Application and Comparison of Robust Linear Regression Methods for Trend Estimation

ANDREAS MUHLBAUER,* PETER SPICHTINGER, AND ULRIKE LOHMANN
Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

(Manuscript received 8 August 2007, in final form 26 May 2009)

ABSTRACT

In this study, robust parametric regression methods are applied to temperature and precipitation time series in Switzerland and the trend results are compared with trends from classical least squares (LS) regression and nonparametric approaches. It is found that in individual time series statistically outlying observations are present that influence the LS trend estimate severely. In some cases, these outlying observations lead to an over-/underestimation of the trends or even to a trend masking. In comparison with the classical LS method and standard nonparametric techniques, the use of robust methods yields more reliable trend estimations and outlier detection.

1. Introduction

Within the framework of trend analysis, simple linear least squares (LS) regression models are widely used and allow for an extrapolation of different atmospheric variables into the future (e.g., Born 1996; Dai et al. 1997; Zerefos et al. 2003; Norris 2005; Solomon et al. 2007). Although linear regression models have been used successfully, a number of difficulties arise with the conceptual framework of linear trend analysis and its applicability to problems of atmospheric and climatic science. The mathematical framework of linear LS regression analysis crucially depends on the assumptions of independent observations and normally distributed error terms with constant variance. However, these assumptions are violated in many applications, which potentially leads to unreliable results of the LS regression (von Storch and Zwiers 1999). Furthermore, statistical outliers in the data pose a problem for the LS trend estimation because the LS estimator can react very sensitively to outlying observations (e.g., Rousseeuw and Leroy 1987; Helsel and Hirsch 1992; Wilcox 1998; von Storch and Zwiers 1999; Trömel and Schönwiese 2005). The problem that a single outlying observation may suffice to severely influence the LS regression estimator makes the LS regression a nonrobust method. As a consequence of outliers in the data, trends may be biased or masked, which may lead to a different interpretation of the data. Statistical outliers also affect significance levels and have major implications for the reliability of confidence intervals and hypothesis tests (Rousseeuw and Leroy 1987; Wilcox 1998). Despite these well-known problems, linear (parametric) LS models are widely used in atmospheric and climatic research, and applications of nonparametric or robust parametric approaches are scarce (Huth and Pokorna 2004). They are more developed in other research fields, for example, robust signal extraction (Davies et al. 2004; Bernholt et al. 2006; Fried et al. 2006), hydrology (Barbosa et al. 2004), and chemistry (Ortiz et al. 1996; Daszykowski et al. 2007).

The main objectives of this paper are to demonstrate how the choice of the regression estimator can affect the results of trend estimation and the interpretation of trends in climatic science. We therefore draw examples from temperature and precipitation records in Switzerland and compare trend results from ordinary LS regression with trend estimates from robust parametric regression models as well as from standard nonparametric approaches.

* Current affiliation: Joint Institute for the Study of the Atmosphere and Ocean/Department of Atmospheric Sciences, University of Washington, Seattle, Washington.

Corresponding author address: Andreas Muhlbauer, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätsstrasse 16, 8092 Zurich, Switzerland.
E-mail: andreasm@atmos.washington.edu

DOI: 10.1175/2009JAMC1851.1

© 2009 American Meteorological Society
2. Methods

a. Classical linear regression

The linear regression model may be written such that

\[ y = X\theta + e, \quad (1) \]

where \( X \) is a matrix of dimension \( n \times p \) that contains the \( p \) vectors of explanatory variables (each of length \( n \)) and \( y \) is the vector of response variables (e.g., von Storch and Zwiers 1999). The matrix \( X \) is referred to as a design matrix, and the explanatory variables are often called the predictors of the model. The vector \( \theta \) contains the \( p \) (unknown) coefficients (also called parameters) of the linear model. The elements \( e_i \) \((1 \leq i \leq n)\) of the error term are considered to satisfy a Gaussian distribution with mean \( \mu = 0 \) and unknown but constant variance \( \sigma^2 \).

