Regional residual plots for assessing the fit of linear regression models

E. Deschepper*, O. Thas, J.P. Ottoy

Ghent University, Department of Applied Mathematics, Biometrics and Process Control, Coupure Links 653, B-9000 Gent, Belgium

Accepted 22 March 2005
Available online 29 April 2005

Abstract

An intuitively appealing lack-of-fit test to assess the adequacy of a regression model is introduced together with a graphical diagnostic tool. The graphical method itself includes a formal testing procedure, and, it is particularly useful to detect the location of lack-of-fit. The procedure is based on regional residuals, using subsets of the space of the independent variables. A simulation study shows that, the proposed procedures in simple linear regression have similar power as those of some popular classical lack-of-fit tests. In case of local departures from the hypothesized regression model, the new tests are shown to be more powerful. Therefore, when it becomes difficult to discriminate between systematic deviations and noise, regional residual plots are very helpful in formally locating areas of lack-of-fit in the predictor space. Data examples illustrate the ability of the new methods to detect and to locate lack-of-fit.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Diagnostic tool; Graphical method; Lack-of-fit; Multiple regression

1. Introduction

An important problem in applied statistics is the examination of the adequacy of parametric regression models to fit the observed data. Residuals are highly informative for this purpose and are widely used in both statistical tests as well as graphical diagnostic tools.
Among the graphical diagnostic tools, the classical residual plot is probably the best known. It is often used as a descriptive method to assess lack-of-fit in a regression analysis. In general, graphical methods allow visualization of possible discrepancies between the fitted model and the data. Nevertheless, judging whether the observed discrepancies are really present or not is often a major problem and systematic departures smaller than the noise level can often not be observed. So far, most graphics introduced to assess the adequacy of regression models are illustrative and indicative, and the results depend on the data analyst. Kuchibhatla and Hart (1996) proposed an approach to lack-of-fit testing wherein graphs of smoothers play both a descriptive and inferential role. This approach is an attempt to obtain a graphical diagnostic tool, including a lack-of-fit test, but a major disadvantage of this procedure is its dependency on the choice of the smoother and its bandwidth.

In this paper, a formal procedure is introduced based on so-called regional residuals, which are defined on subsets of the sample space of the independent variables. A graphical diagnostic tool and a corresponding statistical test are proposed to check how well a parametric linear model for the mean fits to a set of observed data. The graphical method itself includes a formal testing procedure to assess the adequacy of a regression model. In particular, it allows the localization of lack-of-fit in the predictor space. The test statistic is simple and intuitively appealing, and is closely related to the lack-of-fit tests of Lin et al. (2002), who use cumulative and moving sums of residuals, or to earlier work of Stute (1997) or the cusum-based test discussed by Buckley (1991), who consider only cumulative sums of residuals.

Other popular, classical lack-of-fit tests are constructed from nonparametric smoothers. Eubank et al. (1993) and Hart (1997) provide a number of references of smoothing-based tests. Eubank and Hart (1993) concluded that the cusum test is the most powerful for very smooth departures from the no-effect hypothesis, whereas the smoothing-based tests are clearly superior when the alternative is high frequency. In more recent papers, e.g. Kuchibhatla and Hart (1996), Aerts et al. (1999, 2000), omnibus lack-of-fit tests that use data-driven selection criteria are proposed.

In Section 2, the regional residuals are defined on subsets of the sample space of the independent variables. The proposed tests and graphical tools are described in the linear regression context, and a bootstrap approach is proposed for approximating critical values and \( p \)-values. Also in Section 2, the link with the cumulative and moving sums of residuals of Lin et al. (2002) is discussed in more detail. In Section 3 the results of a simulation study are presented. The behavior of the new method is also investigated in case the model assumption of homoscedasticity is violated. Section 4 provides an illustration of the methodology on two data examples, and, finally, the concluding remarks are summarized in Section 5.

2. Plots and tests based on regional residuals

2.1. Regional residuals

In this section the discussion is restricted to the linear regression setting, assuming identically and independently distributed error terms. Given the independent observations
(\(x_i, y_i\), \(i = 1, \ldots, n\), with \(x_i = (x_{1i}, \ldots, x_{(p-1)i}) \in \mathbb{R}^{p-1}\), the model

\[ y_i = r_\beta(x_i) + \epsilon_i, \quad i = 1, \ldots, n \]

is considered, where the \(\epsilon_i\)'s are i.i.d. distributed variables with \(E(\epsilon_i) = 0\) and \(\text{Var}(\epsilon_i) = \sigma^2\). \(r_\beta\) is a real linear regression function defined on \(\mathbb{R}^{p-1}\), with parameter vector \(\beta\).

Consider testing the hypothesis of a linear model, \(r_\beta(x_i) = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ki}\). The residuals are defined by \(e_i = y_i - r_\beta(x_i)\), \((i = 1, \ldots, n)\), where \(\hat{\beta}\) is assumed to be a consistent estimator for \(\beta\), like the least squares estimator. Regional residuals with respect to the \(k\)th covariate \(X_k\) are defined as the average of residuals in the subset \(A_{kij} = [x_{ki}, x_{kj}]\),

\[
R(A_{kij}) = \frac{\sum_{i=1}^{n} e_i I(x_{ki} \leq X_{kl} \leq x_{kj})}{\sum_{i=1}^{n} I(x_{ki} \leq X_{kl} \leq x_{kj})} (i \leq j; i, j = 1, \ldots, n),
\]

where \(n_{kij} = \sum_{i=1}^{n} I(x_{ki} \leq X_{kl} \leq x_{kj})\) is the number of observations in the subset \(A_{kij}\), and the design points are ordered with respect to the \(k\)th covariate \(X_k\). Of course, other directions can be investigated in the same way, e.g. principal components or fitted values.

