Distribution based outlier detection in univariate data

Mark P.J. van der Loo

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (10003)

Statistics Netherlands

The Hague/Heerlen, 2010

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2008–2009	= 2008 to 2009 inclusive
2008/2009	= average of 2008 up to and including 2009
2008/’09	= crop year, financial year, school year etc. beginning in 2008 and ending in 2009
2006/’07–2008/’09	= crop year, financial year, etc. 2006/’07 to 2008/’09 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008310003 X-10

Distribution based outlier detection in univariate data

Mark P. J. van der Loo

Summary:
Two univariate outlier detection methods are introduced. In both methods, the distribution of the bulk of observed data is approximated by regression of the observed values on their estimated QQ plot positions using a model cumulative distribution function. Having obtained a description of the bulk distribution, we give two methods to determine if extreme observations are designated as outliers. In Method I, we determine the value above which less than a certain number of observations (say 0.5) are expected, given the total number of observations and the fitted distribution. In Method II, we devise a test statistic to determine whether an extreme value can be drawn from the same distribution as the bulk data. Both methods have been implemented in the "extremevalues" R package which has been made available via the CRAN web archive. An outlier detection method based Method I using the lognormal distribution has been implemented for the Structural Business Statistics at Statistics Netherlands.

Keywords:
Outliers, Extreme values, Test Statistic

1 Introduction

The detection and handling of outliers, either in sampled or administrative numerical data is an important part of many estimation processes. An outlier can indicate an observation or processing error, or a special element of the population which needs to be treated differently from the bulk in the estimation process.

One problem encountered in outlier detection is that many data which are analysed in practice are skewly distributed, even after a reasonable stratification. Well known examples include economic data, such as turnover values or number of employees per business establishment. Economic data such as turnovers are often spread over several orders of magnitude, which makes the identification of outliers more difficult. As an illustration, consider the distribution of Value Added Tax turnover values in Figure 2, Section 3. The standard box-and-whisker plots of the $\log_{10}$ of turnover values seem to give a reasonable outlier detection on the right side of the distribution, since only a few isolated points lie above the top whiskers. At the left side of the distribution however, the number of outliers seems unreasonable, at least to the eye. Also, it is easy to see that the number of outliers depends here on the chosen data transformation. Namely, in Figure 2, where a logarithmic transformation is applied, a number of 2, 3, 2, 1 and 0 outliers are found in various size classes. If in stead a square root transformation would be used, one would find 6, 7, 5, 5 and 0 outliers respectively. Ideally, to identify outliers, one would like to use a transformation based on the actual distribution of the bulk.

Barnett and Lewis (1994) define an outlier in a set of data as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". Although this seems intuitively appealing, it is not a useful definition since it leaves the identification of an observation as an outlier a matter of subjectivity. Here, we adhere to the view that outliers are generated by a different distribution than the bulk observations. Hawkins (1980) labels this outlier generation *mechanism ii*, as opposed to *mechanism i*, where all observations are thought to be generated by a single distribution. Consequently our method is an operationalisation of this definition.

Various outlier detection methods (often called discordancy tests) have been reported in literature. Many of them are based on a test statistic which takes into account the distance of an observation to a location parameter and the spread of the sample. Examples include Dixon-type statistics (Dixon, 1950, 1953) and Grubbs' statistic (Grubbs, 1950).

In the methods proposed here, the distribution of the bulk of observations is estimated robustly by a suitable model distribution. Outliers are then defined as observations which are unlikely to be generated by the bulk distribution (with an explicit definition of the "degree of unlikelyness"). After obtaining a robust estimate for the bulk distribution, we devise two test statistics. The first is the untransformed observed value, the second is the residual of regression used in the robust estimation procedure. The main advantage of this method is that it gives very robust results once the right distribution is found. In this paper, it is also shown how the correctness of the model distribution can be assessed. The outlier detection methods described here have been implemented as an R package (van der Loo, 2010) which is available from the CRAN website (R Development Core Team, 2008).

The rest of this paper is organized as follows: In section 2 the theory of the method is explained in detail, and an illustration based on artificial data is given. In section 3 the method is applied to Dutch Value Added Tax data and the results are analysed in more detail. Section 4 summarizes our conclusions.

