://pubs.sciepub.com/ajams/1/1/2
DOI:10.12691/ajams-1-1-2

Robust Goodness of Fit Test Based on the Forward Search

Abbas Mahdavi*

Department of Statistics, Faculty of Mathematical Sciences, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran
*Corresponding author: a.mahdavi@vru.ac.ir

Received December 30, 2012; Revised February 04, 2013; Accepted February 28, 2013

Abstract The most frequency used goodness of fit tests are based on measuring the distance between the theoretical distribution function and the empirical distribution function (EDF), but presence of outliers influences these tests strongly. In this study, we propose a simple robust method for goodness of fit test by using the "Forward Search" (FS) method. The FS method is a powerful general method for identifying outliers and their effects on the hypothesized model. The performance and the ability of the procedure to capture the structure of data, even in the presence of outliers, are illustrated by some simulation studies and real data examples.

Keywords: *forward search procedure, goodness of fit test, robust approach, outlier*

1. Introduction

Model testing and verification are important part of statistical inference to obtain information about the population from which the sample is drawn. Most of the parametric tests require a number of assumptions such as normality. These assumptions should be validated before we go advance with other aspects of statistical inference. Goodness of fit test measures the degree of agreement between the distribution of an observed sample data and a hypothetical statistical distribution. In practice, it often happens that an assumption such normality holds approximately in majority of observations, but some observations follow a different pattern or no pattern at all. Such atypical data are called outliers. A single outlier can even have a large disturbing effect on a classical statistical method that is optimal under the classical assumptions.

The Forward Search (FS) approach is a powerful general method that provides diagnostic plots for finding outliers and determining their effect on the fitted models. The FS method starts from a small, robustly chosen subset of the data and increase the subset size until finally all the data are fitted. The outliers enter the model in the last steps and the entrance point of the outliers can be exposed by monitoring some statistics of interest during the process. Initially Hadi [1] and Atkinson [2] presented the method of fitting a model to subsets of an increasing size for multivariate data analysis. Hadi and Simonoff [3] used the FS in regression, and the development of the FS was introduced by Atkinson and Riani [4,5,6] and Atkinson et al. [7] for regression and multivariate procedure with a recent discussion in [8]. Recently the FS method is implemented in wide applications, e.g. ANOVA framework [9], testing normality [10], finding an unknown number of multivariate outliers [11], detecting atypical observations in financial data [12], robust estimation of efficient mean–variance frontiers [13] and benchmark testing of algorithms for very robust regression [14].

The purpose of this article is to adopt the FS method in the goodness of fit test for continues variable. The most frequently used goodness of fit tests are based on measuring the distance between the theoretical distribution function and the empirical distribution function (EDF) such as Kolmogorov-Smirnov [15], Cramer [16], Anderson-Darling [17], but presence of outliers influences these tests strongly. In this paper we try to determine how many and which observations agree with the null hypothesis distribution. In order to adapt the FS for goodness of fit test, we need a way to select an outlier free subset and a test for goodness of fit to be used in the search.

The paper is organized as follows. Section 2 presents the proposed forward search algorithm in goodness of fit test. In Section 3, the performance of the method is illustrated with simulated data and the behavior of our procedure is analyzed. Finally in Section 4 we show an application of the proposed method to real world data by using of the blood clotting data set. Concluding remarks are provided in Section 5.

2. Forward Search in Goodness of Fit Test

Many goodness of tests for testing hypotheses about specified distribution are available in literature, some of them are special purpose tests, and they are appropriate and perform well only for some special situations. Anderson-Darling test gives more attention to the tails of distribution, hence it can be useful for investigated the effects of outliers on the goodness of fit test, since the outliers have low probability when they originate from the same statistical distribution as the other observations. For further results about Anderson-Darling test see [17].

In this paper we use FS method not only to detecting outliers, but also for investigating the effect of outliers on the Anderson-Darling goodness of fit test. The FS method has three steps: the first step is choosing outlier free subset of all observations, the second step presents the plan to progressing in FS and the last step is monitoring statistics during the search. In the following subsections we adopt these three points separately.