Under the null hypothesis of no lack-of-fit, these regional residuals have zero mean and variance equal to \(n_{kij}^{-1} \sigma^2 h^2_{kij}(X)\), where \(h^2_{kij}(X) = (I_{A_{kij}}^T I_{A_{kij}})^{-1} I_{A_{kij}}^T (I_n - H) I_{A_{kij}}\). \(I_{A_{kij}}\) is a \(n \times 1\) inclusion matrix, with \(I_{A_{kij},l} = 1\) if \(x_{kl} \in A_{kij}\), else 0, \(H = X(X^T X)^{-1} X^T\) represents the hat matrix, and \(I_n\) is the \(n \times n\) identity matrix.

Standardization of the regional residuals is an important issue in making the regional residuals comparable among another. However, in practice, the residual variance \(\sigma^2\) is unknown, but can be replaced by a consistent estimator. First the natural estimator, \(S^2 = (n - p)^{-1} \sum_{i=1}^{n} (y_i - r_\hat{\beta}(x_i))^2\) is considered, resulting in the standardized regional residuals

\[
R_S^2(A_{kij}) = \sqrt{n_{kij}} \frac{R(A_{kij})}{Sh_{kij}}. \tag{1}
\]

In the particular case of normally distributed error terms, it is straightforward to show that \(\sqrt{n_{kij}} R(A_{kij}) \overset{H_0}{\sim} N(0, \sigma^2 h^2_{kij}(X))\). Under the null hypothesis of no lack-of-fit, \((n - p)S^2/\sigma^2 \sim \chi^2_{n-p}\). Hence, \(R_S^2(A_{kij}) \overset{H_0}{\sim} \chi_{n-p}\).

2.2. A lack-of-fit test

For all possible intervals of the \(k\)th covariate \(X_k\), \(A_{kij} = [x_{ki}, x_{kj}]\), \(i, j = 1, \ldots, n; i \leq j\), the standardized regional residuals are calculated. Large absolute values of these standardized regional residuals indicate a possible lack-of-fit. To overcome the problem of multiplicity and to obtain a global measure of lack-of-fit, the supremum norm of all the calculated standardized regional residuals is proposed as a test statistic,

\[
T_{k,S}^2 \overset{i \leq j; i,j=1,...,n}{\sup} \left| R_S^2(A_{kij}) \right|. \tag{2}
\]
This test statistic only contains marginal information on lack-of-fit with respect to the $k$th covariate $X_k$, but they can be further combined into one global test statistic $T$, defined as the supremum of the $p - 1$ marginal statistics $T_{k,S}^2(k = 1, \ldots, p - 1)$,

$$T = \sup_{k=1,\ldots,(p-1)} \left( T_{k,S}^2 \right).$$

(3)

The derivation of the asymptotic distribution of $T$ is beyond the scope of this paper. Since the model is assumed to be linear, the asymptotic null distribution of $T$ is independent of the parameters (Hart, 1997), but depends on the design. Therefore, the use of a simulated null distribution for the test statistic is suggested, allowing both fixed and random designs. In what follows this test is called the RRS test.

Bootstrap $p$-values are used in hypotheses testing. Since the model assumptions include homoscedasticity, the ordinary-residual-based bootstrap is performed as follows (Davison and Hinkley, 1997). Bootstrap samples,

$$y^{*}_k = \hat{r}_k(x_k) + \hat{e}_k, \quad k = 1, \ldots, n$$

are constructed by resampling the least squares residuals. In this way, the data-generating distribution obeys the null hypothesis and is based on estimates of the unknown parameters.

2.3. Regional residual plots

In case of simple linear regression, for which the intervals are denoted by $A_{ij} = [x_i, x_j]$ ($i \leq j; i, j = 1, \ldots, n$), the lack-of-fit test can be complemented with a two-dimensional formal graphical tool, which is called a regional residual plot. It is the contour plot obtained by plotting the standardized regional residuals, $R_{S}^2(A_{ij})$, in the $(i, j)$ plane. Although the regional residuals are only defined for $i \leq j$, the regional residual plot is made symmetrical by filling up the half plane $i > j$ with $R_{S}^2(A_{ji})$. An artificial example based on 50 observations is presented in Fig. 1(b). Fig. 1(a) shows the corresponding true and fitted regression models, which show a clear local lack-of-fit in $[0.27;0.51]$. Light gray areas in the regional residual plot refer to large positive standardized regional residuals, exceeding the bootstrap critical value. Similarly, dark gray areas indicate regions with large negative standardized regional residuals. ‘Nonsuspicious’ standardized regional residuals are colored white. Hence, whenever one light or dark gray spot appears in this formal regional residual plot, the null hypothesis of no lack-of-fit is rejected, and, in addition, the plot is able to locate regions of lack-of-fit instead of single outlying observations.

