Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2018; 4(1): 30-5.

Original Article

Robust correlation coefficient goodness-of-fit test for the Gumbel distribution

Abbas Mahdavi[1*]

[1] Department of Statistics, School of Mathematical Sciences, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

ARTICLE INFO

Received 11.09.2017
Revised 12.11.2017
Accepted 02.12.2017
Published 02.01.2018

Key words:
Outlier;
Regression analysis;
Statistical distributions

ABSTRACT

Background & Aim: A single outlier can even have a large disturbing effect on a classical statistical method that is optimal under the classical assumptions. One of the powerful goodness-of-fit tests is the correlation coefficient test, however this test suffers from the presence of outliers.

Methods & Materials: This study provides a simple robust method for test of goodness of fit for the Gumbel distribution [extreme value distribution (EVD) type I family] through using the new diagnostic tool called the "Forward Search" (FS) method. The FS version of this test was introduce in the present study, which is not affected by the outliers.

Results: A new robust method for testing the goodness-of-fit for Gumbel distribution has been presented. The approach gives information about the distribution of majority of the data and the percentage of contamination.

Conclusion: A new robust method for testing the goodness-of-fit for the Gumbel distribution has been presented. The simple and fast method have been used to find distribution of proposed statistic. In addition, using the transformation study, an application to the two-parameter Weibull distribution has been investigated. The performance and the ability of this procedure to capture the structure of data have been illustrated by some simulation studies.

Introduction

The extreme value distributions (EVDs) are widely used in risk management, finance, insurance, economics, hydrology, material sciences, telecommunications, and many other industries dealing with extreme events. The EVD arises from the Fisher-Tippett limit theorem (1) on extreme values or maxima in sample data. Let X_1, X_2, …, X_n be independent and identically distributed (IID) random variables and $M_n=\max(X_1,X_2,\ldots,X_n)$. If there exist constants $a_n > 0$ and $b_n \in$, and some non-degenerate distribution function G, such that, $a_n^{-1}(M_n - b_n) \rightarrow G$ then, G belongs to one of the three types of EVDs: Fréchet, Weibull, and Gumbel. These can be grouped into the generalized EVD. In this study, it has been tried to propose a robust goodness-of-fit test for Gumbel (Type I EVD) distribution.

The assumptions of such models should be validated before progressing with other aspects of statistical inference. In practice, it often happens that such assumptions hold approximately in majority of observations, however 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. One of the basic tools useful for this purpose is the correlation coefficient goodness-of-fit test based on

* Corresponding Author: Abbas Mahdavi, Postal Address: Department of Statistics, Faculty of Mathematical Sciences, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
Email: a.mahdavi@vru.ac.ir

Please cite this article in press as: Mahdavi A. Robust correlation coefficient goodness-of-fit test for the Gumbel distribution. *J Biostat Epidemiol. 2018; 4(1): 30-5*

quantile-quantile (QQ) plot.

The objective in this study was to adopt the Forward Search (FS) method in the goodness of Gumbel distribution. Correlation coefficient test is one of the powerful tests introduced by Filliben (2) to test normality. Kinnison (3) assessed the goodness of fit of EVD type-I based on Filliben's correlation coefficient test and examined its power properties for various alternative models. The correlation coefficient statistic is not a robust statistic and hence, the presence of outliers influences this test strongly. The test involves computing the correlation coefficient between the ranked data and the expected value of the order statistic with the same rank. In this study, it has been tried to determine how many and which observations agree with the hypothesis of Gumbel distribution and also as an application for testing the goodness of fit for Weibull distribution.

The FS approach is a powerful general method providing diagnostic plots for finding outliers and discovering their underlying effects on models fitted to the data and for assessing the adequacy of the model. Riani and Atkinson (4) and Atkinson and Riani (5, 6) developed the FS procedure for regression modeling and multivariate analysis frameworks. The FS method starts from a small, robustly chosen subset of the data. The method increases the subset size using some measures of closeness from fitted model 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 revealed through monitoring some statistics of interest during the process. Recently, the FS method has been implemented in wide applications, for example, analysis of variance (ANOVA) framework (7), testing normality (8), testing the parameters of a normal population (9), and density estimation of a unimodal continuous distribution (10). The study be Atkinson et al. (11) can be referred to for further results.