2.1. Step 1: Choice of the Initial Subset

Starting point of the FS procedure is choosing outlier free subset of observations robustly. If the vector of ordered observations $\mathbf{x}_{(.)}=\left(x_{(1)},x_{(2)},\ldots,x_{(n)}\right)$ comes from a distribution function $F_0(x)$ with known parameters, then it can be shown that the probability distribution functions (pdf) of the random variables

$$Z_{(r)}=F_0\left(x_{(r)}\right),\quad r=1,2,\ldots,n. \tag{1}$$

are given by $Z_{(r)}\sim Beta(r,n-r+1)$, so its mean is

$$\alpha_r=E\left(Z_{(r)}\right)=\frac{r}{n+1}. \tag{2}$$

It is possible to use of transformations of $F_0(\alpha_i)$ to estimate the value of $x_{(i)}$. It means that $\hat{x}_{(i)}=F_0^{-1}(\alpha_i)$, where the evaluation of quantile function may be involves numerical methods and it is done for common distributions in many statistical package. Therefore we obtain the vector $\mathbf{x}_{(.)}=\left(x_{(1)},x_{(2)},\ldots,x_{(n)}\right)$ of estimated expected values of the ordered sample. Now the elements of $\mathbf{x}_{(.)}$ are reordered according their absolute residuals from estimated values $\hat{\mathbf{x}}_{(.)}$, denoted by $\mathbf{x}_{(\mathrm{T})}$.

First the size of initial subset must be specified. A larger initial subset will give more stable estimates and smoother forward plots. Also, it is irrational that more than half of the data be outliers, hence we set the size of initial subset $m_0=\lfloor n/2\rfloor$. The initial subset can be achieve by choosing the first $\lfloor n/2\rfloor$ observations of the $\mathbf{x}_{(\mathrm{T})}$ that are closer to the their estimated values $\hat{\mathbf{x}}_{(.)}$.

2.2. Step 2: Adding Observations during the FS

After choosing initial subset, in the $n-\lfloor n/2\rfloor$ remaining steps all observations must be add to it. At each step, the observations that are closer to the previously fitted model are added to the subset. The parameters of $F_0(x)$ are completely known and we try to find the largest subset of observations that can be distributed from $F_0(x)$, it is not necessary to reorder the observations $\mathbf{x}_{(.)}$ at each step of the search. Therefore we add the next observation of $\mathbf{x}_{(\mathrm{T})}$ to the previously chosen subset in each step of the search.

2.3. Step 3: Monitoring the Search

For detecting and determining the effect of outliers, some interested statistics of interest must be monitor during the search. Let $S^{(m)}$ be the subset of the first m observations of $\mathbf{x}_{(\mathrm{T})}$. The collections of $\mathbf{A}^2_{\mathrm{FS}}$ statistics of the subset $\left\{S^{(m)};m=\lfloor n/2\rfloor,\ldots,n\right\}$ during the FS procedure defined as

$$\mathbf{A}^2_{\mathrm{FS}}=\left(A^2_{S^{(m_0)}},\ldots,A^2_{S^{(m)}},\ldots,A^2_{S^{(n)}}\right), \tag{3}$$

where $A^2_{S^{(m)}}$ denote the value of Anderson-Darling test statistic for the subset $S^{(m)}$ in testing the null hypothesis $H_0:F_{(x)}=F_0(x)$ against the alternative hypothesis $H_1:F_{(x)}\neq F_0(x)$, where $F_0(x)$ is completely known and defines hypothesized distribution function to be test. To obtain the corresponding rejection region of the (3) during the search, the empirical quantiles of this statistic must be estimated by simulation study in each step of the search and for different sample sizes.

