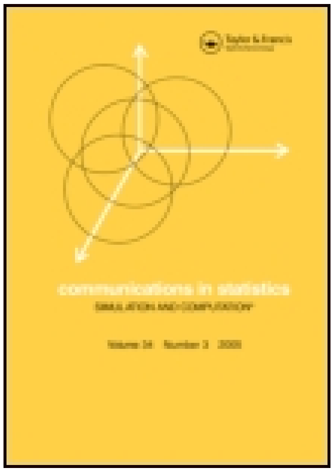


Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Checking Normality and Homoscedasticity in the General Linear Model Using Diagnostic Plots

A. Schützenmeister [a] , U. Jensen [b] & H.-P. Piepho [a]

[a] Bioinformatics Unit, Institute of Crop Science , University of Hohenheim , Stuttgart , Germany

[b] Institute of Applied Mathematics and Statistics , University of Hohenheim , Stuttgart , Germany

Published online: 07 Oct 2011.

To cite this article: A. Schützenmeister , U. Jensen & H.-P. Piepho (2012) Checking Normality and Homoscedasticity in the General Linear Model Using Diagnostic Plots, Communications in Statistics - Simulation and Computation, 41:2, 141-154, DOI: 10.1080/03610918.2011.582560

To link to this article: http://dx.doi.org/10.1080/03610918.2011.582560

Communications in Statistics—Simulation and Computation®, 41: 141–154, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2011.582560

Checking Normality and Homoscedasticity in the General Linear Model Using Diagnostic Plots

A. SCHÜTZENMEISTER[1], U. JENSEN[2], AND H.-P. PIEPHO[1]

[1]Bioinformatics Unit, Institute of Crop Science,
University of Hohenheim, Stuttgart, Germany
[2]Institute of Applied Mathematics and Statistics,
University of Hohenheim, Stuttgart, Germany

Inference for the general linear model makes several assumptions, including independence of errors, normality, and homogeneity of variance. Departure from the latter two of these assumptions may indicate the need for data transformation or removal of outlying observations. Informal procedures such as diagnostic plots of residuals are frequently used to assess the validity of these assumptions or to identify possible outliers. A simulation-based approach is proposed, which facilitates the interpretation of various diagnostic plots by adding simultaneous tolerance bounds. Several tests exist for normality or homoscedasticity in simple random samples. These tests are often applied to residuals from a linear model fit. The resulting procedures are approximate in that correlation among residuals is ignored. The simulation-based approach accounts for the correlation structure of residuals in the linear model and allows simultaneously checking for possible outliers, non normality, and heteroscedasticity, and it does not rely on formal testing.

[Supplementary materials are available for this article. Go to the publisher's online edition of Communications in Statistics—Simulation and Computation® *for the following three supplemental resource: a word file containing figures illustrating the mode of operation for the bisectional algorithm, QQ-plots, and a residual plot for the mussels data.]*

Keywords Constant residual variance; Model checking; Monte Carlo; Multiplicity; Residual plot; Simulation; Simultaneous tolerance band; Tolerance interval; Tolerance region.

Mathematics Subject Classification 62-09; 62F25; 62J5; 62J10.

1. Introduction

A common approach to checking assumptions of the general linear model is to compute residuals and either produce various residual plots, or to subject

Received September 16, 2010; Accepted April 15, 2011

Address correspondence to H.-P. Piepho, Bioinformatics Unit, Institute of Crop Science, University of Hohenheim, Fruwirth-strasse 23, Stuttgart 70599, Germany; E-mail: piepho@uni-hohenheim.de

these to tests of normality and variance homogeneity (homoscedasticity). These procedures strictly assume that residuals have the same distributional properties as the true errors, which is always an approximation, because residuals are linear combinations of the true errors and so are stochastically dependent and may also be heteroscedastic, e.g., in simple linear regression. Least squares estimation of linear models with independent and identically distributed (i.i.d.) errors always results in some non zero covariances between pairs of residuals. This is a consequence of having n residuals, which carry only $(n - p)$ degrees of freedom, where n is the number of observations and p is the rank of the design/model matrix $\mathbf{X}$ (Draper and Smith, 1998, p. 206).

Moreover, the residuals may exhibit supernormality, i.e., the residuals appear to be more normal than the underlying distribution of errors if this is non normal (Atkinson, 1985). This characteristic can directly influence the outcome of statistical tests as well as the interpretation of diagnostic plots for normality or homoscedasticity. Furthermore, when interpreting diagnostic plots, there is always an unavoidable element of subjectivity.

Inference for linear models may be non-robust against violations of both the normality and homoscedasticity assumptions. Bradley (1980, 1984) showed that even for a large number of observations the inference drawn from F-tests and t-tests can be misleading when both assumptions are violated simultaneously, although they are usually robust against violations of only a single assumption in case of a sufficient sample size. Our approach allows assessing both assumptions simultaneously using the same set of simulation results.

Exploiting the fact that studentized residuals are pivotal statistics (Cox and Hinkley, 1974, p. 211; Dufour et al., 1998), the null distribution of a particular set of residuals as well as the null distribution of any test statistic computed from these residuals can be simulated. Piepho (1996a) used studentized residuals to construct a simulation-based test for homoscedasticity within the linear model framework. Dufour et al. (1998) used the same idea and compared eleven normality tests in terms of size and power with their Monte Carlo-based counterparts in linear regressions. The authors showed that the size of these tests is more precisely controlled when p-values are computed by their Monte Carlo (MC) procedure. In the same vein, Atkinson (1981, 1985) suggested computing envelopes in quantile-quantile (QQ)-plots, which are basically simulation-based point-wise tolerance intervals (TI) for each residual. Plotting these envelopes gives the user a general idea how severe potential departures from the assumptions are, e.g., in QQ-plots. Atkinson (1981) simulated a rather small number of data vectors ($N = 19$).