Given the data \((X, y)\), one approximates the unknown parameters \( \theta \) by the estimated parameters \( \hat{\theta} \) so that the residuals \( r = y - \hat{y} \) between the observed values \( y \) and the estimated values \( \hat{y} = X\hat{\theta} \) are minimized.

The classical way to determine the regression coefficients is to introduce the LS estimator (also known as the \( L_2 \) estimator), which corresponds to the minimization of the sum \( Q \) of the \( n \) squared residuals with respect to the coefficients \( \theta \):

\[ Q = \min_{\theta} \sum_{i=1}^{n} r_i^2. \quad (2) \]

The regression coefficients \( \theta \) satisfying the minimization criterion \( (2) \) are given analytically by the normal equation (von Storch and Zwiers 1999)

\[ \theta = (X^TX)^{-1}X^Ty. \quad (3) \]

The LS estimator is commonly used because statistical inferences such as confidence intervals or hypothesis testing are mathematically simple as long as the model error is normally distributed. Confidence intervals may be constructed as discussed in appendix A but are only reliable if the model assumptions are valid and in the absence of statistical outliers (Rousseeuw and Leroy 1987; Wilcox 1998).

b. Robust linear regression

In the presence of outliers, the breakdown point of the chosen estimator plays an important role (Rousseeuw and Leroy 1987; Rousseeuw and van Zomeren 1990). Following the definition of Hampel (1971) and Hodges (1967), the (finite sample) breakdown point of an estimator is the smallest fraction of contamination that may cause the estimates to take on values arbitrarily far away from the uncontaminated sample estimates. In other words, the breakdown point gives the maximum contamination the data may contain to still provide reliable estimates about the model parameters (coefficients) (Rousseeuw and Leroy 1987; Maronna et al. 2006). For the application, this means that the higher the breakdown value is, the more robust is the estimator. Rousseeuw (1984) showed that the breakdown point of the LS estimator is zero. Thus, a single outlier could perturb the linear trend crucially. Two options to overcome this lack of robustness are the least median of squares (LMS) estimator and the least trimmed squares (LTS) estimator, which are described next.

1) The LMS estimator

Based on the ideas of Hampel (1975), Rousseeuw (1984) introduced the LMS estimator by replacing the sum in the minimization criterion \( (2) \) by the median, which is then given by

\[ Q = \min_{\theta} \text{med}_{i} r_i^2. \quad (4) \]

In simple regression applications, the LMS solution is the narrowest strip that covers one-half of the observations (Rousseeuw 1984). The advantage of the LMS estimator is that the breakdown point achieves the highest expectable value of 50%. Thus, the LMS is robust with respect to outlying observations and leverage points (Rousseeuw 1984) and the LMS trend estimate is not as easily biased as the LS trend estimate (Davies et al. 2004). The main disadvantage of many robust regression methods such as the LMS is the absence of a simple analytical relation for the estimates such as \( (3) \) for the LS regression. Therefore, the LMS estimator must be computed iteratively. The convergence rate of the LMS unfortunately decreases rapidly with increasing sample size or complexity of the regression problem (e.g., multivariate regression), which makes this method computationally more expensive than the classical LS method. A further disadvantage of the LMS estimator is connected to its weak statistical efficiency, which makes the construction of confidence bounds much more complex than for the classical LS approach (see appendix A). However, several options exist to improve the statistical efficiency of the LMS as discussed in Rousseeuw and Leroy (1987) and Gervini and Yohai (2002).
2) THE LTS ESTIMATOR