3. Simulation Study

For evaluate the proposed statistic (3), we conduct simulation studies that aim to consider the behavior of this statistic in the presence of outliers and ability of FS for detecting them. Eight samples are considered which are generated in the following way:

- Sample A: 100 observations are generated from a standard normal distribution.
- Sample B: 95 observations are generated from a standard normal distribution and for contamination 5 observations are generated from a $N\left(\mu=5,\sigma^2=1\right)$.
- Sample C: 100 observations are generated from an $Exp\left(\lambda=1\right)$.
- Sample D: 95 observations are generated from an $Exp\left(\lambda=1\right)$ and for contamination 5 observations are generated from a $Uniform\left(a=10,b=11\right)$.
- Sample E: 100 observations are generated from a $Beta\left(a=2,b=5\right)$.
- Sample F: 95 observations are generated from a $Beta\left(a=2,b=5\right)$ and for contamination 5 observations are generated from a $Uniform\left(a=0.9,b=1\right)$.
- Sample G: 100 observations are generated from a $Weibull\left(sh=5,sc=2\right)$.
- Sample H: 95 observations are generated from a $Weibull\left(sh=5,sc=2\right)$ and for contamination 5 observations are generated from a $Uniform\left(a=3,b=4\right)$.

For each sample, we are interested in testing the null hypothesis that majority of observations comes from it. In Figure 1, values of $\mathbf{A}_{FS}^2$ during the search are plotted for samples A-H and compared with corresponding 95% quantile of its distribution obtained from a simulation study with clean data. The null hypothesis is accepted in each step of the search for clean samples A, C, E and G, and it's rejected after entrance of outliers in the last steps for contaminated samples B, D, F and H.

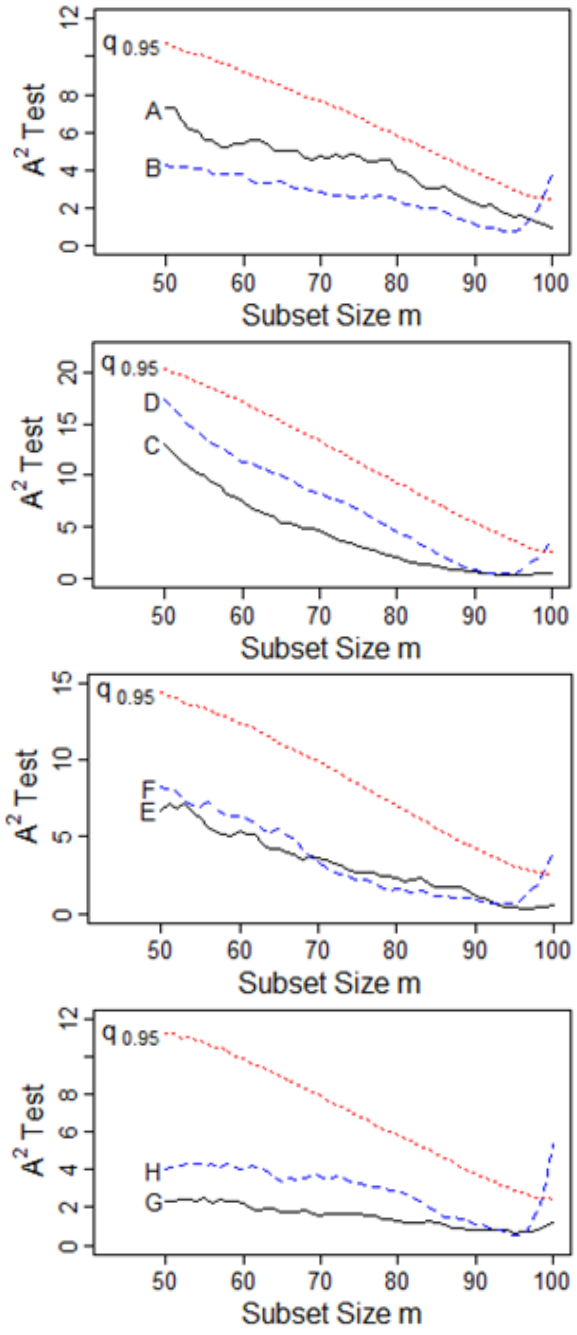

Figure 1. Forward plots of $\mathbf{A}_{FS}^2$ during the search for samples A-H

3.1 Empirical Power of $\mathbf{A}_{FS}^2$

In this subsection, we interest to evaluate the empirical power of our approach. Two following examples are considered:

Example 1: Consider the null hypothesis $H_0 : F(x) = N(0,1)$ against $H_1 : F(x) \neq N(0,1)$. Figure 2 shows the empirical power of $\mathbf{A}_{FS}^2$ against following alternative hypotheses by generating 10000 samples of size 100.