In this article, we propose a simulation-based graphical procedure for checking the normality and homoscedasticity assumptions, which takes into account that residuals may be correlated and heteroscedastic even when the underlying assumptions are met for the errors. We further develop the ideas of Atkinson's envelopes (1981, 1985; Atkinson and Riani, 2000) and Piepho's (1996a) MC test for variance homogeneity. In particular, we show how results of the MC procedure can be used to construct $100(1 - \alpha)\%$ simultaneous tolerance bands (STB). These STBs help to interpret diagnostic plots for normality and homoscedasticity, objectify their interpretation, and also provide asymptotically valid level-α tests.

This article is organized as follows. We start in Sec. 2 with a small example from metabolite profiling (Römisch-Margl et al., 2010), which exemplifies the problems an experimenter faces in interpreting diagnostic plots. In Sec. 3, we present the general idea underlying our procedures. We proceed in Sec. 4 with the construction of the $100(1-\alpha)\%$ STB for normality. In Sec. 5, we present methods for checking homoscedasticity and the identification of outlying observations based on our MC procedure. These methods are exemplified using a previously published dataset.

2. Motivating Example

Römisch-Margl et al. (2010) performed extensive measurements of metabolites in the early stages of the developing maize kernel. They aimed at investigating heterotic patterns of dry matter, starch, sugars, sugar-phosphates, and free amino acids for the B73 × Mo17 hybrid and its parental lines at six developmental stages (8, 12, 16, 20, 25, 30 past pollination). We consider the fructose measurements in the whole kernel at eight days past pollination. Interest was in the differences among genotypes. For this set-up we use the linear model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \tag{1}$$

where μ is the general mean, α_i is the effect of the ith genotype ($i = 1, \ldots, k$), $e_{ij} \sim N(0, \sigma^2)$ is the i.i.d. residual error of the jth observation for the ith genotype ($j = 1, \ldots, n_i$), $n = n_1 + n_2 + \cdots + n_k$ and y_{ij} is the ijth measured metabolic quantity. The standard procedure for checking normality would consist of fitting the model, extracting studentized residuals, and constructing a QQ-plot, as shown in Fig. 1(a). This QQ-plot shows an increasing volatility towards both ends, and it is not clear whether this is within expectation based on the properties of order statistics, or indication of real departure from assumptions. In particular, it is not clear, whether there are any outlying observations. This illustrates the general problem with QQ-plots for a user in deciding whether the pattern of points is indicative of departure from normality or not. The same problem occurs with other residual plots. For this reason, it would be useful to have tolerance bands (TB) such that a QQ-plot can

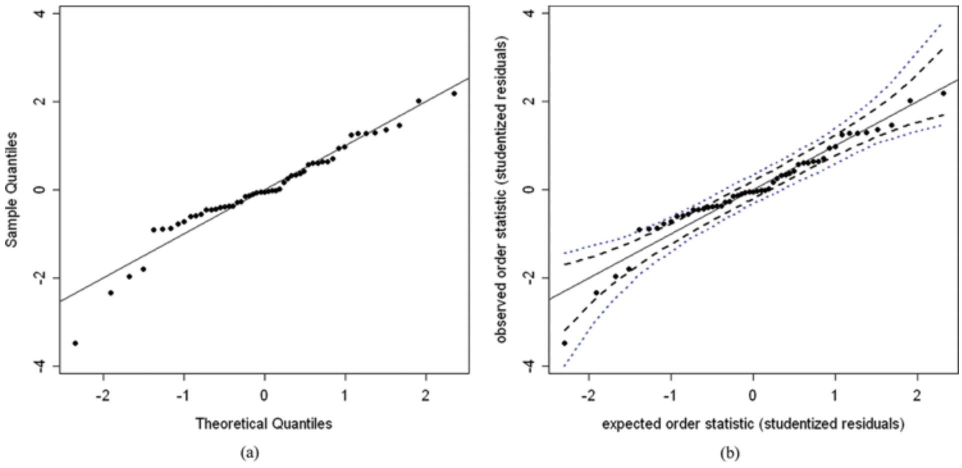


Figure 1. Residual plots of studentized residuals obtained from fitting model (1) to a subset of the metabolite data described in Römisch-Margl et al. (2010). (a) ordinary quantile-quantile plot (QQ-plot). (b) QQ-plot with point-wise 95% tolerance band (TB, dashed lines) and 95% Bonferroni-adjusted TB (dotted lines). (color figure available online)

be judged acceptable whenever all plotted quantiles for the residuals are inside the band. This idea is similar to the envelopes suggested by Atkinson (1981, 1985) for half-normal plots. Atkinson only considers control of the point-wise α level. We here propose to use a simultaneous tolerance band (STB) which has simultaneous coverage probability $(1-\alpha)$.

Our approach is based on the simulation of N datasets, that have the same size (n), the same correlation structure, and the same design matrix $\mathbf{X}$ as the observed data. For each simulated dataset we compute residuals and order them by size. For the ith order statistic there are N simulated residuals. Among these, we compute the $(\alpha/2)-$ and $(1-\alpha/2)-$ quantiles to obtain a $100(1-\alpha)\%$ tolerance interval. These quantiles are denoted here as local. If $N \to \infty$, the local interval attains exact coverage. Note, however, that it controls only the point-wise coverage probability, not the simultaneous coverage probability (see Fig. 1(b), dashed lines). To account for multiplicity, bounds of these intervals could be corrected, e.g., by Bonferroni adjustment where instead of the $(\alpha/2)-$ and $(1-\alpha/2)-$quantiles of the ith order statistic the $(\alpha/2n)-$ and $(1-\alpha/2n)-$quantiles are used, respectively. Bonferroni adjustment guarantees that the simultaneous coverage probability is greater than or equal to $(1-\alpha)$. By the Bonferroni method, each local error level γ is assigned the same value $\gamma = \alpha/n$, which results in the characteristic form of the STB familiar from regression. An example is shown in Fig. 1(b) (dotted lines, 96.77% coverage). The Bonferroni method is known to be conservative (96.77% coverage instead of near 95%), while the point-wise $(1-\alpha)$ TB is far too liberal (32.98% coverage). Some improvement is therefore desirable. Specifically, an improved procedure to compute more narrow STBs compared to the Bonferroni method is required, that accounts for dependencies among residuals. Our proposed method accomplishes that.