A further robust estimator is the so-called LTS estimator (Rousseeuw 1984; Rousseeuw and Leroy 1987) given by

\[ Q = \min_{\theta} \sum_{i=1}^{h} r_{in}^2 \quad r_{1n} \leq \cdots \leq r_{hn}, \]

which minimizes the sum of the first \( h \) smallest squared residuals of a subset \( h \) out of \( n \) data points (Rousseeuw 1984). Note that the residuals \( r_i \) are first squared and then ordered according to their size. If the subset \( h \) is chosen to be \( h = (n/2) + 1 \), the breakdown point of the LTS estimator is the same as for the LMS while the convergence rate is much higher (Rousseeuw 1984; Verboven and Hubert 2005). In contrast to the LMS estimator, the statistical efficiency of the LTS estimator is improved, which is a further benefit of this method (Rousseeuw and van Driessen 2006). For a comprehensive review of the mathematical properties of the LTS estimator, we refer to Agullo et al. (2008). Methods to develop confidence intervals can be found in appendix A or alternatively in Willems and van Aelst (2005).

Throughout this paper we use the freely available statistical software package R (see online at http://www.r-project.org/) to compute the LS, LMS, and LTS estimates for simple linear regression applications.

3. Comparison of robust linear regression with temperature and precipitation time series in Switzerland

In this section, the distinct regression estimators discussed in section 2 are applied to several temperature and precipitation time series in Switzerland for the period of 1864–2007. All time series are quality controlled and homogenized (Begert et al. 2003, 2005). Linear trend estimates based on the classical LS regression and on the robust regression are calculated and compared.

a. Example of annual temperature trends

The time series of the annual mean temperature of the station Lugano serves as a first example to show how sensitively the classical LS trend model may react with respect to single or multiple statistically outlying observations. A brief discussion on the detection of outlying observations is given in appendix B. Figure 1a shows the time series of the annual mean temperature for the station Lugano for the period of 1864–2007.

The LS trend model reveals a linearly increasing trend of \(+0.8^\circ C (100 \text{ yr})^{-1}\) over the given period of 1864–2007, which is highly significant as deduced from the classical \( t \) statistics (see Table 1). The 95% confidence intervals bracket the slope estimate for the LS within the bounds \([0.0061, 0.0103]\), which correspond to a temperature increase between \(+0.6^\circ C\) and \(+1.0^\circ C (100 \text{ yr})^{-1}\). In contrast,

![Figure 1](https://example.com/figure1.png)

**TABLE 1. LS trend estimates.**

<table>
<thead>
<tr>
<th>Example</th>
<th>LS trend line</th>
<th>( t(\text{slope}) )</th>
<th>( p(\text{slope}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual mean temperature Lugano (Fig. 1a)</td>
<td>( y = 0.0082x - 4.26 )</td>
<td>7.66</td>
<td>0.000</td>
</tr>
<tr>
<td>Annual mean temperature Bern (Fig. 1b)</td>
<td>( y = 0.0114x - 14.18 )</td>
<td>9.16</td>
<td>0.000</td>
</tr>
<tr>
<td>Annual precipitation Davos (Fig. 3a)</td>
<td>( y = 0.2164x + 644.74 )</td>
<td>0.61</td>
<td>0.540</td>
</tr>
<tr>
<td>Annual precipitation Chaumont (Fig. 3b)</td>
<td>( y = 1.0182x - 752.02 )</td>
<td>2.64</td>
<td>0.009</td>
</tr>
</tbody>
</table>
the robust parametric LTS and LMS methods show a much weaker trend or almost no trend. The centennial increase in annual mean temperature for Lugano found by employing the LMS and LTS ranges from +0.01°C to +0.6°C, respectively. Note that the robust solutions are not included within the LS 95% confidence bounds. Hence, the robust solutions from LMS and LTS trend lines are different and statistically distinguishable from the LS solution (see also Table 2).

The indicated outliers in annual mean temperature are the early years of the last century, the observations in the last two decades, and the recent years with heat records (e.g., Schär and Jendritzky 2004). The outliers are based on the 97.5th percentile of the normal distribution and are obtained from the robust standardized residuals $r_i/\hat{\sigma}^*$ in Fig. 2, which is described in more detail in Appendix B. Thus, these outliers may be seen as the extreme events in the time series of Lugano. Note that the standardized residuals of the LS differ from the LMS residuals substantially in terms of outlier diagnosis. The LMS residuals show that the extreme observations in the early years and in the last years are influential points and attract the LS trend line crucially. In this case, the outlying observations yield an overestimation of the LS trend with respect to the trends given by the robust estimators. This example also illustrates the clear advantage of using the robust standardized residuals for detecting outlying observations in comparison with the classical standardized LS residuals. The residuals of the LS and LMS differ because the scale estimate $\hat{\sigma}$ itself depends on the estimated trend line (and thus on the underlying regression estimator) and, hence, is a nonrobust measure.