Methods

Correlation coefficient goodness-of-fit test for Gumbel distribution: The correlation coefficient test was introduced by Filliben (2) to test goodness of fit for the normal distribution. Kinnison (3) assessed the goodness of fit of EVD type-I based on Filliben's correlation coefficient test and examined its power properties for various alternative models. A QQ plot is a common and basic technique used for finding a suitable data model. When comparing an observed data to a hypothesized distribution, the plot of the ordered observations versus the appropriate quantiles of assumed distribution should look approximately linear and hence the product moment correlation coefficient (PMCC), which measures the degree of linear association between two random variables, is an appropriate test statistic.

The correlation coefficient goodness-of-fit test for the Gumbel distribution is built as follows. Let X be a random variable from the Type I family distribution.

$$F(x) = \exp\left(-exp\left[\frac{x-\mu}{\sigma}\right]\right); \quad -\infty < x < \infty, \qquad (1)$$

Where, μ and σ are unknown location and scale parameters, respectively. In such location-scale model, there is a simple relationship between the p-quantiles of X and W=(X-μ)/σis the standard Gumbel variable (μ = 0, σ = 1). The p-quantile of X, defined by $P(X \leq x_p) = p$, is

$$x_p + \sigma\log(-\log(p)) \qquad (2)$$

Thus, x_p is a linear function of $w_p = \log(-\log(p))$, the p-quantile of W.

Let $(X_{(1)},X_{(2)},\ldots,X_{(n)})$ be an ordered sample of size n from X, for appropriate p_i;i=1,2,…n (the plotting positions), x_{pi} can be approximated by the i-th order sample $X_{(i)}$ Thus the correlation coefficient statistic, R, for goodness-of-fit test is defined as the correlation between ordered sample $X_{(i)}$ and the p_i-quantile of W, W_{pi}.

Many plotting positions have been proposed in the literature; in this study, the median rank was used due to its robustness property and was therefore used in the case of skewed distributions, like the EVD (12, 13). The median rank, $m_{(i)}$, of the i-th order statistics is given by

$$m_{(i)} = b_{0.5,i,n-1+1}, \qquad (3)$$

Where, $b_{0.5,\alpha,\beta}$ is the median of the beta distribution with parameters α and β.

The distribution of R can be estimated by means of Monte Carlo simulation for different sample sizes. The hypothesized distribution (1) is rejected if the observed value of R is smaller than the critical value.

As an application, in order to use the correlation coefficient test for testing the validity of two-parameter Weibull distribution to the data with cumulative distribution function (CDF):

$$f(x;\alpha,\beta) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right];$$
$$x > 0, \alpha > 0, \beta > 0. \quad (4)$$

The two-parameter Weibull distribution can be transformed to the family of two-parameter Gumbel distribution using a logarithmic transformation. It is necessary to transform into a Gumbel distribution with location parameter $\mu = \log(\beta)$ and scale parameter $\sigma = 1/\alpha$ by taking the logarithm of the data.

FS in correlation coefficient goodness-of-fit test for the Gumbel distribution: Let $x_{(.)} = (x_{(1)}, x_{(2)},\ldots,x_{(n)})$. The vector of ordered observations comes from a Gumbel distribution (1), then it is possible to write

$$x_{(i)} = \mu + \sigma w_{mi} + \varepsilon_i, \quad (5)$$

Where, $w_{mi} = \log(-\log(_{mi}))$, and m_i is the median rank defined in (3).

In this section, the FS method introduced by Atkinson and Riani (5) was used to analyze the behavior of regression model. The FS method is a powerful approach not only to detect outliers, but also to investigate their effects on the estimation of parameters and on aspects of inference about models. The basic idea of the FS approach was to order the observations by their proximity to the fitted model. The FS method was made up of the following three main steps: the starting point was to find the appropriate starting subset of observations, the second step presented the plan to progress in FS, and the last step was to monitor some suitable quantities during the search. In the following subsections, how these three points are performed will be described.