3. Outline of Approach to Model Checking

3.1. *Residuals*

The general linear model, written in standard matrix notation, has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2}$$

where $\mathbf{y}$ is the vector of observed values, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{X}$ is the design/model matrix which corresponds to $\boldsymbol{\beta}$, and $\mathbf{e}$ is a vector of residual errors. The null hypothesis to be tested is that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, which is one prerequisite for standard analysis by the general linear model. Departure from this assumption may hint at outlying observations which should be removed prior to analysis, or there may be heteroscedasticity or non-normality of $\mathbf{e}$, which might be avoided by a suitable data transformation.

Our proposed MC procedure makes use of the ordinary least squares (OLS) residuals $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T$ is the hat matrix. Studentized residuals are computed by

$$\tilde{e}_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, \quad i = 1, \ldots, n, \tag{3}$$

where h_{ii} is the ith diagonal element of $\mathbf{H}$ and

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-\text{rank}(\mathbf{X})} = \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{n-\text{rank}(\mathbf{X})}. \tag{4}$$

The expression $\hat{\sigma}\sqrt{1-h_{ii}}$ is the ith diagonal element of the estimate of the variance-covariance matrix of residuals $var(\mathbf{e}) = (\mathbf{I}-\mathbf{H})\sigma^2$. There are n elements in the vector of observed residuals $\hat{\mathbf{e}}$, which carry only $(n-p)$ degrees of freedom, where p equals rank($\mathbf{X}$). Thus, there are always non zero pair-wise covariances in the variance-covariance matrix $(\mathbf{I}-\mathbf{H})\sigma^2$ (Draper and Smith, 1998, p. 206). Studentized residuals all have unit variance, but unfortunately do not follow Student's t-distribution (Atkinson and Riani, 2000, p. 18). While there are some transformations to univariate t-distributions or beta-distributions of individual residuals (Beckman and Trussell, 1974; Csorgo et al., 1973; Seber and Lee, 2003, p. 267), the joint distribution of residuals is more difficult to handle. We therefore use simulation to obtain the joint distribution of studentized residuals. We here use internally studentized residuals, but one might as well use externally studentized (leave-one-out) residuals (Atkinson, 1985). To simulate the null distribution of studentized residuals for a particular linear model, we compute $\hat{\mathbf{e}}^{\text{MC}} = (\mathbf{I}-\mathbf{H})\mathbf{y}^{\text{MC}}$, where $\mathbf{y}^{\text{MC}}$ is a simulated data vector, and apply (3). Because studentized residuals are pivotal quantities, without loss of generality elements of the random normal vector $\mathbf{y}^{\text{MC}}$ can be drawn from a standard normal distribution $N(0, 1)$. Repeating this step N times results in N simulated sets of studentized residuals (each of size n). When applying formula (3), it is important to re-estimate the residual variance in each simulation. This accounts for the uncertainty in the observed variance when computing the simultaneous tolerance bounds. Otherwise, these bounds could become too narrow, i.e., too liberal. In the following we will mainly suppress the superscript MC when we refer to the vector of simulated residuals whenever it is clear that we use MC residuals.

3.2. *Graphical Methods Based on Studentized Residuals*

We consider three major graphical applications of our approach. The first application aims at facilitating the interpretation of QQ-plots by computing an STB, which simultaneously covers all points with a previously specified probability $(1-\alpha)$. Departure from normality can be detected easily even by the less trained eye if this STB is added to a QQ-plot. The second application aims at checking homoscedasticity and at identifying outlying observations by adding a simultaneous tolerance interval (STI) to residual plots. The third graphical application is designed to assess whether the residual variance is independent of predicted values. It is common that the residual variance increases for increasing predicted values, e.g., in linear regression. Therefore, we regress absolute values (or squares) of studentized residuals on predicted values obtaining N regression lines, where each point on a regression line refers to a specific predicted value of the original data (row in $\mathbf{X}$). This set of regression lines can be used to compute a $100(1-\alpha)\%$ STB.

All three diagnostic/informal procedures rely on an appropriately high number of MC simulations, which, to our experience, should be greater than or equal to 5000. Each vector of studentized MC residuals $\tilde{\mathbf{e}}_{\mathbf{j}}$ $(j = 1, \ldots, N)$ can be ordered to obtain its order statistics, which are denoted for the jth residual vector as $\tilde{e}_{(j,1)} \leq \tilde{e}_{(j,2)} \leq, \ldots, \leq \tilde{e}_{(j,n)}$. Across all N vectors of order statistics, the minima correspond to the set $\{\tilde{e}_{(1,1)}, \ldots, \tilde{e}_{(N,1)}\}$, the maxima correspond to the set $\{\tilde{e}_{(1,n)}, \ldots, \tilde{e}_{(N,n)}\}$. These sets of minima and maxima will be used to construct an STI which can easily be added to ordinary residual plots for checking the homoscedasticity assumption and to identify outlying observations (Sec. 5, 2nd graphical application). In order to check normality and to assess whether the residual variance is independent of predicted values, we make use of all N vectors of order statistics as will be detailed in Sec. 4.1 and Sec. 5 (3rd graphical application).

4. Checking Normality

Consider Fig. 1(b) as an example, where the point-wise 95% TB (dashed lines) is plotted together with the Bonferroni-corrected STB (dotted lines). The point-wise TB (32.98% coverage) results in ten studentized residuals that exceed its bounds, whereas one residual exceeds the bounds of the Bonferroni-corrected STB (96.77% coverage). The exact $100(1-\alpha)\%$ STB (95.00% coverage) would be located in between the liberal point-wise TB and the conservative Bonferroni-corrected STB.

