Global Validation of Linear Model Assumptions

Edsel A. Peña and Elizabeth H. Slate

An easy-to-implement global procedure for testing the four assumptions of the linear model is proposed. The test can be viewed as a Neyman smooth test and relies only on the standardized residual vector. If the global procedure indicates a violation of at least one of the assumptions, then the components of the global test statistic can be used to gain insight into which assumptions have been violated. The procedure can also be used in conjunction with associated deletion statistics to detect unusual observations. Simulation results are presented indicating the sensitivity of the procedure in detecting model violations under a variety of situations, and its performance is compared with three potential competitors, including a procedure based on the Box–Cox power transformation. The procedure is demonstrated by applying it to a new car mileage dataset and a water salinity dataset that has been used earlier to illustrate model diagnostics.

KEY WORDS: Box–Cox transformation; Deletion statistics; Model diagnostics and validation; Neyman smooth test; Outlier detection; Score test.

1. THE LINEAR MODEL AND ITS ASSUMPTIONS

One of the most important models in statistics is the linear model, in which the relationship between an observable $n \times 1$ response vector $\mathbf{Y}$ and an observable $n \times p$ design matrix $\mathbf{X}$ of predictor variables is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, σ is an unknown scale parameter, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of unobservable error variables. Conditionally on $\mathbf{X}$, $\boldsymbol{\epsilon}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$, the $n \times n$ identity matrix. This distributional assumption, together with the linear link specification in (1), are enumerated as four distinct assumptions:

(A1, *Linearity*) $\mathrm{E}\{Y_i|\mathbf{X}\} = \mathbf{x}_i\boldsymbol{\beta}$, where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$.

(A2, *Homoscedasticity*) $\mathrm{var}\{Y_i|\mathbf{X}\} = \sigma^2$, $i = 1, 2, \dots, n$.

(A3, *Uncorrelatedness*) $\mathrm{cov}\{Y_i, Y_j|\mathbf{X}\} = 0$ $(i \neq j)$.

(A4, *Normality*) $(Y_1, Y_2, \dots, Y_n)|\mathbf{X}$, have a multivariate normal distribution.

Assumptions (A3) and (A4) imply that, given $\mathbf{X}$, $Y_i, i = 1, 2, \dots, n$, are independent normal random variables. Without loss of generality, we assume that $\mathbf{X}$ is of full rank with $n > p$, so that $\mathrm{rank}(\mathbf{X}) = p$. Under (A1)–(A4), the maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}$ and σ^2 are given by

$$\mathbf{b} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{Y} \qquad \text{and} \qquad s^2 = \hat{\sigma}^2 = \frac{1}{n}\mathbf{Y}^{\mathrm{t}}(\mathbf{I} - \mathbf{P}[\mathbf{X}])\mathbf{Y}, \tag{2}$$

where $\mathbf{P}[\mathbf{X}] = \mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}$ is the projection operator on the linear subspace generated by the columns of $\mathbf{X}$. The estimator $\mathbf{b}$ in (2) is also the least squares (LS) estimator of $\boldsymbol{\beta}$. The usual procedures for constructing confidence ellipsoids/intervals and for testing hypotheses for $\boldsymbol{\beta}$ and σ^2 rely on the validity of (A1)–(A4). The consequences of the breakdown of any of these four assumptions are well known, and possible remedial measures, such as variable transformations, weighted regression, incorporation of additional predictor variables and, if necessary adoption of nonparametric methods, have also been discussed (see, e.g., Neter, Kutner, Nachtsheim, and Wasserman 1996).

Assessment of whether assumptions (A1)–(A4) are satisfied based on the data $(\mathbf{Y}, \mathbf{X})$ has received considerable attention. Assessment procedures typically involve the standardized residuals $\mathbf{R}$, defined herein according to

$$\mathbf{R} = \frac{1}{s}(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \frac{1}{s}(\mathbf{I} - \mathbf{P}[\mathbf{X}])\mathbf{Y}. \tag{3}$$

Other types of residuals have been used in model validation and diagnostics. The Studentized residuals are $\mathbf{R}' = (R'_1, \dots, R'_n)$ with $R'_i = \frac{n-p}{n}\sqrt{1 - h_{ii}}R_i$, where h_{ii} is the ith diagonal element of $\mathbf{H}$. Other residuals are Theil's (1965) best linear unbiased scalar (BLUS) covariance residuals and recursive or sequential residuals (see Kianifard and Swallow 1996). Here we focus on $\mathbf{R}$, because this is the residual vector that naturally arises from our theoretical development.

Important work in assessing the model assumptions includes that of Tukey (1949) for assessing (A1), Durbin and Watson (1950, 1951) for assessing (A3), and Anscombe (1961) and Anscombe and Tukey (1963) for assessing (A4) and (A2). Many of these methods have been summarized and discussed by Cook and Weisberg (1982) and Atkinson (1985). It should be noted that the residuals are not independent and may have different variances even if (A1)–(A4) hold, in contrast to the iid structure of $\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/\sigma$, which is the counterpart of $\mathbf{R}$ when the model parameters $\boldsymbol{\beta}$ and σ are known. The impact (especially the nonnegligible change on the distributional properties of the residuals in even large samples) of substituting estimators for unknown parameters to obtain the residuals has been duly noted (Durbin and Watson 1950; Anscombe and Tukey 1963; Theil 1965; Atkinson 1985).