Its use is illustrated in Fig. 1(b). At the $\alpha = 0.05$ significance level, the light gray areas show a significant underestimation of the data in many intervals situated in the lower range of the predictor variable, which contain the first half of the interval where the lack-of-fit was introduced. Similarly, the dark gray areas indicate a significant overestimation of the data in small intervals containing the second half of the interval where the lack-of-fit was introduced. In this way, the data analyst can identify areas in the predictor space that deserve special attention. Note that the plots only provide an idea of over- or underestimation of the observations in that specific area. They are not a real tool that suggest how to ameliorate the model.
The ability to formally locate areas of lack-of-fit is a major advantage of the regional residual plot over the ordinary scatter plot, or plots contrasting the null model with a smoother. In the latter, local discrepancies might be observed thanks to the ability of the human eye to detect patterns, but it is hard to judge whether the observed discrepancies are really present or just noise. The regional residual plot actually complements these plots by indicating which observed discrepancies are statistically significant.

Especially in multiple regression, where graphical display of the regression model and the observed data is hardly possible, the marginal regional residual plots, introduced in the next subsection, can be very helpful.

2.4. Marginal regional residual plots

In case of more than one predictor variable, marginal regional residual plots are constructed by plotting the standardized regional residuals in each point of the \((i, j)\) plane of the selected covariate \(X_k\), but still assigning a light or dark gray color whenever the \(\alpha\)-level critical value of the global test statistic \(T\) (Eq. (3)) is exceeded. So, whenever one light or dark gray spot appears in any marginal regional residual plot, the global null hypothesis of no lack-of-fit is rejected at the \(\alpha\) significance level. In addition, the marginal plots show in which variables a region of lack-of-fit occurs and where this area is located. These marginal plots include a lack-of-fit test itself and allows to conclude in a formal way whether the multiple linear regression model is appropriate or not. The usefulness of these marginal plots in localizing lack-of-fit is illustrated in Section 4.
2.5. Related test statistics

The proposed test statistic is closely related to the one proposed by Stute (1997), who studied the process

$$R(x) = n^{-1/2} \sum_{i=1}^{n} I(X_i \leq x) e_i,$$

where \( I(X_i \leq x) = I(X_{1i} \leq x_1, \ldots, X_{(p-1)i} \leq x_{p-1}) \). This test statistic is constructed by considering the supremum norm. In what follows, the test will be abbreviated as the CS test. Since the process \( R \) accumulates all the residuals associated with covariate values less than \( x \), it tends to be dominated by residuals with small covariate values. This problem can be overcome by considering moving sums or moving averages of residuals with respect to one covariate \( X_j, j = 1, \ldots, p - 1 \) as proposed by Lin et al. (2002), who also use the supremum norm to obtain a global measure of lack-of-fit. The moving sums of residuals are calculated over blocks of fixed size \( b \),

$$W_j(x; b) = n^{-1/2} \sum_{i=1}^{n} I(x - b < X_{ji} \leq x) e_i.$$

However, the moving sums are based on blocks of the same size, so the number of observations in the blocks can be quite different when the covariate values are not evenly distributed. Therefore, also moving averages were studied,

$$\overline{W}_j(x; b) = \frac{n^{-1/2} \sum_{i=1}^{n} I(x - b < X_{ji} \leq x) e_i}{\sum_{i=1}^{n} I(x - b < X_{ji} \leq x)}.$$

The power of these tests depends on the choice of \( b \). Larger values of \( b \) will lead to more powerful tests when a lack-of-fit is situated over the entire range of the predictor variable \( X_j \), while smaller values of \( b \) are needed to detect local deviations (Section 3). A test based on moving sums with fixed block size \( b \) is denoted by the MSb test. The method proposed in this paper solves this problem by considering all possible intervals obtained with respect to each covariate \( X_j, j = 1, \ldots, p - 1 \), which results in powerful tests in case of both global and local lack-of-fit.

2.6. Nonparametric variance estimators

Ideally, a variance estimator that is consistent under both null and alternative hypotheses is used. Unfortunately, using \( S^2 \), the residual variance is overestimated under a lack-of-fit situation, which might result in low power. The use of variance estimators which are more robust against deviations of the null model may therefore be more appropriate. In what follows, the focus will be on some difference-based estimators in nonparametric simple regression. Difference-based type of variance estimators are attractive from a practical point of view, as they are computationally simple and often have a small bias for small
sample sizes. Rice (1984) introduced a simple local residual estimator,
\[ \hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2. \]

Another popular choice of a nonparametric estimator of residual variance is the estimator proposed by Gasser et al. (1986), based on local linear fitting. For equally spaced data, the estimator is defined as
\[ \hat{\sigma}_G^2 = \frac{2}{3(n-2)} \sum_{i=3}^{n} \left( \frac{1}{2} y_{i-2} - y_{i-1} + \frac{1}{2} y_i \right)^2. \]

By replacing the residual variance estimator \( S^2 \) by these estimators, the standardized regional residuals are
\[ R_R(A_{ij}) = \sqrt{n_{ij}} \frac{R(A_{ij})}{\hat{\sigma}_R h_{ij}} \quad \text{and} \quad R_G(A_{ij}) = \sqrt{n_{ij}} \frac{R(A_{ij})}{\hat{\sigma}_G h_{ij}}. \]

The corresponding test statistics \( T_R \) and \( T_G \) become
\[ T_R = \sup_{i \leq j; i, j = 1, \ldots, n} \left| R_R(A_{ij}) \right| \quad \text{and} \quad T_G = \sup_{i \leq j; i, j = 1, \ldots, n} \left| R_G(A_{ij}) \right|. \] (4)

Bootstrap simulated null distributions (Section 2.2) of these test statistics will be used for the power study in Section 3. These tests are abbreviated as the RRR and the RRG test.

