Distribution based outlier detection with the extremevalues package

Mark P.J. van der Loo[1]

1. Statistics Netherlands, PO box 24500, 1490 HA The Hague, the Netherlands
Contact: m.vanderloo@cbs.nl

Keywords: Economic data, outliers, QQ-plot, distribusion-based outlier detection

Outlier detection is performed by statistical agencies, such as Statistics Netherlands, to identify observations that either contain errors or have to be treated differently in the estimation process.

The **extremevalues** package is an implementation of the outlier detection methods described in van der Loo (2010), which were recently developed to detect outliers in economic data. The method can be applied when an approximate data distribution is known. For example, in the reference it is shown that certain types of economic data distributions (Value Added Tax turnover values) often resemble a lognormal distribution.

Distribution based outlier detection takes advantage of this knowledge in two steps: first, the distribution's parameters are determined robustly, by fitting a (possibly transformed) subset of the data to the QQ-plot positions for the model distribution. Second, a test is performed to deceide whether the smallest or largest values are outliers. Our method involves two test options: the first option computes a value above (below) which less than ρ (say 1.0) observations are expected, given the sample size N. The second option tests the hypothesis that a small (large) value can be drawn from the model distribution using its fit residual as a test statistic.

The extremevalues package currently supports outlier detection, assuming the normal, lognormal, Pareto, exponential or Weibull distribution as a model. It also includes a number of plotting facilities which can be used to graphically analyze the outlier detection results. See Figure 1 for an example.

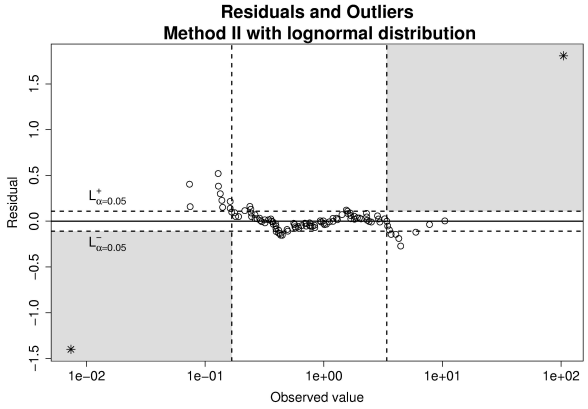

Figure 1: Results of outlier detection on a simulated dataset, using the second test method. Outliers (indicated with a $*$) are detected using the fit residuals as a test statistic. The horizontal lines indicate $\alpha = 0.05$ probability levels, assuming normally distributed residuals. Points between the vertical dotted lines were used in the fit.

References

M.P.J. van der Loo (2010). Distribution based outlier detection for univariate data, Discussion paper 10xxxx *Statistics Netherlands, the Hague, (in press)*
`http://www.cbs.nl/en-GB/menu/methoden/onderzoek-methoden/discussionpapers/archief/2010/default.htm`.

M.P.J. van der Loo (2010) **extremevalues**: a package for outlier detection in unvariate data, R package version 2.0
`http://cran.r-project.org`,`http://www.markvanderloo.eu`