To compute an approximate $100(1-\alpha)\%$ STB, we propose to use a bisection algorithm to adapt the point-wise tolerance level γ in order to achieve joint coverage of approximately $100(1-\alpha)\%$ for all N vectors of studentized residuals. For the kth iteration the bisection algorithm (Press et al., 1989, p. 277) can be outlined as follows (Initialization: $\gamma_0 = \alpha$, $\gamma_1 = \alpha/2$).

1. Compute local $100(1-\gamma_k)\%$ tolerance intervals for each quantile of the order statistic among all N values, i.e., the ith local interval is $[Q^i_{100(\gamma_k/2)\%};\ Q^i_{100(1-\gamma_k/2)\%}]$, $i = 1, \ldots, n$, where γ_k is the point-wise nominal tolerance level of the kth iteration, $Q^i_{100(\gamma_k/2)\%}$ and $Q^i_{100(1-\gamma_k/2)\%}$ are the $100(\gamma_k/2)\%$ and $100(1-\gamma_k/2)\%$ sample quantiles for the ith order statistic.
2. Compute the value m/N (coverage), where m is the number of studentized residual vectors located entirely within the area defined by the point-wise tolerance intervals, which constitute the STB, i.e., none of their elements exceeds these bounds.
3. The algorithm terminates if:

   a. $\delta \in [0;\ \varepsilon]$, $\delta = m/N - (1-\alpha)$, where ε is a previously defined convergence tolerance, or

   b. the previously specified maximum number of iterations is reached. In this case, that γ_k is used which minimizes $\delta = m/N - (1-\alpha)$, $\delta > 0$.

   If neither condition (a) nor condition (b) is fulfilled, compute an updated γ_k by

$$\gamma_{k+1} = \begin{cases} \gamma_k - \dfrac{|\gamma_k - \gamma_{k-1}|}{2}, & \text{if } \dfrac{m}{N} - (1-\alpha) < 0 \\ \gamma_k + \dfrac{|\gamma_k - \gamma_{k-1}|}{2}, & \text{if } \dfrac{m}{N} - (1-\alpha) > 0, \end{cases}$$

   go to step 1 and proceed with iteration $(k+1)$.

Here, we used a simple bisection algorithm for finding appropriate local tolerance limits. Certainly, other algorithms could have been used to achieve the same goal. The key-feature of the algorithm is to check the coverage in each iteration and to adapt the local tolerance level γ accordingly. See supplemental material for a plot depicting the iterative approach to the final value of γ (Fig. S1).

Figure 2(a) depicts a possible graphical display of the $100(1-\alpha)\%$ STB for the metabolite data from Sec. 2, calculated with the bisection algorithm. The simulated coverage for this example was 95.02%. In fact, our procedure provides a valid level-α test for normality, if we reject normality whenever at least one point exceeds the bounds of the $100(1-\alpha)\%$ STB. In this setup, the STB represents the acceptance region of the null hypothesis. We would like to stress that the main purpose of the STB is to provide assistance in interpreting residuals plots, and that availability of a valid level-α test is simply a welcome by-product of the way our STB is constructed. For the example shown in Fig. 2(a), we would accept normality, despite one point violating the bounds of the STB (▲). It is located mid-range and the remaining points do not point to non-normality, i.e., the overall point-pattern nicely fits inside the 95.02% STB.

The construction of the $100(1-\alpha)\%$ STB benefits from a higher number of simulations, such that the coverage probability becomes exactly $100(1-\alpha)\%$ on average for $N \to \infty$. The smoothness of the $100(1-\alpha)\%$ STB also increases with N. Figure 2(b) depicts the effect of an inadequately small number of simulations ($N = 250$). The coverage for this example (96.4%) is higher than for the first example in Fig. 2(a) (95.02%), where $N = 10{,}000$ simulations were performed. Contrarily, there are more residual points exceeding the bounds of the STB in Fig. 2(b) compared to Fig. 2(a). This exemplifies the need of performing appropriately many simulations to approximate the joint null distribution of diagnostic features (studentized residuals here) in order to avoid misleading conclusions. See the supplemental material section for additional examples, where the number of simulations used for constructing the $100(1-\alpha)\%$ STB was varied (Fig. S2).

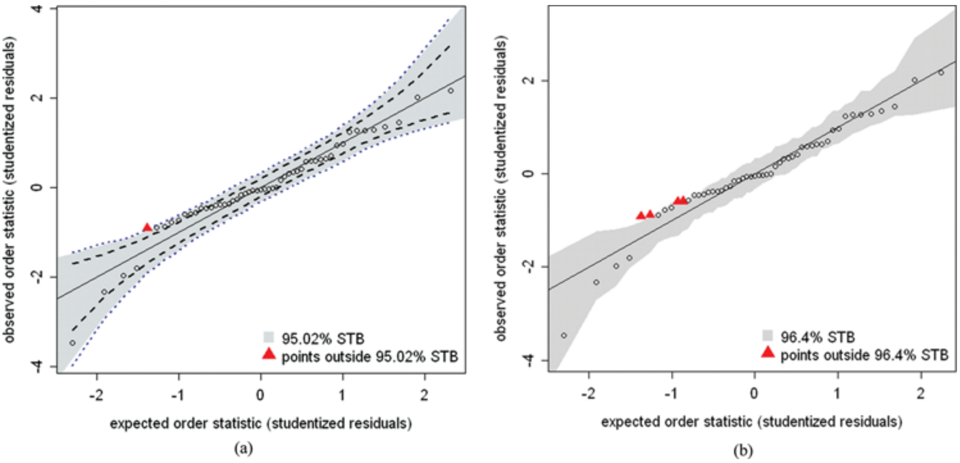


Figure 2. Simultaneous tolerance bands (STB) for studentized residuals of the metabolite data. (a) 95.02% STB computed from $N = 10{,}000$ simulations. There is a single residual exceeding these bounds which is indicated as triangle (▲). Dashed lines correspond to the point-wise 95% tolerance band, dotted lines correspond to the Bonferroni-corrected 95% tolerance band. (b) 96.4% STB computed form $N = 250$ simulations, with four residuals exceeding these bounds (▲). (color figure available online)

5. Homoscedasticity and Outliers

Here, we make use of a dataset, which comprises 82 measurements of mussels from New Zealand (Atkinson and Riani, 2000, p. 116, citing Cook and Weisberg, 1994, p. 161). Here, we consider the linear regression of variable M (mass of a mussel) onto variable S (mass of a mussel's shell), denoted as $M \sim S$. Often, one observes an increasing variance of residuals with increasing predicted values. This can be seen in the residual plot (Fig. 3(a)), where studentized residuals were plotted against predicted values. The regression of absolute values of studentized residuals onto predicted values makes this obvious, since the regression line has a relatively high positive slope (Fig. 3(b)).