2.7. Heteroscedasticity

To deal with heteroscedastic errors, a wild bootstrap procedure will be used. As described in Flachaire (2004) wild bootstrap samples are drawn from the data-generating process \( y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i v_i^* \), where \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are the least squares parameter estimates, \( e_i \) the least squares residual, and \( v_i^* \) a random variable with expectation \( E(v_i^*) = 0 \), variance \( E(v_i^*^2) = 1 \) and third moment \( E(v_i^*^3) = 1 \). Mammen (1993) suggests the most popular choice for the distribution of \( v_i^* \),
\[ F_1 : v_i^* = \begin{cases} -\left(\sqrt{5} - 1\right)/2 & \text{with probability } p = \left(\sqrt{5} + 1\right)/\left(2\sqrt{5}\right), \\ \left(\sqrt{5} + 1\right)/2 & \text{with probability } 1 - p. \end{cases} \] (5)

Recently, Davidson and Flachaire (2001) have shown that the Rademacher distribution
\[ F_2 : v_i^* = \begin{cases} 1 & \text{with probability } 0.5, \\ -1 & \text{with probability } 0.5 \end{cases} \] (6)

may lead to better results than the popular version \( F_1 \).
2.8. Nonlinearity

Consider testing the hypothesis of a nonlinear model, \( r_{\beta}(x_i) \). In this case, the regional residuals, the test statistics and the regional residual plots can be constructed in exactly the same way as described in the previous sections. In general, the null distribution of the test statistic in Eqs. (2) and (3) now depend on the unknown value of the parameter \( \beta \) (Hart, 1997). However, \( p \)-values can be obtained by the use of the bootstrap discussed in Section 2.2.

3. Simulation results

To learn about the small sample power characteristics of the proposed tests in simple linear regression, a simulation study is performed comparing the empirical powers of the RRS, RRR and RRG tests, with those of the closely related tests of Lin et al. (2002) and three classical lack-of-fit tests. The supremum test with cumulative sums of residuals (Stute, 1997; Lin et al., 2002) is abbreviated as the CS test. Since the power of the tests based on moving sums depends on the choice of \( b \), the fixed block size, three different block sizes are included in the study, corresponding to the range of the lowest 10%, 30% and 50% of the covariate values, represented as the MS10, MS30 and the MS50 test. The first classical lack-of-fit test is a generalization of the Von Neumann (1941) test, described by Hart (1997) (the \( H \) test). This test is included in the study as it is an omnibus consistent test for testing lack-of-fit and, according to Hart (1997), this test performs well under the same conditions as assumed in this paper. The \( H \) test statistic, \( T_H \), is a variance ratio of a model-based estimator of variance, \( \hat{\sigma}_M^2 = S^2 \), and a reasonable estimator of variance under the alternative hypothesis.

The cusum test statistic, \( T_B \), discussed by Buckley (1991) (abbreviated as the \( B \) test) is based on the sums \( S_j \) of all the classical residuals for which \( x \prec x_j \). As this test also uses the idea of sums of residuals in several intervals, it is included in the study. In particular, \( T_B \) is the ratio of a sensitive to a robust variance estimator.

Finally, the smoothing-based lack-of-fit test proposed by Kuchibhatla and Hart (1996) (abbreviated as the \( KH \) test) is considered. This test statistic, \( T_{KH} \), is based on a trigonometric series regression estimator whose data-driven smoothing parameter is the point at which the series is truncated. This test is known to be powerful in detecting high frequency departures from the hypothesized regression model.

In this study, the asymptotic null distribution of test statistics \( T_H \) and \( T_B \) is used, while the bootstrap procedure discussed in Section 2.2 was used for all other tests. Calculations were performed using R and C++. To reduce the computing time for the estimation of power of the bootstrap tests, a Monte Carlo power study was set up based on the simple linear extrapolation method proposed in Boos and Zhang (2000). For each scenario, \( O = 1000 \) data sets are generated under the alternative, resulting in \( O \) estimated \( p \)-values, \( \hat{p}_{1,1}, \ldots, \hat{p}_{1,O} \), each of which is obtained from resampling \( I = 59 \) times (bootstrap). A linear extrapolation procedure further results in a bias-adjusted power estimate. A sufficiently accurate approximation of the nominal level is observed in nearly all cases (Tables 2, 3, and 5).
Table 1
Estimated power in case of global and local lack-of-fit ($\gamma = 12.5$, $19$ and $36$, $\lambda = 0.5$, $n = 50$) situated in the lower and mid range of the predictor variable

<table>
<thead>
<tr>
<th>Location</th>
<th>$\gamma$</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RRS</td>
</tr>
<tr>
<td>Lower</td>
<td>12.5</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>0.252</td>
</tr>
<tr>
<td>Mid</td>
<td>12.5</td>
<td>0.911</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.611</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>0.189</td>
</tr>
</tbody>
</table>

In what follows, three main questions are discussed:

- How good is the small sample performance of the new tests as compared to classical ones, assuming homoscedasticity and Gaussian error terms?
- How do the three new tests perform in case of heteroscedasticity?
- How do they behave for heavy-tailed residuals?