(a) $N\left(\mu = 1, \sigma^2 = 1\right)$

(b) $N\left(\mu = 0, \sigma^2 = 0.5\right)$

(c) $N\left(\mu = 1, \sigma^2 = 0.5\right)$

(d) $t_{(df=1)}$

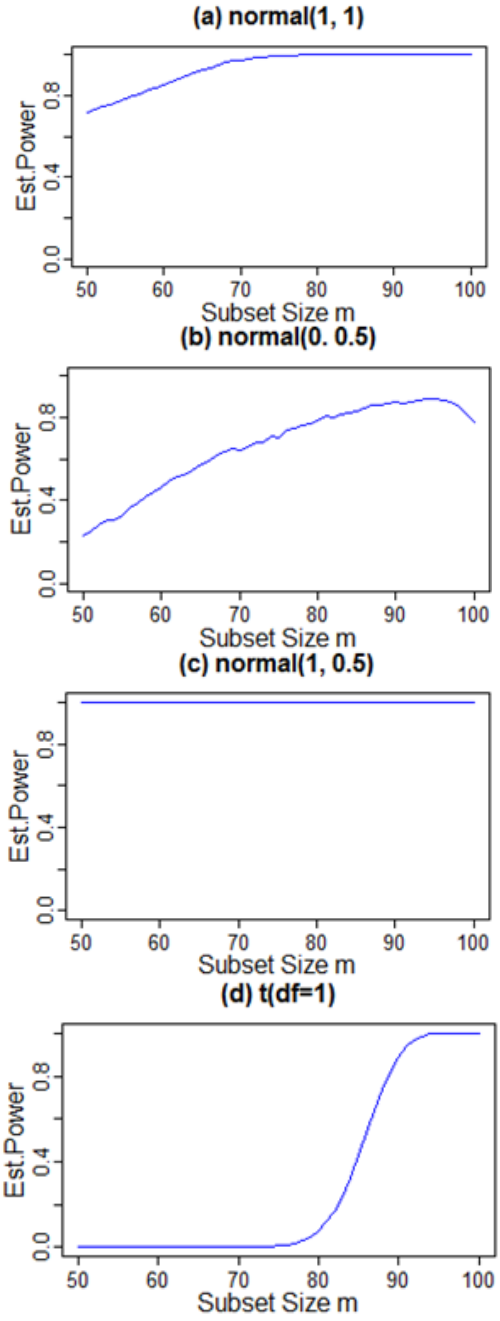

Figure 2. Empirical power of $\mathbf{A}_{FS}^2$ for example 1 versus alternative distributions (a-d)

Example 2: Here, we analysed the estimated power of testing the null hypothesis $H_0 : F(x) = Exp(1)$

against $H_1 : F(x) \neq Exp(1)$, by regarding the following alternative distributions, the results are shown in Figure 3.