2 Theory

Throughout this paper a real random variable Y is assumed, of which N realisations y_i have been obtained and ordered so that $y_1 \leq y_2 \leq \ldots \leq y_N$.

2.1 Parameter estimation by regression on QQ plot positions

For the purpose of outlier detection, assume that the observations y_i are generated by a model probability density, with cumulative density function (cdf) $F(Y|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters specifying F. The value of $\boldsymbol{\theta}$ can be estimated robustly from the bulk of the observations by minimizing the sum of squares:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i\in\Lambda} [g(y_i) - g(F^{-1}(\hat{F}_i|\boldsymbol{\theta}))]^2, \tag{1}$$

where Λ indexes a subset of the observations y_i and g is a monotonic function, differentiable on the range of Y. Here, we use

$$\Lambda = \{i \in \{1, 2, \ldots, N\} \,|\, F_{\mathsf{min}} \leq \hat{F}_i \leq F_{\mathsf{max}}\}, \tag{2}$$

where $0 \leq F_{\min} < F_{\max} \leq 1$ are to be determined by the user and $\hat{F}_i$ are plot positions as used in quantile-quantile (QQ) plots, based on the sorted observations. The above equation is based on the notion that the plot positions $\hat{F}_i$ can be considered estimates of the cumulative probability value at a given y_i. The plotting positions can be calculated as [see Makkonen (2008)]

$$\hat{F}_i = \frac{i}{N+1}. \tag{3}$$

The method was implemented for the exponential, Weibull, lognormal, Pareto and normal distribution. With the exception of the exponential distribution, solving Eq. (1) with a suitable transformation g yields linear regression equations of the form

$$\mathbf{b} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{x}, \tag{4}$$

where $\mathbf{b}$ is a 2-dimensional vector containing functions of the distribution parameters, $\mathbf{A}$ is a $|\Lambda| \times 2$ matrix containing functions of $\hat{F}_i$ and $\mathbf{x}$ is a $|\Lambda|$-dimensional vector containing functions of y_i. Table 1 shows the result for all five distributions.

Table 1. Results of solving Eq. (1) for various model distributions. The first and second column give the name and variable range of the models. The third column shows the function g from Eq. (1). The fourth column shows the form of the cdf of the model and the subsequent column shows the interpretation of the symbols in Eq. 4 for various models. We use the notations $\hat{\mathbf{F}} = (\hat{F}_i | i \in \Lambda)$, $\mathbf{y} = (y_i | i \in \Lambda)$, $||\mathbf{x}||$ *for the Euclidean vector norm of* $\mathbf{x}$ *and* $\cdot$ *for the standard inner product. All functions work elementwise on vectors,* Id *is the identity function,* $\ln$ *the natural logarithm and* erf *the error function as defined in Abramowitz and Stegun (1972).*

Model	range	g	cdf	$\mathbf{b}$	$\mathbf{A}$	$\mathbf{x}$
Normal	$\|y\| < \infty$	Id	$\frac{1}{2} + \frac{1}{2}\mathrm{erf}\{(y-\mu)^2/\sqrt{2}\sigma)\}$	$(\hat{\mu}, \hat{\sigma})'$	$[\mathbf{1}, \sqrt{2}\mathrm{erf}^{-1}(2\hat{\mathbf{F}} - \mathbf{1})]$	$\mathbf{y}$
Lognrm.	$y > 0$	$\ln$	$\frac{1}{2} + \frac{1}{2}\mathrm{erf}\{(\ln y-\mu)^2/\sqrt{2}\sigma)\}$	$(\hat{\mu}, \hat{\sigma})'$	$[\mathbf{1}, \sqrt{2}\mathrm{erf}^{-1}(2\hat{\mathbf{F}} - \mathbf{1})]$	$\ln \mathbf{y}$
Weibull	$y \geq 0$	$\ln$	$1 - \exp\{-(y/\lambda)^k\}$	$(\ln\hat{\lambda}, \hat{k}^{-1})'$	$[\mathbf{1}, \ln\ln(\mathbf{1} - \hat{\mathbf{F}})^{-1}]$	$\ln \mathbf{y}$
Pareto	$y \geq y_m$	$\ln$	$1 - (\frac{y_m}{y})^\alpha$	$(\ln\hat{\lambda}, -\widehat{\alpha^{-1}})'$	$[\mathbf{1}, \ln(\mathbf{1} - \hat{\mathbf{F}})]$	$\ln \mathbf{y}$
Exp.	$y \geq 0$	Id	$1 - \exp\{-\lambda y\}$	$\hat{\lambda} = -\|\|\ln(\mathbf{1} - \hat{\mathbf{F}})\|\|^2/\ln(\mathbf{1} - \hat{\mathbf{F}})' \cdot \mathbf{y}$		