A fixed, equidistant design with one covariate will be used.

3.1. Homoscedasticity and Gaussian error terms

To address the first issue, focus on the null hypothesis of a linear model, $r_{\beta}(x_i) = \beta_0 + \beta_1 x_i$, where the vector $x$ of the independent variate is fixed by design and $x_i \in [0, 1]$. Continuous data are generated, $Y_i \sim N(r_{\beta}(x_i), \sigma^2)$ with $\sigma^2 = 0.1$ and with an equidistant design $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$, for different sample sizes, $n = 20, 50$ and $100$. High frequency alternatives are studied with both global and local lack-of-fit. The lack-of-fit is introduced as one period of a sine function,

$$r_{\beta}(x_i) = 5 - 2x_i + \lambda \sin(\gamma x_i) I \{\delta_1 \leq i \leq \delta_2\},$$

where the amplitude, $\lambda = 0.10, \ldots, 0.90$, determines the strength of the lack-of-fit. The period, $2\pi/\gamma$, with $\gamma = 12.5$, 16, 19, 24 or 36, determines the length of the interval where the lack-of-fit occurs, varying from global departures ($\gamma = 12.5$) to local departures ($\gamma = 36$). Finally, $\delta_1$ and $\delta_2$ are the lower and upper bounds of the interval which depend on the period of the sine function. Fig. 1(a) shows a simulated data set with $\sigma^2 = 0.06^2$, $\lambda = 0.2$, $n = 50$, $\gamma = 24$, $\delta_1 = 0.27$ and $\delta_2 = 0.51$.

In Table 1, the empirical power of the RRS, RRR and RRG tests, are compared with those of the closely related tests of Lin et al. (2002). A lack-of-fit of $\lambda = 0.5$ is introduced for three different lengths of intervals, $\gamma = 12.5$, global departures, $\gamma = 19$, and $\gamma = 36$, local departures, situated in the lower and mid range of the predictor variable ($n = 50$). The results clearly show the dependency of the power on the choice of the fixed block size $b$. 
Larger values of $b$ will lead to more powerful tests when a lack-of-fit is situated over a larger range of the predictor variable, while smaller values of $b$ are needed to detect more local deviations. Also the inferior performance of the CS test in case the lack-of-fit is situated in the mid-range instead of the lower range of the predictor space can be observed. On the other hand, the RRS, RRR and RRG tests performed well in all cases. Similar results were found in all other simulations presented further in this section. Therefore, only the results of the RRS, RRR and RRG tests, together with those of the classical lack-of-fit tests will be shown in the remainder of this paper.

The main results of the power study with $n = 50$ are visualized using power curves. Fig. 2 displays the power curves for several alternatives. A distinction is made between large (global), intermediate and small (local) intervals of lack-of-fit and intervals situated at the start of the predictor range (low-range), in the middle (mid-range) or at the end (high-range). The plots for the high-range are not shown as they are similar to these of the low-range.

When comparing the three new and the three classical tests under different conditions of lack-of-fit, the following conclusions can be made. In case of a global lack-of-fit, all tests have good power characteristics (Fig. 2), with a slight advantage for the smoothing-based KH test, and a rather bad performance of the cusum-based B test in the mid-range. It may be concluded that for global departures from the simple linear regression model the power of the new tests are comparable to those of the classical tests.

For lack-of-fit intervals of medium length, hardly any difference in performance can be seen between the smoothing-based KH test and the regional-residual-based tests. As the length of the interval decreases, it becomes more difficult to discriminate between systematic deviations or noise. In this case, the regional-residual-based tests have the best power independent of the location, in particular the RRR test. Notice the complete power breakdown of the cusum B test and the poor performance of the KH test. In contrast, the power of the three new tests decrease only very slowly with decreasing length of the lack-of-fit interval. This means that for local departures from the simple linear regression model, our tests perform much better in comparison with the three classical tests.

In order to study the effect of the sample size, data were simulated with sample sizes 20, 50 and 100. Some results are presented in Table 2. The scenario with lack-of-fit strength $\lambda = 0.5$ was chosen as a representative. In general, the previous conclusions seem to remain valid. In particular, for all sample sizes studied, the power of the three regional residual tests are quite similar, with a minor power advantage of the RRR test, especially in small samples. The power advantage of this test can be explained by the fact that the Rice estimator of variance has the smallest bias in this particular case (Dette et al., 1998).