(e) $Exp(\lambda = 0.5)$

(f) $Gamma(r = 2, \lambda = 1)$

(g) $Gamma(r = 2, \lambda = 0.5)$

(h) $Weibull(sh = 2, sc = 1)$

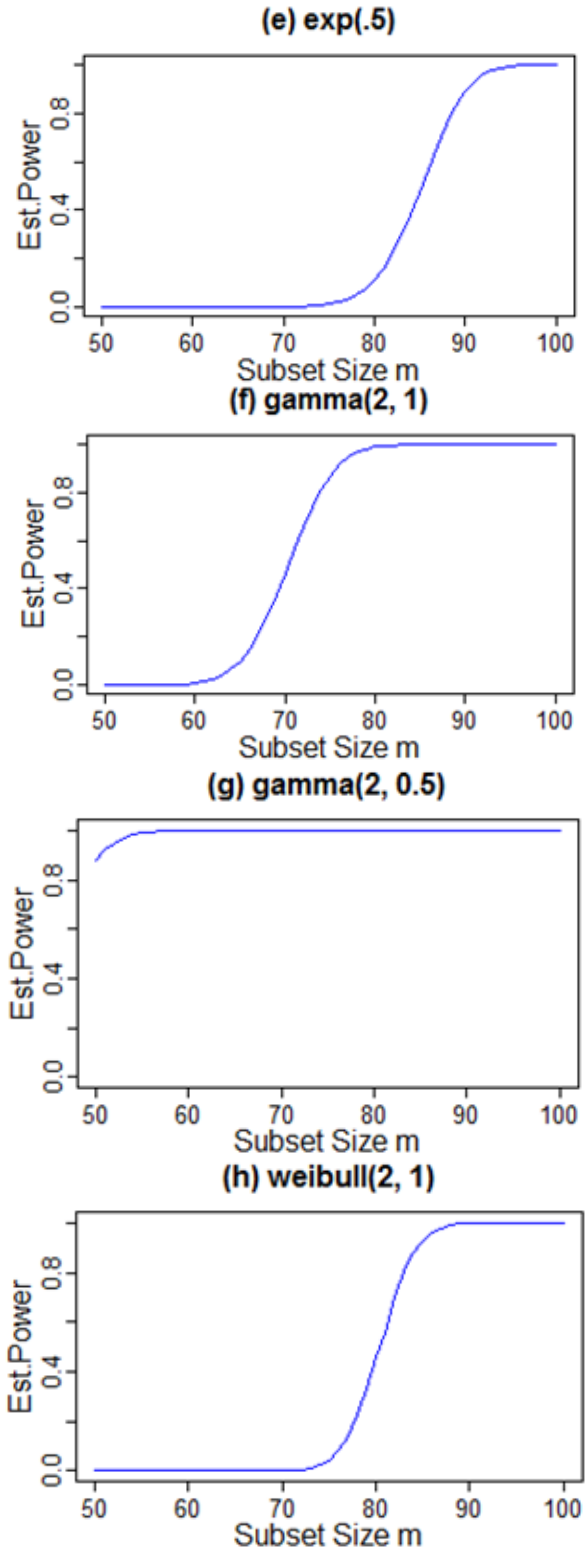

Figure 3. Empirical power of $\mathbf{A}^2_{\mathrm{FS}}$ for example 2 versus alternative distributions (e-h)

After analysis the two mentioned examples, we concluded that the power of our proposed procedure in first steps of the search is low for close alternatives. It is due to our aim is to find the largest subset of observations that can be distributed as the null hypothesis and for close alternatives this can occurs in the first steps of the search. Therefore, the larger sample size provides safer procedure to detect and investigate the effect of outliers.

4. The Blood Clotting Data

To investigate the performance of the proposed approach to real-world data, we use the blood clotting dataset relating to blood clotting activity (PCA) is measured for 158 Norway rats before (baseline) and four days after injection of an anticoagulant, published by Heiberg [18]. This data set contains 91 instances for female gender, for our purposes, we use the blood clotting activity at baseline (PCA0) for the female gender.

The histogram and boxplot for PCA0, plotted in Figure 4, show some observations that could be regard as outliers. To test normality assumption for these observations, we use the robust M-estimator [19] to estimate parameters, that are $\hat{\mu}_1 = 73.16$ and $\hat{\sigma}_1 = 16.33$. Therefore, we perform Test 1 as: the null hypothesis $H_0 : F_{(x)} = N(\mu_1, \sigma_1)$ against $H_1 : F_{(x)} \neq N(\mu_1, \sigma_1)$ and the results of forward search plotted in Figure 5 (a).