The assumption of normally distributed error and homoscedasticity (constant variance) is not satisfied in this specific example (especially for the years after 1990) as can be seen more clearly from the LMS residuals or the normal probability plot (Figs. 2e,f). These violations in the model assumptions question not only the reliability of the estimates obtained by the classical LS method, but also the inferences such as the significance in terms of the $t$ statistics, the coefficient of determination, or the confidence intervals.

The linearly increasing annual mean temperature trend for the station Bern (Fig. 1b) is affirmed independently by all three methods, suggesting that the warming observed at this station is a robust signal. The LS trend estimate gives a temperature increase of approximately +1.2°C (100 yr)$^{-1}$. The 95% confidence intervals are [0.0090, 0.0140] and, hence, bracket the centennial temperature increase between +0.9°C and +1.4°C. Furthermore, the LS confidence intervals include the LMS and LTS solutions (see Table 2). In fact, the LMS and LTS slope estimates are very close to the LS slope estimate and yield temperature trends of +1.2°C and +1.3°C (100 yr)$^{-1}$, respectively. Given the LS uncertainty range, one may interpret the robust trend estimates as statistically not distinguishable.

The LMS residuals unmask several observations as statistically outlying, which again influences the LS trend estimate toward these points (see also Fig. 2). Based on the LS residuals, many of the outliers would have been equally identified, although the observation 1868 would not have been identified from the LMS standardized residuals and, in fact, is not an extreme event. This observation attracts the LS trend estimate and may explain why the LS trend is slightly lower than the robust trends. Furthermore, it can be seen from the normal probability plot (see Fig. 2) that the assumption of normally distributed residuals is violated. However, the few outlying observations together with the violation of the model assumptions do only marginally affect the LS trend estimate in this example.

**b. Example of annual precipitation trends**

In the second example, we compare the trends in the time series of annual precipitation for the stations Davos and Chaumont for the period of 1864–2007 (Fig. 3). All precipitation trends are subsequently given as percentage change per 100 years with respect to the 1961–90 average.

The LTS and LMS trend estimators both support a statistically significant (95% confidence level) increasing linear trend of approximately +8% for the annual precipitation in Davos (Fig. 3a). In contrast, the LS estimator only reveals a very weak trend in annual precipitation of approximately +2% (100 yr)$^{-1}$ that is not statistically significant at the 95% confidence level. Note that the confidence intervals for the LS slope bracket the LS precipitation trend between −4% and +8%. Thus, the LS confidence intervals barely include the LMS and LTS solutions but would also allow for negative trends.

---

**Table 2.** LMS and LTS trend estimates. The given $t$ and $p$ values refer to the LTS slope estimates.

<table>
<thead>
<tr>
<th>Example</th>
<th>LMS trend line</th>
<th>LTS trend line</th>
<th>$t$(slope)</th>
<th>$p$(slope)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual mean temperature Lugano (Fig. 1a)</td>
<td>$y = 0.0001x + 1.11$</td>
<td>$y = 0.0055x + 0.88$</td>
<td>5.26</td>
<td>0.000</td>
</tr>
<tr>
<td>Annual mean temperature Bern (Fig. 1b)</td>
<td>$y = 0.0109x - 12.96$</td>
<td>$y = 0.0128x - 16.81$</td>
<td>10.58</td>
<td>0.000</td>
</tr>
<tr>
<td>Annual precipitation Davos (Fig. 3a)</td>
<td>$y = 0.8750x - 670.10$</td>
<td>$y = 0.9062x - 695.62$</td>
<td>2.58</td>
<td>0.011</td>
</tr>
<tr>
<td>Annual precipitation Chaumont (Fig. 3b)</td>
<td>$y = 1.2275x - 1094.81$</td>
<td>$y = 1.3225x - 1329.78$</td>
<td>3.47</td>
<td>0.001</td>
</tr>
</tbody>
</table>
From the LMS and LS standardized residuals shown in Fig. 4 several statistically outlying observations can be identified. However, a subset of outliers between 1860 and 1930 that is only identified based on the LMS standardized residuals influences the LS trend line remarkably and, thus, masks the increasing precipitation trend of the station Davos.