In general, the new tests have similar power as the KH test and perform better than the H and B tests when global lack-of-fit occurs. The power of regional-residual-based tests even exceeds those of the classical tests in case of local lack-of-fit. The major advantage of the new procedures, which is the ability of the regional residual plots to formally locate lack-of-fit, is illustrated in Fig. 3(a). In this graph, each point $(i, j)$ corresponds to a particular interval for which a probability $P_{ij}$ is estimated and plotted. $P_{ij}$ is the probability that the corresponding standardized regional residual is larger than the 5% critical value of the global test $T_R$. This probability is estimated by the ratio of the number of times the standardized regional residual exceeds the simulated critical value of $T_R$ and the total number of simulation runs (5000).
The study was performed under the condition that the lack-of-fit is introduced in two small intervals over the $x$-range, in $[0.19; 0.35]$ and $[0.79; 0.95]$ with $\lambda = 0.7$ and $\sigma^2 = 0.1$. Fig. 3(b) shows an example of the local lack-of-fit simulated under these conditions. It is clearly observed in Fig. 3(a) that mainly the regional residuals calculated over intervals including the area of lack-of-fit, are responsible for the rejection of the null hypothesis.
Table 2
Estimated power with $r(x_i) = 5 - 2x_i + \lambda \sin(\gamma x_i) I_{\delta_1 \leq i \leq \delta_2}$ for various sample sizes ($n = 20, 50$ and $100$), various interval lengths ($\gamma = 12.5, 19$ and $36$) and in case of no lack-of-fit ($\lambda = 0.0$) and of lack-of-fit ($\lambda = 0.5$) in the low-range

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\gamma$</th>
<th>$n$</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RRS</td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td>20</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.053</td>
</tr>
<tr>
<td>0.5</td>
<td>12.5</td>
<td>20</td>
<td>0.385</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>0.5</td>
<td>19</td>
<td>20</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.995</td>
</tr>
<tr>
<td>0.5</td>
<td>36</td>
<td>20</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>0.726</td>
</tr>
</tbody>
</table>

Fig. 3. (a) Contour plot showing the estimated probability to reject the null hypothesis of no lack-of-fit in each interval $[x_i, x_j]$; (b) true (dashed line) and fitted (solid line) regression model ($n = 50$); local lack-of-fit situated in $[0.19; 0.35]$ and $[0.79; 0.95]$ with $\lambda = 0.7$ and $\sigma^2 = 0.1$.

This figure also suggests that the power of the Lin et al. (2002) tests depend on the choice of the length of the interval.

3.2. Heteroscedasticity and Gaussian error terms

As in the simulation study of Dette and Munk (1998), a simulation study is set up with three different models for the standard deviation to study the loss of efficiency in using the
Table 3
Estimated power for various variance functions I–III, in case of no lack-of-fit

<table>
<thead>
<tr>
<th>Model</th>
<th>c</th>
<th>Tests</th>
<th>RRS</th>
<th>RRR</th>
<th>RRG</th>
<th>H</th>
<th>B</th>
<th>KH</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.0</td>
<td>0.049</td>
<td>0.052</td>
<td>0.052</td>
<td>0.048</td>
<td>0.056</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.053</td>
<td>0.055</td>
<td>0.054</td>
<td>0.054</td>
<td>0.053</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.055</td>
<td>0.054</td>
<td>0.057</td>
<td>0.053</td>
<td>0.053</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>0.5</td>
<td>0.059</td>
<td>0.068</td>
<td>0.067</td>
<td>0.054</td>
<td>0.048</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.088</td>
<td>0.077</td>
<td>0.072</td>
<td>0.061</td>
<td>0.049</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>0.5</td>
<td>0.051</td>
<td>0.058</td>
<td>0.059</td>
<td>0.054</td>
<td>0.057</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.056</td>
<td>0.059</td>
<td>0.058</td>
<td>0.057</td>
<td>0.051</td>
<td>0.055</td>
<td></td>
</tr>
</tbody>
</table>

proposed procedures under heteroscedastic errors,

\[
\sigma(x) = \sigma \exp(cx) \quad \text{monotone, model I,} \tag{8}
\]

\[
\sigma(x) = \sigma(1 + c \sin(10x))^2 \quad \text{high frequency, model II,} \tag{9}
\]

\[
\sigma(x) = \sigma(1 + cx)^2 \quad \text{unimodal, model III,} \tag{10}
\]

where different values for \(c = 0, 0.5, 1.0\) are used, \(\sigma^2 = 0.1\), and the error is assumed to be standard normal.

To deal with heteroscedastic errors, the two wild bootstrap procedures, discussed in Section 2.7, can be used. When using the popular distribution \(F_1\), suggested by Mammen (1993), for the random variables \(v^*_i\), instead of the Rademacher distribution \(F_2\), the size distortion was larger and the power smaller in all cases (results not shown). Therefore, the Rademacher distribution is recommended.

The empirical size and power results are presented in Tables 3 and 4. To make the power of all tests comparable, all estimated rejection probabilities are based on the wild bootstrap method with distribution \(F_2\). Under the no lack-of-fit null hypothesis, sufficiently accurate approximations to the nominal level for the bootstrap method are observed in nearly all cases. All possible scenarios of lack-of-fit discussed in Section 3.1 are reconsidered here. Only some representative results of global and local lack-of-fit are shown in Table 4. For all tests, the power clearly decreases in case of increasing heteroscedasticity. The KH and B test tend to achieve the best power in case of global lack-of-fit, although the RRR test often performs almost equally well. In case of local lack-of-fit, the RRR and H tests outperform all other tests.