We consider two graphical procedures to check homoscedasticity. One uses a $100(1-\alpha)\%$ STI, which can be added to residual plots, the other one makes use of a $100(1-\alpha)\%$ STB for the regression line obtained from the regression of absolute values of studentized residuals onto predicted values.

For the first graphical procedure we make further use of the set of N simulated samples to construct a $100(1-\alpha)\%$ STI. Any residuals not falling into this interval would be indicative of heteroscedasticity or could be outliers. Examples are shown in Figs. 3(a)–5(a), where points highlighted as triangles (▲) correspond to residuals falling outside the QQ-plot with $100(1-\alpha)\%$ STB for normality (Supplemental material, Fig. S3). The horizontal (dashed) lines in these plots correspond to the interval

$$[Q_{100(\gamma/2)\%}; Q_{100(1-\gamma/2)\%}],$$

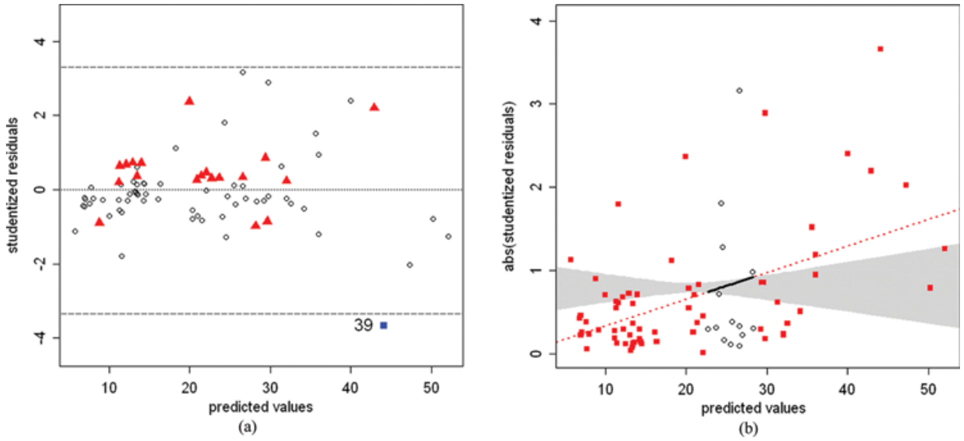


Figure 3. Residual plots for the regression $M \sim S$ of the mussels data (Atkinson and Riani, 2000, p. 116). (a) Plot of studentized residuals vs. predicted values with $100(1-\alpha)\%$ simultaneous tolerance interval (STI) for homoscedasticity and outlier detection, points highlighted as squares (■) fall outside the STI, points highlighted as triangles (▲) fall outside th 95.00% STB for normality in the QQ-plot (Supplemental martial, Fig. S3a); (b) Plot showing the regression of absolute values of studentized residuals on predicted values with $100(1-\alpha)\%$ STB for homoscedasticity. Squares (■) correspond to residuals, where the regression line for the observed data is located outside the 95.01% STB (dotted). (color figure available online)

where $Q_{100(\gamma/2)\%}$ and $Q_{100(1-\gamma/2)\%}$ are the bounds of the $100(1-\alpha)\%$ STI. We numerically search for a tolerance level γ such that at least $100(1-\alpha)\%$ of all simulated, studentized residual vectors are enclosed by these bounds, i.e., $100\alpha\%$ of all vectors have at least one residual falling outside. To achieve this, we simply make use of the sets of minimum and maximum residuals $\{\tilde{e}_{(1,1)}, \ldots, \tilde{e}_{(N,1)}\}$ and $\{\tilde{e}_{(1,n)}, \ldots, \tilde{e}_{(N,n)}\}$, i.e., 1st and nth order statistics, taken from the set of MC studentized residual vectors $\tilde{\mathbf{e}}_j (j = 1, \ldots, N)$ (see Sec. 3.2), and apply the bisection algorithm (see Sec. 4.1). Figure 4(a) depicts the residual plot with 95.00% STI of the regression model $\log(M) \sim \log(S)$. There are two residual points outside the STI (observations 8 and 48), which are indicated as asterisk ($*$). They are also located outside the $100(1-\alpha)\%$ STB for normality (Supplemental material, Fig. S3b). Since both points exceed the bounds of the $100(1-\alpha)\%$ STB for normality and the $100(1-\alpha)\%$ STI for homoscedasticity, they can be considered as truly outlying.