Table 2. Calculation of the outlier detection limit for Method I. Note that $\delta_\mp = 1 - \delta_\pm$.

F	$\ell_\rho^\pm$
Normal	$\sqrt{2}\sigma\mathrm{erf}^{-1}[\pm(1-2\rho_\pm/N)]+\mu$
Lognrm.	$\exp\left\{\sqrt{2}\sigma\mathrm{erf}^{-1}[\pm(1-2\rho_\pm/N)]+\mu\right\}$
Weibull	$\lambda[-\ln(\delta_\mp \mp N/\rho_\pm)]^{1/k}$
Pareto	$y_m(\delta_\mp \mp \rho_\pm/N)^{-1/\alpha}$
Exp.	$-\lambda^{-1}\ln(\delta_\mp \mp \rho_\pm/N)$

2.2 Detection Method I

Given a model distribution with parameters $\boldsymbol{\theta}$, the expected number of observations $\rho_\pm$ above ($+$) or below ($-$) a value $\ell^\pm$ is given by

$$\rho_\pm = N[\delta_\pm \mp F(\ell^\pm|\boldsymbol{\theta})], \tag{5}$$

where $\delta_\pm = 1$ for the upper symbol and 0 for the lower. Mindful of Eq. (5) the following definition of an outlier is suggested: given the distribution parameters $\boldsymbol{\theta}$, an observation will be called an *outlier with respect to* $\rho_\pm$ when it is above ($+$) or below ($-$) the value where less then $\rho_\pm$ observations are expected, conditional on the total number of observations N.

Solving $\ell^\pm$ from Eq. (5) and replacing $\boldsymbol{\theta}$ with its estimate, we get

$$\ell_\rho^\pm = F^{-1}\left(\delta_\pm \mp \frac{\rho_\pm}{N}\middle|\hat{\boldsymbol{\theta}}\right), \tag{6}$$

The label ρ is added to distinguish the limit from the limit defined in the next section. Since F is non-decreasing, the value of ℓ_ρ^+ (ℓ_ρ^-) decreases (increases) with increasing ρ, and therefore less observations will be identified as outlier when ρ increases. If one chooses $\rho_\pm < 1$, then $\ell_\rho^\pm$ represents the value above (below) which less than 1 observation is expected. In Table 2 the limits are calculated for various distributions used here.

2.3 Detection Method II

Based on the estimation method discussed in Subsection 2.1, we can deduce a second outlier definition as follows. Given a model distribution F with estimated $\hat{\boldsymbol{\theta}}$ and an observation y_j, where $j \notin \Lambda$. Assume the following null hypothesis:

$$H_0 : y_j \text{ is generated by } F(Y|\hat{\boldsymbol{\theta}}). \tag{7}$$

The largest and smallest values for which H_0 is rejected are identified as outliers. To make this precise, we use the test statistic E whose realizations ε are given by

$$\varepsilon_j = g(y_j) - g(F^{-1}(\hat{F}_j|\hat{\boldsymbol{\theta}})), \tag{8}$$

with g as in Eq. (1). A large observation y_k, with $k > \max\{\Lambda\}$ is called an *outlier with respect to* α_+ if

$$\begin{aligned}&\varepsilon_k \geq \ell_\alpha^+ = F_E^{-1}(1-\alpha_+|\hat{\boldsymbol{\phi}}) \text{ and}\\ &\qquad k = N \text{ or } y_{k+1} \text{ is also an outlier.}\end{aligned} \tag{9}$$