3.3. Homoscedasticity and nonGaussian error terms

Finally, the performance of the tests is investigated when dealing with heavy-tailed error distributions. To address this issue, the same settings are adopted as in Section 3.1, but, as in Dette and Munk (1998), \(t\)-distributed error terms with 4 degrees of freedom instead of normally distributed error terms are considered. First the error terms are rescaled to obtain the same variance as in Section 3.1. The estimated power based on the wild bootstrap are
Table 4
Estimated power for various variance functions I–III, in case global lack-of-fit ($\gamma = 12.5$, $\lambda = 0.5$) and local lack-of-fit ($\gamma = 36$, $\lambda = 0.9$)

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$\lambda$</th>
<th>Model</th>
<th>$c$</th>
<th>Tests</th>
<th>RRS</th>
<th>RRR</th>
<th>RRG</th>
<th>H</th>
<th>B</th>
<th>KH</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>0.5</td>
<td>I</td>
<td>0.0</td>
<td>0.97</td>
<td>0.972</td>
<td>0.948</td>
<td>0.853</td>
<td>0.893</td>
<td>0.971</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>0.782</td>
<td>0.791</td>
<td>0.74</td>
<td>0.585</td>
<td>0.714</td>
<td>0.793</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.375</td>
<td>0.404</td>
<td>0.37</td>
<td>0.317</td>
<td>0.423</td>
<td>0.459</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>0.5</td>
<td>0.516</td>
<td>0.552</td>
<td>0.492</td>
<td>0.426</td>
<td>0.749</td>
<td>0.665</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.211</td>
<td>0.227</td>
<td>0.201</td>
<td>0.177</td>
<td>0.436</td>
<td>0.271</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>0.5</td>
<td>0.513</td>
<td>0.534</td>
<td>0.496</td>
<td>0.408</td>
<td>0.514</td>
<td>0.579</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.155</td>
<td>0.184</td>
<td>0.173</td>
<td>0.179</td>
<td>0.245</td>
<td>0.218</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.9</td>
<td>I</td>
<td>0.0</td>
<td>0.418</td>
<td>0.703</td>
<td>0.661</td>
<td>0.71</td>
<td>0.073</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>0.315</td>
<td>0.497</td>
<td>0.472</td>
<td>0.5</td>
<td>0.071</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.135</td>
<td>0.237</td>
<td>0.234</td>
<td>0.285</td>
<td>0.061</td>
<td>0.188</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>0.5</td>
<td>0.274</td>
<td>0.344</td>
<td>0.309</td>
<td>0.314</td>
<td>0.058</td>
<td>0.264</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.164</td>
<td>0.171</td>
<td>0.159</td>
<td>0.147</td>
<td>0.057</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>0.5</td>
<td>0.185</td>
<td>0.302</td>
<td>0.296</td>
<td>0.338</td>
<td>0.069</td>
<td>0.237</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.071</td>
<td>0.116</td>
<td>0.126</td>
<td>0.174</td>
<td>0.058</td>
<td>0.114</td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Estimated power for heavy-tailed data, using wild F2 bootstrap, in case of no lack-of-fit ($\lambda = 0.0$) and lack-of-fit ($\gamma = 12.5, 19, 36, \lambda = 0.3, 0.6$)

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$\lambda$</th>
<th>Tests</th>
<th>RRS</th>
<th>RRR</th>
<th>RRG</th>
<th>H</th>
<th>B</th>
<th>KH</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5</td>
<td>0.0</td>
<td>0.049</td>
<td>0.054</td>
<td>0.056</td>
<td>0.051</td>
<td>0.047</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.49</td>
<td>0.556</td>
<td>0.507</td>
<td>0.408</td>
<td>0.527</td>
<td>0.595</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.938</td>
<td>0.969</td>
<td>0.957</td>
<td>0.937</td>
<td>0.95</td>
<td>0.982</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.3</td>
<td>0.282</td>
<td>0.364</td>
<td>0.335</td>
<td>0.274</td>
<td>0.201</td>
<td>0.369</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.841</td>
<td>0.911</td>
<td>0.869</td>
<td>0.84</td>
<td>0.642</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.3</td>
<td>0.092</td>
<td>0.154</td>
<td>0.147</td>
<td>0.129</td>
<td>0.059</td>
<td>0.111</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.271</td>
<td>0.429</td>
<td>0.418</td>
<td>0.458</td>
<td>0.101</td>
<td>0.329</td>
<td></td>
</tr>
</tbody>
</table>

presented in Table 5, and are similar to those in the homoscedastic case with normal errors, except that some power loss is reported due to the use of the wild bootstrap procedure instead of the residual-based bootstrap. Therefore, it may be concluded that the performance of the new tests is quite robust for heavy-tailed data.
4. Data examples

4.1. Windmill data

The lack-of-fit tests and corresponding formal regional residual plots introduced in Section 2.3 are applied on the windmill data (Montgomery and Peck, 1982) (Fig. 4(a)). These data contain information on the Direct Current (DC) Output, the dependent variable $y$, and the wind velocity (miles/h), the independent variable $x$. It has been suggested that a reciprocal transformation on $x$ is appropriate. It will be shown how the new test detects and locates the lack-of-fit when a linear model is fitted to this data. To approximate the percentiles of the test statistic, a residual-based bootstrap algorithm described in Kuchibhatla and Hart (1996), is used based on 100,000 bootstrap samples drawn from the classical residuals. In this way, the critical value at the $\alpha = 0.05$ level of significance was estimated as 3.62. This value is used to construct the regional residual plot (Fig. 4(b)). The fact that there are light and dark gray areas in the regional residual plot, provides enough evidence that there is a lack-of-fit at the 0.05 level of significance. The value of the test statistic $T_R$ amounts 7.48 ($p = 0.00$). The plot also shows that the null hypothesis of no lack-of-fit is rejected in rather large intervals, mainly containing design points from the mid-range and even larger intervals, including almost the entire range. The lack-of-fit is also detected in very small intervals at the very low- and high-range of $x$. It is clear that the overestimation in the low- and high-range and the underestimation of the data points in the mid-range of the predictor variable is statistically significant.