The second graphical procedure for checking homoscedasticity is based on the regression of absolute (or squared) values of studentized residuals onto predicted values of the linear model. For the kth simulated dataset we store the predicted values of the regression $abs(\tilde{\mathbf{e}}_k) \sim \hat{\mathbf{y}}_k$ in row k of the $(N \times n)$ matrix $\mathbf{P}$. Each column of $\mathbf{P}$ corresponds to a specific predicted value for the original data and is associated with a specific row in the design/model matrix $\mathbf{X}$. In the simple linear regression set-up considered in the mussels example, each shell mass value is assigned a specific predicted value of the response variable (mass). Thus, over N simulations we obtain N sets of predicted values (rows in $\mathbf{P}$), where each element is due to the regression of absolute (or squared) values of studentized residuals onto predicted values $\hat{\mathbf{y}}$ of the linear model. We then apply the bisection algorithm from Sec. 4.1 to compute a

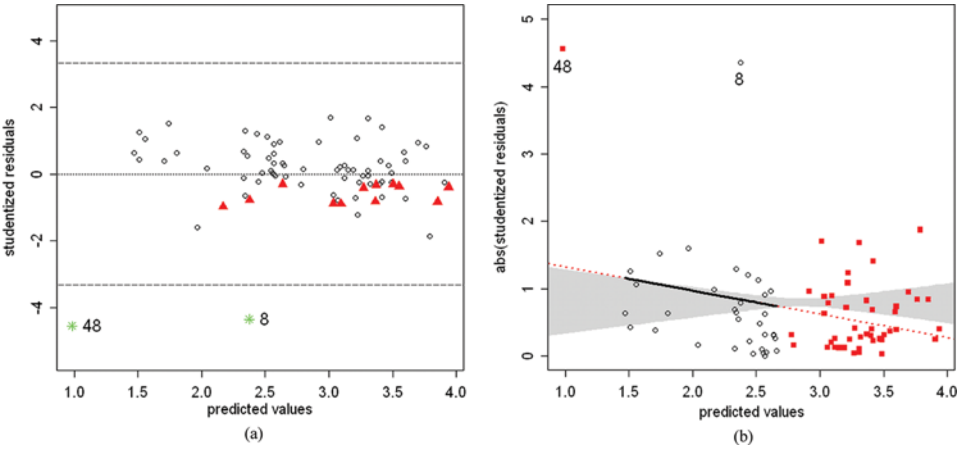
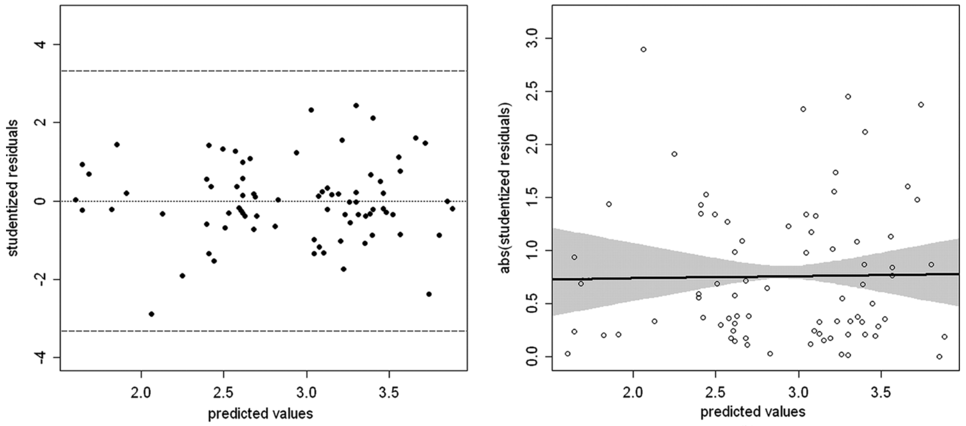


Figure 4. Residual plots for the regression $\log(M) \sim \log(S)$ of the mussels data (Atkinson and Riani, 2000, p. 116). (a) Plot of studentized residuals vs. predicted values with $100(1-\alpha)\%$ simultaneous tolerance interval (STI) for homoscedasticity and outlier detection, points highlighted as asterisk ($*$) fall outside of this STI as well as outside of the STB for normality (Supplemental material, Fig. S3b), points highlighted as triangles (▲) only fall outside the STB for normality. (b) Plot of the regression of absolute values of studentized residuals on predicted values with 95.00% STB for homoscedasticity. Squares (■) correspond to residuals, where the regression line is located outside the STB (dotted). (color figure available online)

$100(1-\alpha)\%$ STB for the regression line, i.e., a local tolerance interval is assigned to each predicted value (row in $\mathbf{X}$). This STB encloses $100(1-\alpha)\%$ of all N regression lines and can be added to the plot of $abs(\tilde{\mathbf{e}}) \sim \hat{\mathbf{y}}$. This gives an informative plot regarding homoscedasticity of the residuals as shown in Figs. 3(b)–5(b). We plotted residuals as squares (■), when they belong to those parts of the regression line located outside the STB, shown as dotted line. Two residual points, which are located outside the STI in Fig. 4(a), appear to have a major impact on the fitted line. Observation 48 is responsible for the large negative slope of the regression line in Fig. 4(b). Observations 48 and 8 both shift the regression line towards zero, since their relatively large values downsize the remaining residuals. This is due to the fact that the variance of studentized residuals has an expected value equal to one, and both residuals account for a substantial proportion of this variance.

For the mussels data we started with the complete dataset and the regression model $M \sim S$, which exhibited non normality of the studentized residuals as well as heteroscedastic residuals, since the variance of the residuals increases with increasing predicted values (Fig. 3). The residual points falling outside the STB for normality were located mid-range, which is a hint that something is wrong with the model (Supplemental material, Fig. S3a). Log-transformation remedied this partly, since the residual variance was stabilized (Fig. 4(a)). Subsequently removing the observations which correspond to the two largest residuals resulted in plots, which did not exhibit violations of either normality or homoscedasticity when the model $\log(M) \sim \log(S)$ was considered. None of the residuals are located outside the STI in the residual plot (Fig. 5(a)), and the fitted line for the regression of absolute values of studentized residuals onto predicted values is located entirely within the corresponding STB (Fig. 5(b)). Note that there is no indication of non-normality of the residuals either (Supplemental material, Fig. S3d).



(a)

(b)

Figure 5. Residual plots for the regression $\log(M) \sim \log(S)$ of the mussels data (Atkinson and Riani, 2000, p. 116), where observations 48 and 8 were removed. (a) Residual plot with $100(1-\alpha)\%$ simultaneous tolerance interval (STI) for homoscedasticity and outlier detection. (b) Plot showing the regression of absolute values of studentized residuals onto predicted values with 95.05% STB for homoscedasticity.