The positive precipitation trend for 1864–2007 estimated for the station Chaumont (Fig. 3b) can be reproduced by all three methods and, hence, is a robust trend. Again, as in the case of the temperature from the previous example, the slope estimates and intercept values differ slightly for the different regression methods and bracket the precipitation increase between +8% (LS), +10% (LMS), and +11% (LTS) per century.

The 95% confidence bounds for the LS slope include the LTS and LMS solution. However, several dry years that yield outlying observations influence the LS trend estimate and may explain why the LS estimator underestimates the precipitation trend with respect to the robust estimators.

4. Discussion and conclusions

The results of section 3 show that the LS estimator can react very sensitively to outlying observations and, thus, can affect trend estimation results and interpretation. In general, the influence of these statistical outliers on the LS estimator tends to be higher toward the boundaries than in the center part of the time series. This general
and problematic feature is especially persistent and dangerous in temperature trends in which a strong increase has been observed during the last two decades. Future climate scenarios also suggest an increase in the variability and an increase of rare and extreme events such as heat waves and heavy precipitation (Katz and Brown 1992; Schär et al. 2004; Seneviratne et al. 2006). The occurrence of such extreme events in turn affects the amount of statistically outlying observations. Our examples demonstrate the vulnerability of the LS estimator to these outlying observations and emphasize the necessity of using robust estimators in climatic science. Because robust parametric estimators such as the LTS or LMS are not easily biased in the slope estimate (Davies et al. 2004), we encourage the use of robust estimators in climate-related work to reduce the effect of outliers on trend estimates.

We also compared the classical LS and robust trends against trends derived from nonparametric approaches such as the Spearman rank correlation coefficient (SRCC; Sachs 1984; Hess et al. 2001), the Mann–Kendall test (Gilbert 1987), and Sen’s nonparametric estimate of slope (Sen 1968; Hollander and Wolfe 1973). In many cases the trends found with parametric and nonparametric methods are very similar and are mostly included within the 95% confidence interval of the LS, which is a result also found by Huth and Pokorna (2004). The SRCC and the Mann–Kendall test indicate a positive trend in all examples with a high level of confidence. In qualitative terms, these nonparametric trends correspond to the trend signs found with the LS method, which questions the reliability of the SRCC and the Mann–Kendall test in the presence of outlying observations. Sen’s slope estimator tends in many cases more toward the trend estimates of the robust methods (when compared with the LS trend) and corroborates the trend signs and magnitudes found by applying the robust parametric methods.

Some of our examples from section 3 raise the more general question of whether linear trend models are adequate in applications of climatic science. In particular, temperature time series often show a considerable amount of nonlinearity. For the annual mean temperature the robust methods give a trend that differs remarkably from the LS trend estimate. This suggests that the LS trend estimate is severely attracted by the data points that belong to the warmer period from 1980 to 2007 whereas the robust methods are only weakly influenced. Because the data points in this period are not outliers but rather belong to another population, the linear trend models may not properly represent the variability in the data. In contrast, nonlinear methods (e.g., Miksovsky and Raidl 2006) may be the better choice to account for the variability and support the upward temperature trend in the late twentieth century.