Here, F_E is the cdf for the residual distribution with parameters $\boldsymbol{\phi}$, estimated using observations indexed with Λ. The second condition ensures that an observation is only identified as an outlier when all larger observations are outliers too. Similarly, a small value y_i with $i < \min\{\Lambda\}$ is called an outlier with respect to α_- when

$$\begin{aligned}&\varepsilon_i \leq \ell_\alpha^- = F_E^{-1}(\alpha_-|\hat{\boldsymbol{\phi}}) \text{ and}\\ &\qquad i = 1 \text{ or } y_{i-1} \text{ is also an outlier.}\end{aligned} \tag{10}$$

Assuming that the residuals are normal distributed with mean 0 and variance σ_E^2, we get for the upper and lower limits

$$\ell_\alpha^\pm = \sqrt{2}\hat{\sigma}_E \text{erf}^{-1}\{\pm(1-2\alpha_\pm)\} \tag{11}$$

with $\hat{\sigma}_E^2 = |\Lambda|^{-1}\sum_{j\in\Lambda}\varepsilon_j^2$. and erf^{-1} the inverse of the error function. In this formulation, more observations on both the left and right-hand side of the distribution will be identified as outliers as α decreases.

2.4 An illustration

To illustrate the detection methods, a set of 100 $\log-\mathcal{N}(\mu = 0, \sigma = 1)$ distributed numbers were generated, the realized mean and standard deviation being -0.317 and 1.199 respectively (Note that for the lognormal distribution, $\mu(Y) = \text{E}(\ln Y)$ and $\sigma^2(Y) = \text{Var}(\ln Y)$). Two outliers were added to the random numbers: the first set to 0.1 times the smallest random number, the second set to 10 times the largest random number, yielding values of 0.007889 and 104.8071 respectively. Next, using all 102 “observations”, the values $\hat{F}_i$ were calculated, and $\hat{\mu}$ and $\hat{\sigma}$ were estimated using $F_{\mathsf{min}} = 0.1$ and $F_{\mathsf{max}} = 0.9$. In this case we found $\hat{\mu} = -0.331$ and $\hat{\sigma} = 1.230$, close to the realized value of the random sample. This is also reflected in the corresponding R^2 value which equals 0.9950.

Next, outliers were determined with both methods. For Method I we used $\rho_+ = \rho_- = 1$, yielding $\ell_{\rho=1}^- = 0.04073$ and $\ell_{\rho=1}^+ = 12.6568$. The top panel of