Step 1: Choice of the initial subset: Starting point of the FS procedure was to choose outlier free subset of observations robustly. To start the FS approach, the size of initial subset had to be specified. This size could be as small as p = 2. Therefore, the search was performed over subsets of P observations to find the best subset of observations. The initial subset could be achieved by the use of robust regression estimator least median of squares (LMS) regression estimator proposed by Rousseeuw (14).

Step 2: Progressing in the search: At each step of the search, the procedure added to the subset the observation that was closest to the previously fitted model. Let $S^{(k)}$ be a subset of size k, the FS moves to $S^{(k+1)}$ in the following way: after the least square regression model is fitted to the $S^{(k)}$ subset, all observations are ordered according to their square residuals, now the k + 1 observations are chosen with the smallest square residuals. This procedure is repeated until all observations are entered into the model.

Step 3: Monitoring the search: For detecting and determining the effect of outliers, some statistics of interest must be monitored during the search. The FS version of correlation coefficient, R_{FS}, is defined as a collection of R (correlation coefficient) statistics computed for different subsets of $X_{(.)}$ and corresponding units of w_{mi}. Let $w_{(k)}$ $w^{(k)}$ be the units of w_{mi} corresponding to the subset $S^{(k)}$, then the R_{FS} is defined as $R_{FS} = (R_{S^{(p)},w^{(p)}}, \ldots, R_{S^{(k)},w^{(p)}}, \ldots, R_{S^{(n)},w^{(n)}}$ (6).

The empirical quantiles of (6) during the search can be estimated by simulation in each step of the search. In any step of the search, the acceptance region lies between the value of the chosen quantile and 1.

Results

Simulation study: In order to evaluate the proposed procedure, simulation studies were conducted with the aim to consider the behavior of statistic (6) in the presence of outliers and ability of FS to detect them. Consider 6 samples were generated in the following way:

- Sample A: 100 observations were generated from a standard Gumbel distribution.
- Sample B: 95 observations were generated from a standard Gumbel distribution and for contamination, 5 observations were generated from a $N(\mu = 3, \sigma = 1)$.