6. Discussion

The proposed methodology has several applications within the general linear model framework. Our simulation-based approach facilitates the interpretation of informal/graphical procedures for model checking, which are usually preferred over formal tests. Diagnostic plots are routinely used in the analysis of linear models and most statistical software provides various diagnostic plots as standard output. Adding simultaneous tolerance bounds (STB, STI) to residual plots objectifies their interpretation with regard to assessing normality, homoscedasticity, and the presence of outliers. This allows even less experienced practitioners to assess such plots more easily and with more confidence. Since any interpretation of diagnostic plots is more or less subjective, our simultaneous tolerance bounds can be seen as a means for minimizing biases from subjective assessment.

QQ-plots can be accompanied by an STB to check for normality. Residuals from a fitted linear model are expected to be located entirely within this area with probability $100(1-\alpha)\%$ (see Sec. 4.1). Although an STB inherently defines a valid level-α test, we suggest to use the STB mainly as a tool in deciding about the assumed normality of the studentized residuals and not as formal test. If we are willing to use the 95.02% STB of Fig. 2(a) as level-α test, we would have to reject normality. However, we would accept normality in this case, because there is only one mid-range residual located slightly outside the STB. This residual is clearly not an outlier. Evidence against normality is minor, since the overall point-pattern fits nicely inside the STB.

Residuals, which are located outside of the STB and which are extreme, i.e., either the first or the last order statistic are likely to be outliers. If these points appear as outliers in the residual plot with STI for homoscedasticity as well, the corresponding observations should be removed. Otherwise, the STB helps in deciding how severe the departure from normality is.

The envelopes suggested by Atkinson (1981, 1985) were a first attempt to provide some guidance for the interpretation of residual plots. These envelopes used point-wise tolerance intervals and only few simulations. As shown in Sec. 2, point-wise coverage results in bounds, which are very liberal, since they do not account for multiplicity. The (joint) coverage of the point-wise TI for the metabolite data was equal to 32.98%, for example. The small number of simulations suggested by Atkinson ($N = 19$) results in very erratic bounds for the tolerance band, which gives only a very imprecise approximation of the joint null distribution of this diagnostic feature. We remedied the first of both problems by using the simultaneous coverage probability, thus accounting for multiplicity, and the second problem by using a sufficiently large number of simulations. We suggest using at least $N = 5{,}000$ simulations, which, to our experience, should result in a sufficiently well approximated null distribution of diagnostic features, e.g., a QQ-plot of studentized residuals. In order to decide whether the number of simulations were sufficiently large in a given case, it is helpful to repeat the MC approach with the same number of simulations. In case that the simultaneous tolerance bounds (STI, STB) are similar on both simulations, N can be deemed sufficiently large. Otherwise, we recommend to increase N. Our method benefits from a larger number of simulations, because exact coverage probability $100(1-\alpha)\%$ is approached as $N \rightarrow \infty$, making the STB less erratic and less conservative. Since modern computers become rapidly faster, computational costs will not be problematic even for larger values of N. Furthermore, MC-approaches are tailored for parallel computing.

The second graphical tool was an STI to assess homoscedasticity as shown in Sec. 5. The bounds of this interval are most informative, when added to plots of residuals vs. predicted values. Frequently, such plots exhibit an increasing variance of the residuals for increasing predicted values. This interval should give some assistance in judging how severe this is. It can also be thought of as a tool to identify possible outliers, since suspiciously deviant residuals are likely to fall outside this tolerance interval.

Moreover, we proposed an STB for plots of absolute or squared residuals vs. predicted values as a third graphical tool. If the residual variance increases with increasing predicted values, this plot makes this relation easily recognizable. It turned out to be sensitive for outlying observations as well, even in case the slope of the corresponding regression is not extreme compared to the null distribution of slopes. Large absolute (or squared) residuals shift the remaining residuals towards zero, which results in a regression line that violates the lower bound of the corresponding STB. This is obvious in the plot of the regression $\log(M) \sim \log(S)$ for the mussels data, where only observation 48 was removed (Supplemental material, Fig. S4). The slope of the regression of absolute values of studentized residuals onto predicted values is slightly negative. The regression line is moved towards zero, because observation 8 accounts for a large proportion of the total variance, thus shifting all other residuals towards zero. Along with the QQ-plot and the STB for normality, these types of plot can help in deciding whether a data transformation is needed or not and which residuals are suspiciously large compared to the null distribution of residuals.

All three graphical procedures account for the correlation structure of the residuals, which is due to having n residuals and only $(n - p)$ residual degrees of freedom (Draper and Smith, 1998). In complex experimental designs with large number of free parameters (p), this may have severe impact on the residual analysis. There exist alternative approaches which also account for the correlation structure. One could use $(n - p)$ orthogonal residuals, e.g., linear unbiased residuals (LUS) or best linear unbiased residuals (BLUS), which are uncorrelated by construction. Note that LUS residuals are identical to *recursive residuals* in case a Cholesky factorization is used to construct them (Cook and Weisberg, 1982, pp. 34–36). Seber (1977, p. 172, 310) mentioned orthogonal residuals, which are close to BLUS. There are at least three disadvantages using orthogonal residuals. Firstly, the identification of residuals with observations becomes blurred (Cook and Weisberg, 1982), because orthogonal residuals are linear combinations of the observed OLS residuals. Second, we must expect an amplification of the *supernormality* effect (Atkinson, 1985). Finally, some power is lost in detecting non-normality and/or heteroscedasticity when using only $(n - p)$ orthogonal residuals. The latter was confirmed when we applied the Shapiro-Wilk normality test once to orthogonal residuals, once to studentized residuals, and once as MC test. For seven experimental designs with different complexity, it had the best empirical power when it was applied as MC test, studentized residuals came second, and orthogonal residuals yielded the smallest power (results not shown). Four other tests of normality (Anderson-Darling, Shapiro-Francia, Cramer-von Mises, Lillifors-Kolmogorov-Smirnov), which are reviewed in Thode (2002), all part of the R-package nortest, revealed that MC-testing should be preferred over testing only the set of raw or studentized residuals. The more complex an experimental design, the more power is gained by using MC tests.