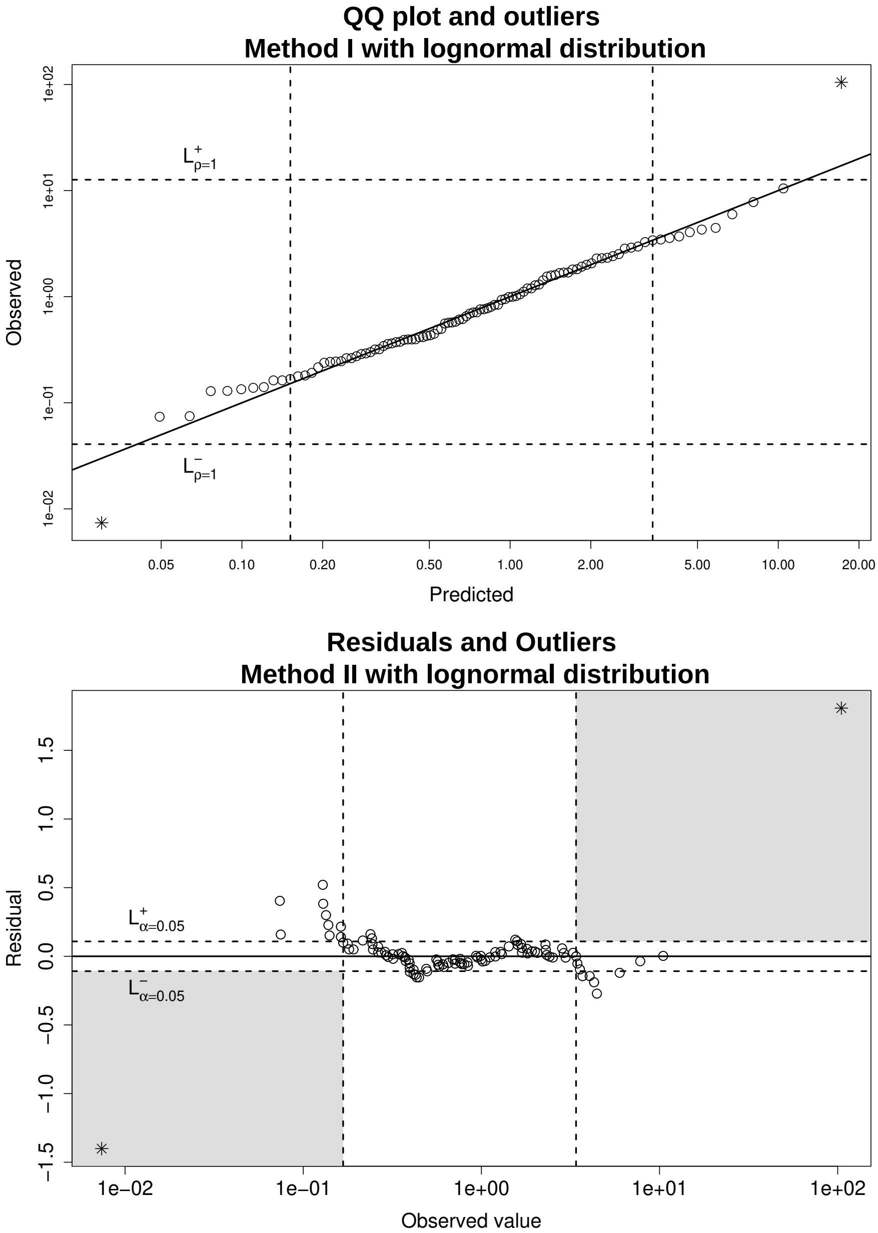

Figure 1. Results of outlier detection with Method II on simulated lognormal distributed data with two added outliers. Filled points indicate data used in the fit. Outliers are indicated with a $\ast$*. Upper panel: observed values versus predicted values, the continuous line indicating the perfect fit. Lower pane: residuals against observed values. The grey areas indicate the outlier region. Points between the vertical dashed lines were used in the fit, The horizontal dashed lines indicate the levels* $\ell^{\pm}_{\alpha}$*.*

Figure 1, shows the result of Method I outlier detection as well as an overview of fit quality by means of a QQ-plot. Points between the dashed vertical lines were used in the fit. The dashed horizontal lines indicate the levels $\ell^{\pm}_{\rho=1}$ and points above (below) the dashed horizontal lines are classified as outliers (indicated with $\ast$).

For Method II we used $\alpha_+ = \alpha_- = 0.05$. The bottom panel of Figure 1 combines an overview of the detection results as well as the fitting quality by means of a residual plot. As in the top panel, data points between vertical lines were used in the fit. Based on these data we find $\hat{\boldsymbol{\phi}} = (\hat{\mu}_E, \hat{\sigma}_E) = (0.000, 0.0659)$ giving $\ell^{\pm}_{\alpha=0.05} = \pm 0.153$. The horizontal dashed lines in the lower pane of Figure 1 indicate the levels $\ell^{\pm}_{\alpha=0.05}$, and the gray areas indicate where outliers can occur. The artificially added outliers are indeed classified as such by the method and are marked with a $\ast$. The values in the upper left rectangle in the lower pane have residuals larger than $\ell^{+}_{0.05}$, but they are no outliers since they are situated on the left side of the distribution (*i.e.* they are small even though they have large residuals). Similarly, values in the bottom right rectangle have residuals lower than $\ell^{-}_{0.05}$, but are not identified as extremely small values since they are on the right side of the distribution.

3 Application to Dutch VAT data

To further test the methods, they were applied to monthly Value Added Tax data of Dutch supermarkets (SBI code 47.11, see also SBI08 (2008)). VAT data is used by Statistics Netherlands in estimation processes for economic growth and the Structural Business Statistics, amongst others. Here, only outliers on the right side of the distribution (called "right outliers") are treated. For technical reasons, such as deductions, some VAT values can be negative. After removing records with nonpositive values from the dataset, 14880 records remained which were further devided into five size classes $h = 1, 2, \ldots, 5$ of sizes 463, 443, 297, 223, and 119 respectively. The size classification is based on the number of employees and enterprizes in higher size class have more employees. In Figure 2 an overview of the data is given in the form of box-and-whisker plots of the $\log_{10}$ of the turnover values.