Fig. 5 shows the scatter plot of the DC Output and a reciprocal transformation on the wind velocity, with a fitted linear regression model. When the analysis is repeated in this case, the formal regional residual plot (Fig. 5(b)) no longer shows evidence of a lack-of-fit. The value $T_R = 1.60$ is much smaller than the simulated critical value of 2.67 ($p = 0.85$).

4.2. US temperatures data

To illustrate the tests and corresponding marginal regional residual plots introduced in Section 2.4, data from the Data and Stories Library (DASL) is used. The US Temperatures data (Peixoto, 1990) gives the normal average January minimum temperature, $y$, in degrees Fahrenheit (1931–1960) with the longitude, $x_1$, and latitude, $x_2$, of 56 US cities. The longitude of the US cities is measured in degrees west of the prime meridian and the latitude in degrees north of the equator. The data file can be found at http://lib.stat.cmu.edu/DASL/Stories/USTemperatures.htm. This page also includes partial regression plots of latitude and longitude. They show that the relationship between January temperature and latitude, after removing the effects of longitude, is linear and negative. However, after removing the effects of latitude, the relationship between January temperature and longitude is cubic polynomial. This conclusion will be confirmed by using the formal regional residual plots and in addition, one will be able to locate the lack-of-fit with respect to longitude.

Consider a linear model in longitude and latitude

$$y = \beta_{00} + \beta_{10} x_1 + \beta_{01} x_2 + e. \quad (11)$$
Fig. 4. (a) Windmill data (Montgomery and Peck, 1992); $y =$ direct current (DC) output; $x =$ wind velocity (miles/h); (b) formal regional residual plot; dark gray areas indicate an overestimation in the low- and high-range; light gray areas an underestimation of the data points in the mid-range.

Fig. 5. (a) Windmill data; $y =$ direct current (DC) output; $x =$ reciprocal transformation on wind velocity; (b) formal regional residual plot.

Fig. 6 shows the US Temperature data and the fitted plane. The percentiles of the test statistic were approximated using 100,000 bootstrap samples drawn from the classical residuals, resulting in a critical value of 3.94 at the $\alpha = 0.05$ significance level. The calculated values of the test statistics $T_{1,S^2}$ and $T_{2,S^2}$ from the data sample are 5.82 and 3.60, respectively. Thus, $T = \max \left( T_{1,S^2}, T_{2,S^2} \right) = 5.82$, which corresponds to a $p$-value of 0.000. A clear lack-of-fit is detected and the marginal regional residual plots for the two predictor variables, longitude and latitude (Fig. 7(a) and (b), respectively), can be used to localize this lack-of-fit.
Fig. 6. US temperature data (Peixoto, 1990) $y$ = average January minimum temperature in degrees Fahrenheit (1931–1960); $x_1$ = longitude of the US city in degrees west of the prime meridian; $x_2$ = Latitude in degrees north of the equator.

Fig. 7. US Temperature data: (a) Marginal regional residual plot for longitude; the dark gray areas show that the underestimation of the data in the high-range of longitude is statistically significant; (b) marginal regional residual plot for latitude; no regions of lack-of-fit are found.

No significant lack-of-fit is found in the marginal regional residual plot of latitude, which confirms the earlier stated linear relationship between January temperature and latitude. However, there is a clear lack-of-fit detected for the variable longitude. Fig. 7(a) shows that the underestimation of the data in the high-range of longitude is statistically significant. There can be concluded that the relationship between January temperature and longitude is not linear.
The solution proposed by Peixoto (1990), a cubic polynomial in longitude,
\[ y = \beta_{00} + \beta_{10}x_1 + \beta_{01}x_2 + \beta_{11}x_1x_2 + \beta_{20}x_1^2 \\
+ \beta_{21}x_1^2x_2 + \beta_{30}x_1^3 + \beta_{31}x_1^3x_2 + \epsilon \quad (12) \]
results in marginal regional residuals plots that display no lack-of-fit. Both values, \( T_{1,S^2} = 3.23 \) and \( T_{2,S^2} = 3.07 \), are smaller than the critical value 3.59 \( (p = 0.18) \). This confirms that model (12) accurately predicts the average January minimum temperature.

5. Conclusions

Different lack-of-fit tests and corresponding regional residual plots are proposed to assess the fit of both simple and multiple linear regression models. Simulations in simple linear regression strongly suggest that the power of the proposed testing procedures are at least comparable to the power of popular classical methods. With the Rice variance estimator \( \hat{\sigma}_R^2 \) good empirical power is obtained for alternatives with both global and local lack-of-fit. This test seems to behave similarly as the KH test, except for cases with local lack-of-fit, where the proposed test performs even better. A major advantage of the new procedures is the ability to locate lack-of-fit in a formal graphical way. Even in case of violations of the model assumption of homoscedasticity the new tests still behave well compared with other classical tests. The use of the wild bootstrap is recommended in practice, as it handles adequately heteroscedasticity and nonnormality of the error terms.

References