In conclusion, the comparisons of ordinary and robust regression methods show that outlying observations may bias the LS trend estimate and lead to over-/underestimation of trends or even trend masking. Hence, trend estimation results and interpretation can be affected, which suggests the use of robust estimators. Based on our findings, the benefits of using robust parametric regression methods are twofold. First, robust parametric regression methods may be used in addition to the classical LS method to check its reliability and reproducibility. Second, the robust standardized residuals provide a useful and simple diagnostic tool to identify outlying observations more reliably than with the standardized residuals of the classical LS approach.

Acknowledgments. We thank the anonymous reviewers for their valuable comments and suggestions. We
are grateful to W. Stahel, M. Mächler (ETH Seminar für Statistik), and A. Ruckstuhl (Zürcher Hochschule Winterthur) for helpful discussions. MeteoSwiss is kindly acknowledged for sharing the data.

**APPENDIX A**

**Construction of Confidence Intervals**

a. *LS estimator*

A confidence interval for the $j$th estimate $\hat{\theta}_j$ ($1 \leq j \leq p$) may be constructed by calculating the scale estimate $\hat{\sigma}^2$ (estimated variance) from the residuals $r$:

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^{n} r_i^2. \quad (A1)$$

With the estimated variance–covariance matrix $\hat{\sigma}^2(XX^T)^{-1}$ the $(1 - \alpha) \times 100\%$ confidence bounds for the estimate $\hat{\theta}_j$ are (e.g., Rousseeuw and Leroy 1987)

$$\hat{\theta}_j \pm t_{1-\alpha/2,n-p} \sqrt{\hat{\sigma}^2(XX^T)^{-1}}_{jj} \quad j = 1, \ldots, p. \quad (A2)$$

Here, $t_{1-\alpha/2,n-p}$ denotes the $1 - \alpha/2$ quantile of a Student’s distribution with $n - p$ degrees of freedom and the probability of error $\alpha$. The subscript $jj$ denotes the $j$th diagonal element of the variance–covariance matrix. Note that the construction of the confidence intervals involves the assumption of independent and normally distributed error terms.

b. *LMS estimator*

Given the probability of error $\alpha$, one may compute the $(1 - \alpha) \times 100\%$ confidence intervals for the $j$th LMS estimates $\hat{\theta}_{j,LMS}$ ($1 \leq j \leq p$) in a manner similar to (A2):
with the robust LMS scale estimate $\hat{\sigma}^2_{\text{LMS}}$ and the residuals $r_i$ ($1 \leq i \leq n$) obtained from the LMS regression (Rousseeuw and Leroy 1987; Rousseeuw and van Zomeren 1990); $V(LMS, F)$ denotes the asymptotic variance, which depends on the chosen estimator and the error statistics. Here, the estimator is the LMS and $F$ denotes the true cumulative distribution function (cdf) of the error with $f$ being the corresponding probability density function (pdf). If one assumes, for example, a normal distributed error, then $F$ is the normal cumulative distribution $\Phi$ and $f$ is the normal probability density function $\phi$. The asymptotic variance of the LMS $V(LMS, \Phi)$ with breakdown $0 \leq \beta \leq 0.5$ is then given by D. J. Olive (2008, unpublished manuscript, p. 238, available online at http://www.math.siu.edu/olive/ol-bookp.htm):

$$V[LTS(\beta), \Phi] = \frac{1}{1 - \beta - 2k\Phi(k)} k = \Phi^{-1}\left(1 - \frac{\beta}{2}\right).$$

Note that in the limit

$$\lim_{\beta \to 0} V[LTS(\beta), \Phi] = 1$$

the asymptotic variance $V[LTS(\beta), \Phi]$ of the LTS estimator is equal to unity and the LTS estimator has the same zero breakdown as the classical LS estimator. For a breakdown value of $\beta = 0.5$, the asymptotic variance of the LTS estimator is obtained from (A7) to be $V[LTS(0.5), \Phi] = 14.02$, which corresponds closely to the value given by Rousseeuw and Leroy (1987, p. 191).