Next, right outliers were identified with the two methods using all five model distributions. In each case we used $F_{\mathsf{min}} = 0.1$, $F_{\mathsf{max}} = 0.9$. In the case of Method I, the outlier detection limit ℓ^{+}_{ρ} was determined by $\rho = 0.5$. In the case of Method II, the maximum residual ℓ^{+}_{α} was determined by $\alpha = 0.05$. The parameters were equal for all five size classes. The resulting number of outliers for the different detection methods are shown in table 3. The variation of number of outliers as the method is varied from Method I to Method II is not the same for every used distribution. When the normal or Weibull distributions are used, Method II yields more outliers than Method I, while the lognormal, Pareto and exponential distributions are relatively insensitive to the used detection method.

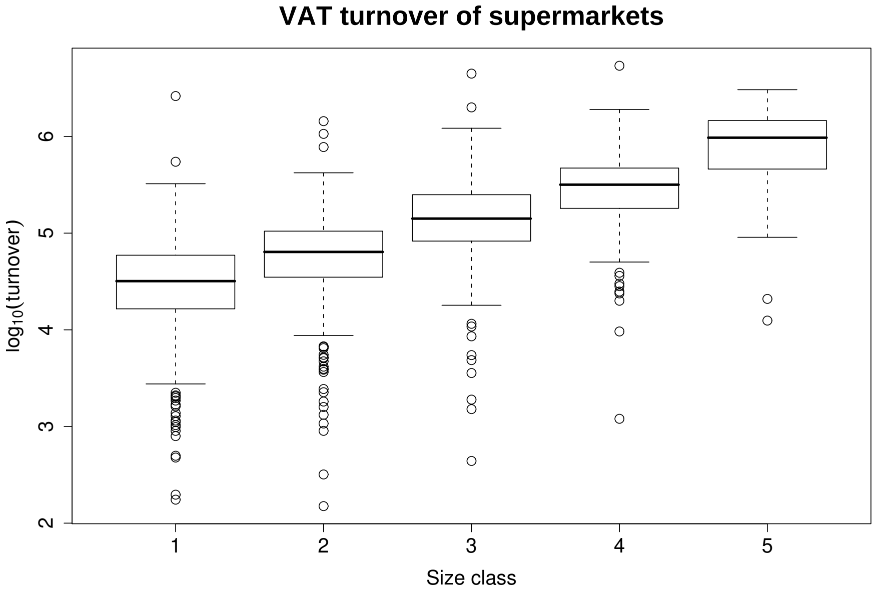

Figure 2. Boxplots of the logarithm (in base 10) of montly VAT turnover of 1545 Dutch supermarkets diveded into size classes.

The number of detected outlier depends strongly on the used distribution. If the normal or Weibull distribution is used, the number of outliers varies from 0 to 46 over the size classes while no outliers are detected at all when the Pareto distribution is used. Roughly, the number of detected outliers decreases in the order normal - Weibull - exponential - lognormal - Pareto distribution. Note that for Method I, where the value of ℓ_ρ^+ depends directly on the shape of the tail, this is consistent with the fact that for large enough y, the density distributions obey $F'_{\mathsf{nrm}} < F'_{\mathsf{exp}} < F'_{\mathsf{lnrm}} < F'_{\mathsf{pto}}$ (Explicitly, writing $y = \ln z$, one can show that $\ln F'_{\mathsf{nrm}} \to -e^{2z}$, $\ln F'_{\mathsf{exp}} \to -e^z$, $\ln F'_{\mathsf{nrm}} \to -z^2$,and $\ln F'_{\mathsf{pto}} \to -z$). The only exception to this is the Weibull distribution, which has asymptotic behaviour similar to the exponential distribution. As shown further on, this is due to the fact that the Weibull distribution overfits the observed data. The above observations, as well as the question which distribution to use in practice, can be explained by investigating the robustness of the methods for the parameter settings. In the top panel of Figure 3 the R^2 values of the model fits are plotted against the maximum bulk QQ plot position F_{max} for observations in size class 3. In the lower panel, the number of outliers resulting from different F_{max} values are plotted. Remember that a lower F_{max} implies that less values are used in the fit.