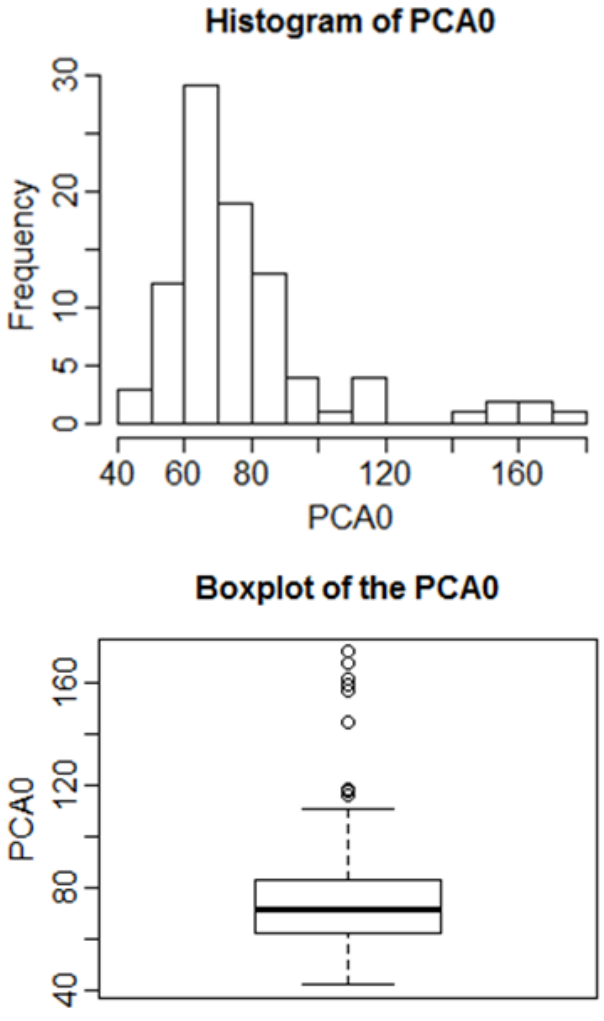

Figure 4. Histogram and boxplot for PCA0

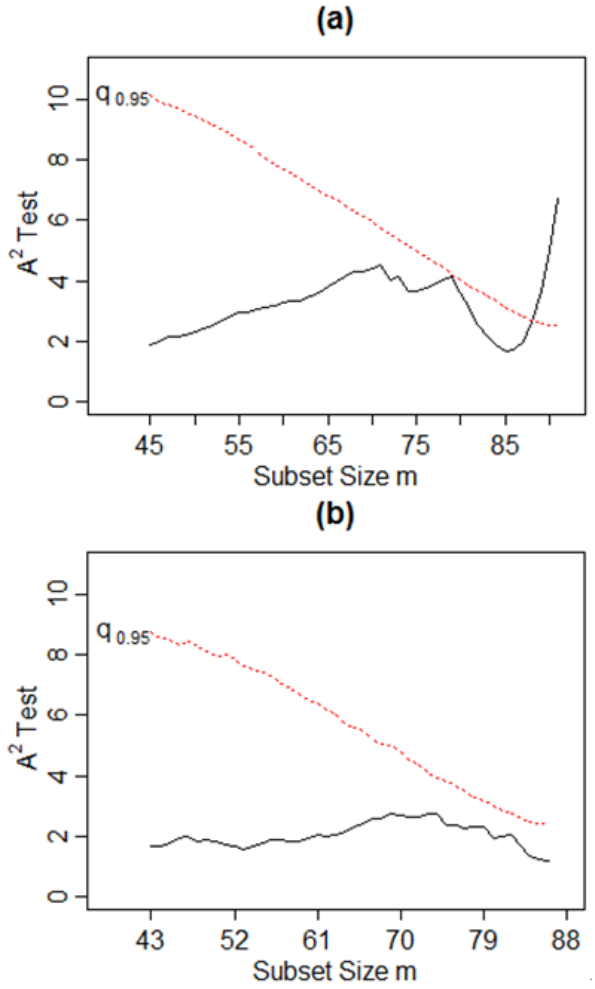

Figure 5. (a) Forward search results for Test 1 and (b) Forward search results for Test 2