A preliminary robust scale estimate for the LTS estimator $\sigma^*_\text{LTS}$ may be obtained such that

$$\sigma^*_\text{LTS} = d_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^{h} r_{i,n}^2}$$

with

$$d_{h,n} = \left\{1 - \frac{2n}{h} \Phi^{-1}(\gamma)\phi[\Phi^{-1}(\gamma)]\right\}^{-1/2}, \quad \gamma = \frac{h + n}{2n}.$$  

Note that $d_{h,n}$ is a constant that depends only on the sample size $n$ and the size of the subset $h \leq n$ of the LTS regression to make the LTS scale estimate $\sigma^*_\text{LTS}$ consistent with the Gaussian model (Rousseeuw and Leroy 1987). For small sample sizes (e.g., $n \leq 30$), the LTS estimator may underestimate the scale of the residuals and thus may flag too many observations as outlying. To overcome this problem a correction factor may be applied to (A10) as proposed by Pison et al. (2002). With the preliminary scale estimate $\sigma^*_\text{LTS}$, the robust LTS scale estimate $\hat{\sigma}_{\text{LTS}}$ may be computed in a similar manner to the robust LMS scale estimate $\hat{\sigma}_{\text{LMS}}$ according to (A5).

An alternative and nonparametric approach for the construction of robust LTS confidence intervals that is independent of the underlying error distribution is given.

Here, $\sigma^2_{\text{LTS}}$ is the robust LTS scale estimate and $V(LTS, F)$ is the asymptotic variance of the LTS estimator with a cdf of the error according to $F$. For a normally distributed error, $F$ is the normal cdf $\Phi$ with corresponding normal pdf $\phi$. The asymptotic variance of the LTS $V[LTS(\beta), \Phi]$ with breakdown $0 \leq \beta \leq 0.5$ is then given by D. J. Olive (2008, unpublished manuscript, p. 238, available online at http://www.math.siu.edu/olive/ol-bookp.htm):

$$V[LTS(\beta), \Phi] = \frac{1}{1 - \beta - 2k\Phi(k)} k = \Phi^{-1}\left(1 - \frac{\beta}{2}\right).$$

Note that in the limit

$$\lim_{\beta \to 0} V[LTS(\beta), \Phi] = 1$$

the asymptotic variance $V[LTS(\beta), \Phi]$ of the LTS estimator is equal to unity and the LTS estimator has the same zero breakdown as the classical LS estimator. For a breakdown value of $\beta = 0.5$, the asymptotic variance of the LTS estimator is obtained from (A7) to be $V[LTS(0.5), \Phi] = 14.02$, which corresponds closely to the value given by Rousseeuw and Leroy (1987, p. 191).

A preliminary robust scale estimate for the LTS estimator $\sigma^*_\text{LTS}$ may be obtained such that

$$\sigma^*_\text{LTS} = d_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^{h} r_{i,n}^2}$$

with

$$d_{h,n} = \left\{1 - \frac{2n}{h} \Phi^{-1}(\gamma)\phi[\Phi^{-1}(\gamma)]\right\}^{-1/2}, \quad \gamma = \frac{h + n}{2n}.$$  

Note that $d_{h,n}$ is a constant that depends only on the sample size $n$ and the size of the subset $h \leq n$ of the LTS regression to make the LTS scale estimate $\sigma^*_\text{LTS}$ consistent with the Gaussian model (Rousseeuw and Leroy 1987). For small sample sizes (e.g., $n \leq 30$), the LTS estimator may underestimate the scale of the residuals and thus may flag too many observations as outlying. To overcome this problem a correction factor may be applied to (A10) as proposed by Pison et al. (2002). With the preliminary scale estimate $\sigma^*_\text{LTS}$, the robust LTS scale estimate $\hat{\sigma}_{\text{LTS}}$ may be computed in a similar manner to the robust LMS scale estimate $\hat{\sigma}_{\text{LMS}}$ according to (A5).

An alternative and nonparametric approach for the construction of robust LTS confidence intervals that is independent of the underlying error distribution is given.