The reliability of the outlier detection methods described here, depends on the adequacy of the model distribution, as well as on the robustness of the outlier numbers as a function of fit parameters. For example, it can be seen that the Weibull distribution gives fairly good fits for all tested F_{max} values, with R^2-

Table 3. Right outliers in VAT turnover data, determined using Methods I and II, and various model distribuitions. Outliers were determined using $F_{\mathsf{min}} = 0.1$, $F_{\mathsf{max}} = 0.9$ $\rho_+ = 0.5$ (Method I), and $\alpha_+ = 0.05$ (Method II).

	nrm		lnrm		wbl		pto		exp	
h	I	II	I	II	I	II	I	II	I	II
1	17	46	2	2	7	30	0	0	4	7
2	14	44	2	3	5	44	0	0	3	5
3	9	30	2	4	5	30	0	0	3	5
4	8	22	1	1	5	14	0	0	1	3
5	1	3	0	0	0	0	0	0	0	0

values wich are second only to that of the lognormal distribution. However, it can be seen in the lower panel of Figure 3 that the number of outliers increases drastically when F_{max} is lowered from 0.9 to 0.6, especially for Method II. This indicates that the Weibull distribution is too flexible, giving a good fit at the used bulk observations but with bad extrapolation properties to higher lying observations. When Method II is used together with the Weibull distribution, basically all points which are not used in the fit are identified as outliers since they have large residuals. This can be considered an overfitting by the Weibull distribution. The normal distribution shows similar behaviour, although with lower overall R^2-values.

Both the Pareto distribution and the exponential distribution yield a fairly constant number of outliers as a function of F_{max}. However, both have R^2-values which depend strongly on F_{max}. This indicates that neither of these distributions can be seen as a proper description of the bulk distribution, since for perfectly Pareto (exponentially) distributed variables even the use of a few $\hat{F}_i$-values should give a reasonable estimate.

The only distribution for which both the quality of fit and the number of identified outliers is (almost) independent of the chosen F_{max} value is the lognormal distribution. In fact, the number of detected outliers is completely independent of F_{max} in this case and the lognormal distribution yields the highest R^2-values for the whole range of F_{max} values. Finally, we note that tests involving business types different from supermarkets gave similar results.

It is therefore concluded that 1) the lognormal distribution yields a reasonable description for the right-hand side of the distribution of Dutch VAT turnover values and 2) the values which are identified as outliers are indeed outliers in the sense that they are unlikely to be drawn from the same distribution as the