The MC set-up may also be used to perform tests for homoscedasticity, e.g., the Levene-test (Levene, 1960; Piepho, 1995a, 1996a) or tests based on the F-statistic for non zero slope in the regression of absolute values (squares) of studentized residuals onto predicted values. We expect the same gain in power as for the normality tests when applying them as MC-tests.

It should be mentioned that the graphical procedures as well as significance tests lack power when they are most needed, i.e., when samples are very small. Conversely, when sample size is very large, minute but irrelevant departures may become highly significant. It is in these instances that diagnostic plots are most helpful in deciding whether some departure is serious or minor.

We would like to stress that the proposed method only supplements the large tool box of strategies which check residuals of linear models for outliers and the several assumptions made, and that it does not, of course, solve all problems. For instance, it may fail to identify influential observations with high leverage which might cause fitting of an inadequate model. Such observations can be identified and removed by other well-known methods (Atkinson, 1985; Atkinson and Riani, 2000; Cook and Weisberg, 1982; Draper and Smith, 1998).

The focus of this article is on checking assumptions, which underlie linear model analysis. But both normality and homoscedasticity may be of interest in themselves. For example, when a specific crop is considered for planting, the farmer may want to know about the stability of this crop, i.e., the risk of the yield falling below a critical value. There are approaches to risk assessment, which are based on normality of crop yields across different environments (Bürkert et al., 2001; Eskridge, 1990; Piepho, 1998). Similar methods are used in quality control for manufacturing (Nelson, 1982; Meeker and Escobar, 1998). Moreover, variances may themselves be of intrinsic interest. For example when estimating inter-laboratory consensus values it is useful to include tests of homoscedasticity (Piepho, 1996b). In plant breeding, genotype-specific variances are often used to assess the phenotypic stability of crop cultivars (Denis et al., 1997; Piepho, 1995b, 1998).

All computations were carried out in the freely available statistical language R (http://cran.r-project.org) but it is not hard to implement our approach with other statistical packages, e.g., SAS (SAS Institute, Cary, North Carolina). For the future we plan to implement our MC procedure in a low-level programming language which should clearly increase the performance, i.e., reduce the computational time. This would allow using this method even for datasets which consist of many variables to be analysed separately as, e.g,. in metabolomic profiling.

Acknowledgments

We thank the anonymous reviewers for helpful comments. The last author acknowledges support within the DFG project PI 377/12-1.

References

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* 68:13–20.

Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford: Oxford University Press.

Atkinson, A. C., Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

Beckman, R. J., Trussell, H. J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association* 69:199–201.

Bradley, J. V. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society* 16:333–336.

Bradley, J. V. (1984). The complexity of nonrobustness effects. *Bulletin of the Psychonomic Society* 22:250–253.

Bürkert, A., Bationo, A., Piepho, H.-P. (2001). Efficient phosphorus application strategies for increased crop production in sub-Saharan West Africa. *Field Crops Research* 72:1–15.

Cook, R. D., Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Cook, R. D., Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.

Cox, D. R., Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Csorgo, M., Seshadri, V., Yalovsky, M. (1973). Some exact tests for normality in the presence of unknown parameters. *Journal of the Royal Statistical Society B* 35:507–522.

Denis, J.-B., Piepho, H.-P., van Eeuwijk, F. A. (1997). Modelling expectation and variance for genotype by environment data. *Heredity* 79:162–171.

Draper, N. R., Smith, H. (1998). *Applied Regression Analysis*. New York: Wiley.

Dufour, J.-M., Farhat, A., Gardiol, L., Khalaf, L. (1998). Simulation-based finite sample normality tests in linear regression. *Econometrics Journal* 1:154–173.

Eskridge, K. M. (1990). Selection of stable cultivars using a safety first rule. *Crop Science* 30:369–374.

Levene, H. (1960). Robust tests for equality of variances. In: Olkin, I., Ghurye, S. G., Hoeffding, W. G., Mann, H. B., eds. *Contributions to Probability and Statistics*. Essays in honor of Harold Hotelling. Stanford, CA: Stanford University Press, pp. 278–292.

Nelson, W. (1982). *Applied Life Data Analysis*. New York: Wiley.

Meeker, W. O., Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: Wiley.

Piepho, H.-P. (1995a). A robust test for homoscedasticity in a two-way layout. *Biometrical Journal* 37:151–160.

Piepho, H.-P. (1995b). Detecting, interpreting and handling heteroscedasticity in yield trial data. *Communications in Statistics B—Simulation and Computation* 24:243–274.

Piepho, H.-P. (1996a). A Monte Carlo test for variance homogeneity in linear models. *Biometrical Journal* 38:461–473.

Piepho, H.-P. (1996b). Weighted estimates of interlaboratory consensus values. *Computational Statistics & Data Analysis* 22:471–479.

Piepho, H.-P. (1998). Methods for comparing the yield stability of cropping systems. *Journal of Agronomy & Crop Science* 180:193–213.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1989). *Numerical Recipes in PASCAL*. Cambridge: Cambridge University Press.

Römisch-Margl, L., Spielbauer, G., Schützenmeister A., Schwab, W., Piepho, H.-P., Genschel, U., Gierl, A. (2010). Heterotic patterns of sugar and amino acid components in developing maize kernels. *Theoretical and Applied Genetics* 120:369–381.

Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.

Seber, G. A. F., Lee, A. J. (2003). *Linear Regression Analysis*. 2nd ed. New Jersey: Wiley.

Thode, H. C. (2002). *Testing for Normality*. New York: Marcel Dekker.