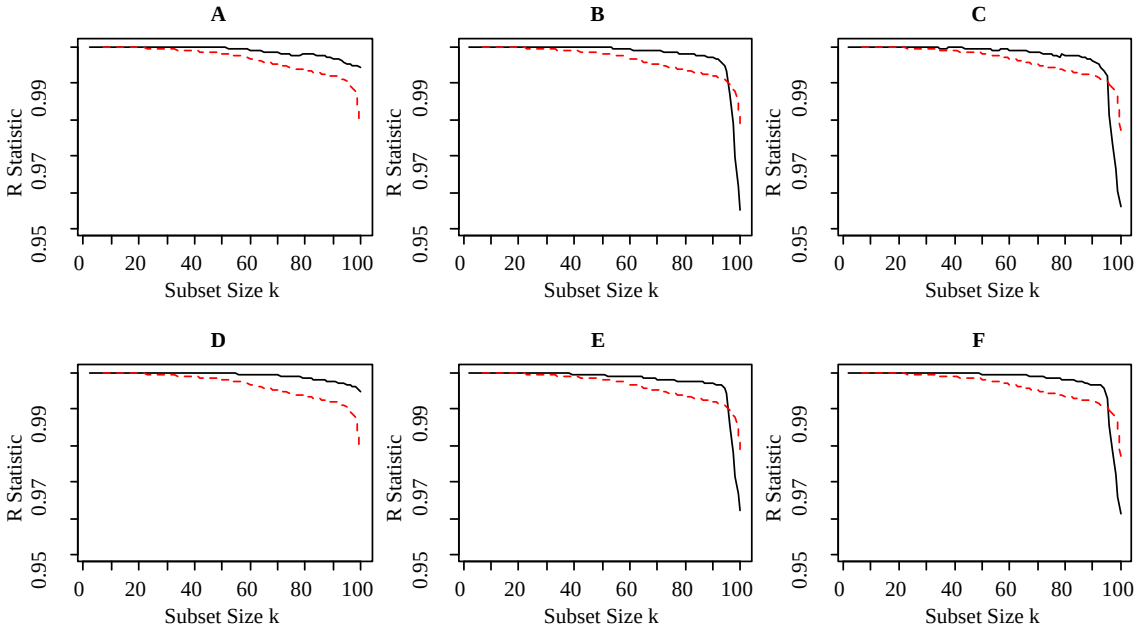

Figure 1. Forward plots of R $_{\text{forward search}}$ (R_{FS}) during the search for samples A-F.

• Sample C: 95 observations were generated from a standard Gumbel distribution and for contamination, 5 observations were generated from a $\text{Uniform}(a = 3, b = 4)$.

• Sample D: 100 observations were generated from a $\text{Weibull}(\alpha = 2, \beta = 2)$.

• Sample E: 95 observations were generated from a $\text{Weibull}(\alpha = 2, \beta = 2)$ and for contamination, 5 observations were generated from a $N(\mu = 10, \sigma = 1)$.

Sample F: 95 observations were generated from a $\text{Weibull}(sh = 2, sc = 2)$ and for contamination, 5 observations were generated from a $\text{Uniform}(a = 9, b = 10)$.

In Figure 1, the values of R_{FS} during the search have been plotted for samples A-F and compared with estimated corresponding 5% quantile (dashed line) of its distribution obtained from ordered observations method discussed in the next section. The null hypothesis of Gumbel distribution was accepted in each step of the search for clean sample A and it was rejected after entrance of outliers in the last steps (step 96 onwards) for contaminated samples B and C, indicating 5 observations were outliers. Moreover, the same results for samples D-F were summarized as follows: the null hypothesis of two-parameter Weibull distribution was accepted in any step of the search for clean sample D and it was rejected from step 96 onwards for samples E and F.

Discussion

The empirical null distribution of (6) can be found by simulating numerous samples generated from a standard Gumbel distribution. Since the FS is a reiterative algorithm, this way of estimating distribution is very time consuming. Atkinson and Riani (15) proposed the method of ordered observations to estimate the distribution of outlier test statistic. In the following subsection, this simple and fast method will be described briefly.

Method of ordered observations: The FS orders all observations in each steps of the search. In the absence of outliers, when moving from $S^{(k)}$ to $S^{(k+1)}$, most of the time, only one new observation joins the subset and this ordering does not change much during the search. Hence, the observations can be ordered only once according to square residuals resulting from the chosen initial subset, denoted by $X_{(ord)}$. In the step k of the search, only the first k observations of $X_{(ord)}$ are chosen.