The null hypothesis of Test 1 accepted in each step except the last steps due to entrance of outliers after steps 85 onwards, indicating that 6 observations are important outliers. After removing these outliers, the maximum likelihood estimate of parameters are $\hat{\mu}_2 = 72.25$ and $\hat{\sigma}_2 = 15.73$. Now we have the estimated parameters that are more optimal respect to M-estimator with total observations and hence we define Test 2 to verify the distribution of the remained observations as follow: the null hypothesis $H_0 : F_{(x)} = N(\mu_1, \sigma_1)$ against $H_1 : F_{(x)} \neq N(\mu_1, \sigma_1)$. Figure 5 (b) shows the ability of our proposed approach to detecting outliers and finding the largest subset of observations that verify the null hypothesis.

5. Concluding Remarks

In this paper, a new robust approach to the goodness of fit test for continuous distributions has been presented. Our work concerns the effect of outliers in goodness of fit and the robust FS method is implemented to individuate the outliers. At every step of the FS, the proposed robust statistics are computed and this search method divides the group of outliers from the other observations by a graphical approach. Hence, it is able to determine whether the majority of data is distributed as the null hypothesis distribution. In order to illustrate the application and the advantage of the FS approach we conducted simulation studies. Furthermore, we showed an application of the proposed approach to real data.

References

[1] Hadi, A. S., Identifying multiple outliers in multivariate data, *Journal of the Royal Statistical Society, Series B*, 54. 761-771. 1992.

[2] Atkinson, A. C., Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, 89. 1329-1339. 1994.

[3] Hadi, A. S., Simonoff, J. S., Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, 88. 1264-1272. 1993.

[4] Atkinson, A. C., Riani, M., *Robust Diagnostic Regression Analysis*, Springer, New York, 2000.

[5] Atkinson, A. C., Riani, M., Forward search added-variable t-tests and the effect of masked outliers on model selection, *Biometrika*, 89(4). 939-946. 2002.

[6] Atkinson, A. C., Riani, M., The Forward search and data visualization, *Computational Statistics*, 19. 29-54. 2004.

[7] Atkinson, A. C., Riani, M., Cerioli, A., *Exploring Multivariate Data with the Forward Search*, Springer, New York, 2004.

[8] Atkinson, A. C., Riani, M., Cerioli, A., The forward search: theory and data analysis, *Journal of the Korean Statistical Society*, 39. 117-134. 2010.

[9] Bertaccini, B., Varriale, R., Robust Analysis of Variance: an approach based on the Forward, *Computational statistics and data analysis*, 51. 5172-5183. 2007.

[10] Coin, D., Testing normality in the presence of outliers, *Statistical Methods & Applications*, 17. 3-12. 2008.

[11] Riani, M., Atkinson, A.C., Cerioli, A., Finding an unknown number of multivariate outliers, *Journal of the Royal Statistical Society Series*, B 71. 447-466. 2009.

[12] Bellini, T., Detecting atypical observations in financial data: the forward search for elliptical copulas, *Advances in Data Analysis and Classification*, 4. 287-299. 2010.

[13] Grossi, L., Laurini, F., Robust estimation of efficient mean-variance frontiers, *Advances in Data Analysis and Classification*, 5. 3-22. 2011.

[14] Torti, F., Perrotta, D., Atkinson, A.C., Riani, M., Benchmark testing of algorithms for very robust regression: FS, LMS and LTS, *Computational statistics and data analysis*, 56. 2501-2512. 2012.

[15] Kolmogorov, A. N., Sulla Determinazione Empirica di Una Legge di Distribuzione, *Giornale dell'Istituto Italiano degli Attuari*, 4. 83-91. 1933.

[16] Cramér, H., On the composition of elementary errors, *Scandinavian Actuarial Journal*, 11. 141-180. 1928.

[17] Anderson, T. W., Darling, D. A., A Test of Goodness of Fit, *Journal of the American Statistical Association*, 49. 765-769. 1954.

[18] Heiberg, A-C., Project at The Royal Veterinary and Agricultural University, 1999.

[19] Huber, P.J., *Robust Statistics*, John Wiley & Sons, New York, 1981.