Here, $\sigma^2_{\text{LTS}}$ is the robust LTS scale estimate and $V(LTS, F)$ is the asymptotic variance of the LTS estimator with a cdf of the error according to $F$. For a normally distributed error, $F$ is the normal cdf $\Phi$ with corresponding normal pdf $\phi$. The asymptotic variance of the LTS $V[LTS(\beta), \Phi]$ with breakdown $0 \leq \beta \leq 0.5$ is then given by D. J. Olive (2008, unpublished manuscript, p. 238, available online at http://www.math.siu.edu/olive/ol-bookp.htm):

$$V[LTS(\beta), \Phi] = \frac{1}{1 - \beta - 2k\Phi(k)} k = \Phi^{-1}\left(1 - \frac{\beta}{2}\right).$$

Note that in the limit

$$\lim_{\beta \to 0} V[LTS(\beta), \Phi] = 1$$

the asymptotic variance $V[LTS(\beta), \Phi]$ of the LTS estimator is equal to unity and the LTS estimator has the same zero breakdown as the classical LS estimator. For a breakdown value of $\beta = 0.5$, the asymptotic variance of the LTS estimator is obtained from (A7) to be $V[LTS(0.5), \Phi] = 14.02$, which corresponds closely to the value given by Rousseeuw and Leroy (1987, p. 191).

A preliminary robust scale estimate for the LTS estimator $\sigma^*_\text{LTS}$ may be obtained such that

$$\sigma^*_\text{LTS} = d_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^{h} r_{i,n}^2}$$

with

$$d_{h,n} = \left\{1 - \frac{2n}{h} \Phi^{-1}(\gamma)\phi[\Phi^{-1}(\gamma)]\right\}^{-1/2}, \quad \gamma = \frac{h + n}{2n}.$$  

Note that $d_{h,n}$ is a constant that depends only on the sample size $n$ and the size of the subset $h \leq n$ of the LTS regression to make the LTS scale estimate $\sigma^*_\text{LTS}$ consistent with the Gaussian model (Rousseeuw and Leroy 1987). For small sample sizes (e.g., $n \leq 30$), the LTS estimator may underestimate the scale of the residuals and thus may flag too many observations as outlying. To overcome this problem a correction factor may be applied to (A10) as proposed by Pison et al. (2002). With the preliminary scale estimate $\sigma^*_\text{LTS}$, the robust LTS scale estimate $\hat{\sigma}_{\text{LTS}}$ may be computed in a similar manner to the robust LMS scale estimate $\hat{\sigma}_{\text{LMS}}$ according to (A5).
by the robust bootstrap method discussed in Willems and van Aelst (2005).

**APPENDIX B**

**Outlier Detection**

For detecting statistical outliers we examine the standardized residuals. We plot the standardized residuals $r/\hat{\sigma}$ for the LS (Figs. 2a,b and 4a,b) and the standardized residuals $r/\hat{\sigma}^*$ for the LMS (Figs. 2c,d and 4c,d) against the explanatory variable as suggested by Draper and Smith (1966). For the standardized LMS residuals, the robust scale estimate $\hat{\sigma}^*$ is computed as described in appendix A [(A4)]. Observations are classified as outliers if the values of $r/\hat{\sigma}^*$ taken from the standardized LMS residuals exceed the limit of 1.96, which corresponds to the 97.5th percentile of a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. The normal probability plots are shown in Figs. 2e,f and 4e,f. The trend estimates as well as the corresponding statistics for accepting or rejecting the null hypothesis (no trend) are shown in Table 1.

Note that to be significant at the 95% level the absolute value of $r$ (slope) has to exceed the critical value of the 97.5th percentile of a Student’s distribution with $n-p$ degrees of freedom, which is roughly $t_5(0.975, n-p) = 2$ in all of the examples. The value of $p$ (slope), which is the probability of erroneously rejecting the null hypothesis, has to be lower than $\alpha = 0.05$.

**REFERENCES**