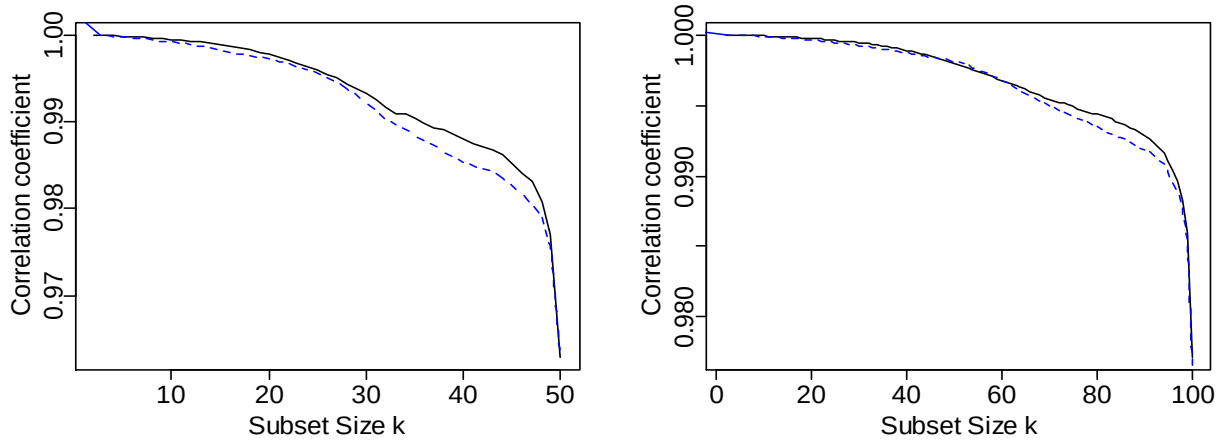

Figure 2. 5% bounds of the empirical distribution (continuous lines) and the estimated distribution using the ordered observations method (dashed lines) for sample sizes n = 50 (left panel) and n = 100 (right panel)

Figure 2 shows the 5% bounds of the empirical distribution and the estimated distribution using the ordered observations method for sample sizes n = 50 and n = 100.

The analysis of figure 2 indicates that the method of ordered observations approximate the 5% quantile of (6) very well except the middle of the search and by increasing the sample size, this approximation was improved. To specify the acceptance region, the lower bounds of (6) are required and hence, only the 5% quantile of (6) was curved in figure 2.

Conclusion

In this study, a new robust method has been presented to test the goodness-of-fit for Gumbel distribution. The approach provides information on the distribution of majority of the data and the percentage of contamination. At every step of the FS, the proposed statistic was computed and a cut-off point divided the group of outliers from the other observations with a graphical approach. In order to illustrate the application and the advantage of the FS approach, some artificial examples were used. In addition, the simple and fast method was used to find distribution of the proposed statistic. Furthermore, an application of the proposed approach to the goodness-of-fit test was shown for the two-parameter Weibull distribution.

Conflict of Interests

Authors have no conflict of interests.

Acknowledgments

This study was supported by the research council of Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.

References

1. Fisher RA, Caleb Tippett LH. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical Proceedings of the Cambridge Philosophical Society 1928; 24(2): 180-90.
2. Filliben JJ. The probability plot correlation coefficient test for normality. Technometrics 1975; 17(1): 111-7.
3. Kinnison R. Correlation coefficient goodness-of-fit test for the extreme-value distribution. Am Stat 1989; 43(2): 98-100.
4. Riani M, Atkinson AC. Robust diagnostic data analysis: Transformations in regression. Technometrics 2000; 42(4): 384-94.
5. Atkinson AC, Riani M. Forward search added-variable t-tests and the effect of masked outliers on model selection. Biometrika 2002; 89(4): 939-46.

6. Atkinson AC, Riani M. The forward search and data visualisation. Comput Stat 2004; 19(1): 29-54.
7. Bertaccini B, Varriale R. Robust analysis of variance: An approach based on the forward search. Comput Stat Data Anal 2006; 51(10): 5172-83.
8. Coin D. Statistical methods and applications. Stat Methods Appt 2008; 17(1): 3-12.
9. Mahdavi A, Towhidi M. Robust tests for testing the parameters of a normal population. Journal of Sciences, Islamic Republic of Iran 2014; 25(3): 273-80.
10. Mahdavi A, Towhidi M. Density estimation of a unimodal continuous distribution in the presence of outliers. Iranian Journal of Science and Technology, Transactions A: Science 2017; 1-12.
11. Atkinson AC, Riani M, Cerioli A. The forward search: Theory and data analysis. J Korean Stat Soc 2010; 39(2): 117-34.
12. D'Agostino RB. Goodness-of-fit-techniques. Boca Raton, FL: CRC Press; 1986.
13. Castillo E, Hadi AS, Balakrishnan N, Sarabia JM. Extreme value and related models with applications in engineering and science. Hoboken, NJ: Wiley; 2004.
14. Rousseeuw PJ. Least median of squares regression. J Am Stat Assoc 1984; 79(388): 871-80.
15. Atkinson AC, Riani M. Distribution theory and simulations for tests of outliers in regression. J Comput Graph Stat 2006; 15(2): 460-76.