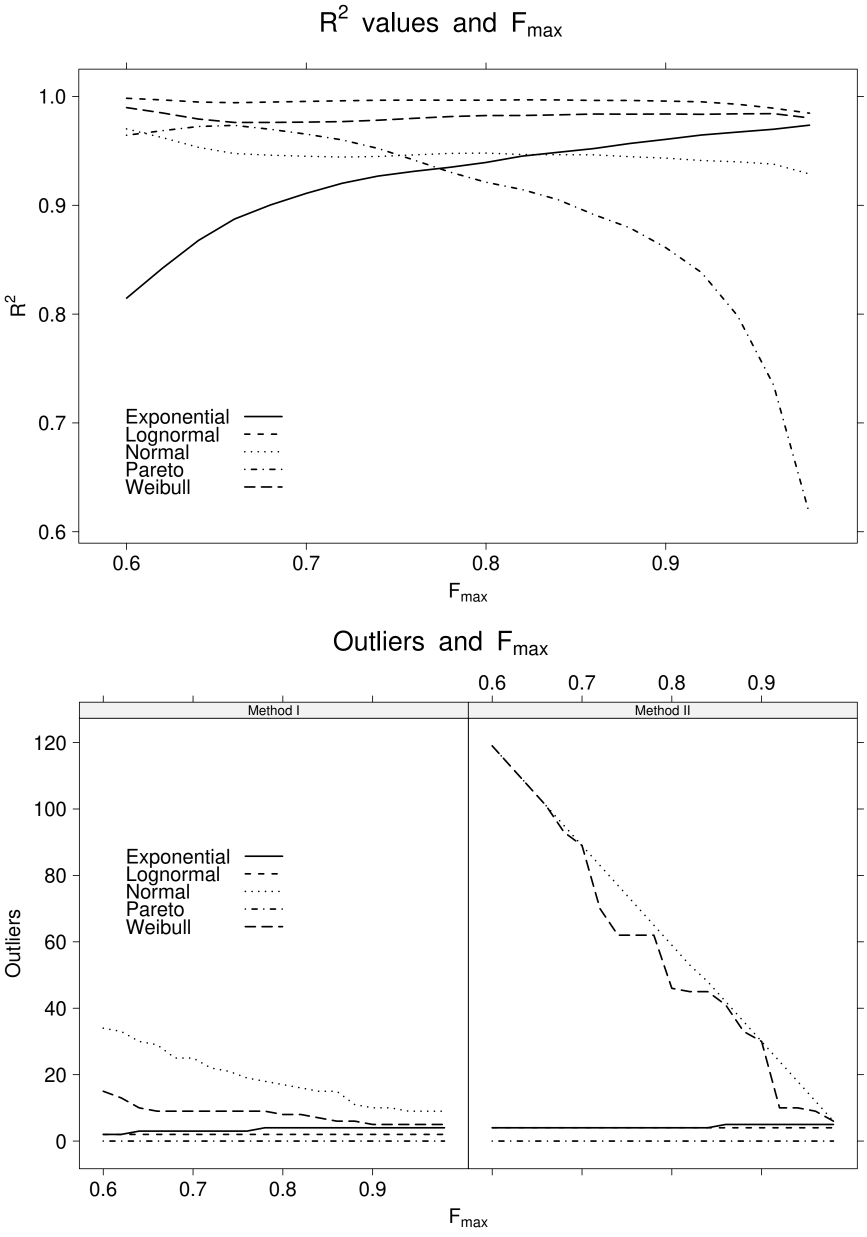

Figure 3. Dependence of R^2 values (upper panel) and number of detected outliers (lower panel) on the value of the bulk upper limit F_{max} for various methods and distributions applied to VAT data in size class 3.

bulk.

4 Conclusion

An outlier detection method is shown which uses a model distribution to describe the bulk, and identifies outliers as observations which are unlikely to be drawn from the same distribution. Also, a method to approximate model parameters based on regression on the QQ plot positions is pointed out. The methods have proven to be robust against chosen parameter settings once the right model distribution is chosen and similar parameter settings can be used for different size classes and business types. Also, the methods offer conceptual advantage since the technical definition of an outlier used here is a close translation of the concept of "not belonging to the bulk". Finally we mention that an outlier detection method based on Method I with the lognormal distribution, is now implemented as part of the new production process for the Structural Business Statistics at Statistics Netherlands. An implementation of this work has been submitted to the CRAN archive as the R package "extremevalues" (version 2.0).

References

M. Abramowitz and I.A. Stegun, editors. *Handbook of mathematical functions: with formulas graphs and mathematical tables*. Dover Publications Inc., New York, ninth edition, 1972.

V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley Series in Probability & Statistics. Wiley, New York, 3rd edition, 1994.

W. J. Dixon. Analysis of extreme values. *Annals of Mathematical Statistics*, 21:488–506, 1950.

W.J. Dixon. Processing data for outliers. *Biometrics*, 9:74–89, 1953.

F.E. Grubbs. Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21:27–58, 1950.

D. M. Hawkins. *Identification of outliers*. Monographs on Applied Probability and Statistics. Chapman and Hall, 1980.

L. Makkonen. Bringing closure to the plotting position controversy. *Communications in Statistics*, 37:460–467, 2008.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

SBI08, 2008. URL http://www.cbs.nl. The SBI08 enterprise index is Statistics Netherlands' implementation of the NACE enterprise classification standard.

M. P. J. van der Loo. *extremevalues: Outlier detection in univariate data*, 2010. URL http://www.r-project.org. R package version 2.0.