In the assessment of (A1)–(A4) through graphical methods, the impact of the aforementioned substitution is mostly ignored, potentially giving rise to inaccurate assessment. Moreover, even apart from this issue, the interpretation of graphical methods is highly subjective, for although a picture is worth a thousand words, beauty is in the eye of the beholder. Furthermore, a particular plot is used to assess a specific assumption, and sometimes the synergistic impact of combinations of violations of

Edsel A. Peña is Professor, Department of Statistics, University of South Carolina, Columbia, SC 29208 (E-mail: *pena@stat.sc.edu*). Elizabeth H. Slate is Professor, Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, SC 29425 (E-mail: *slateeh@musc.edu*). Peña was supported by National Science Foundation grant DMS-01-02870, NIH grant GM056182, NIH COBRE grant RR17698, and the USC/MUSC Collaborative Research Program. Slate was supported by NIH grant CA077789, NIH COBRE grant RR17696, DAMD grant 17-02-1-0138, and the USC/MUSC Collaborative Research Program. The authors thank the associate editor and editor for their very useful and perceptive comments, suggestions, and changes that led to significant improvements in the manuscript.

© 2006 American Statistical Association
Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000637

(A1)–(A4) is not clear. It is therefore beneficial to augment these plots with a numerical measure of the degree to which (A1)–(A4) are violated.

Formal significance tests for (A1)–(A4) involve testing the null hypothesis (H_0) versus the alternative hypothesis (H_1), where

$$\begin{aligned} &H_0: \text{Assumptions (A1)–(A4) all hold,} \\ &H_1: \text{At least one of (A1)–(A4) does not hold.} \end{aligned} \tag{4}$$

The typical structure of such a test is to define a statistic $S(\mathbf{R})$ whose sampling distribution is known under H_0 and such that departures from H_0 will manifest in terms of larger values of $S(\mathbf{R})$. Given an observed residual vector $\mathbf{R} = \mathbf{r}$, one calculates the p value via $p = \Pr\{S(\mathbf{R}) > S(\mathbf{r})|H_0\}$, and the decision to reject H_0 is based on the magnitude of p. However, existing formal significance tests are typically tests for a specific assumption, and hence are not simultaneous or global tests for the four assumptions (A1)–(A4). For instance, there are tests for the normality assumption (Anscombe and Tukey 1963), tests for link misspecifications (Tukey 1949), tests for heterogeneity of variances (Cook and Weisberg 1983; Bickel 1978; Anscombe 1961), and tests for the uncorrelatedness or independence of the error components (Durbin and Watson 1950, 1951; Theil and Nagar 1961). (See also Kianifard and Swallow 1996 for procedures that use the recursive residuals for significance testing of the different assumptions.) The difficulty with these tests is that each is designed to detect departures from one assumption, and the impact of violations of other assumptions on this test, as well as its sensitivity against these violations are not apparent. Hence, when a specific test indicates a violation, it might be due to the violation of another assumption that affects this test. For example, a test for normality could be affected by a misspecified link function or dependent error components. One may decide to perform tests for each of the different assumptions, but this will lead to an increase in the type I error probability when the results of these tests are combined—although some corrective measure, such as a Bonferroni adjustment, could be implemented to alleviate this inflation. There is therefore the need for a global test for all of assumptions (A1)–(A4) that controls the type I error rate and could be used especially if the analyst does not have an idea of which set of assumptions are violated. If such a test indicates that at least one of the assumptions is not satisfied, then directional tests may be used to identify the assumptions that have been violated. Knowing the set of assumptions that has been violated is important for instituting appropriate remedial measures, such as variable transformations, adjustments in the link function, use of lagged values, or other strategies.

In this article we propose such a global test. An important consideration in our proposal is that the procedure should be simple and easy to implement, but at the same time should be theoretically justifiable. Our procedure is based on the residual vector $\mathbf{R}$, and the theoretical development of the procedure relies on the idea of Neyman's (1937) smooth test (Thomas and Pierce 1979; Rayner and Best 1986, 1989). The components of the global test also can be used as directional tests for determining the assumptions that have been violated. Because functions of $\mathbf{R}$ generally have complicated distributional properties, asymptotic distributional properties for the global test are ascertained. For small sample sizes, computer-intensive methods may be used to determine p values. We also discuss deletion statistics based on the global statistic that can be used to identify outlying or influential observations. Moreover, the mathematical framework for the test procedure is quite general and allows for generation of a broad class of tests by changing the embedding functions (see Sec. 3). However, with the goals of obtaining an easy-to-implement procedure and recovering some currently used directional tests, we have confined ourselves to a particular set of embedding functions. As a reviewer pointed out, a better procedure may arise through a different choice of embedding functions, but possibly at the cost of greater complexity. Even with this potential limitation, the performance of the proposed procedure is still commendable, as seen in the simulation studies in Section 5 and the applications in Section 6.

There is a deeper foundational issue regarding model building in relation to the validation of the model assumptions and the additional inferences made, such as testing hypotheses, constructing of confidence intervals about the regression parameters, or predicting future observations. For example, suppose that the linear model assumptions are validated through formal and/or graphical methods using the observed data, so this validation process is subject to error, and then regression parameters are estimated using the same data and through procedures derived under the linear model and its assumptions. How should one assess the properties of these estimators in light of this two-step process? There is a growing body of literature and ongoing active research on the more general, but related, area of inference after model selection (see, e.g., Hjort and Claeskens 2003; Claeskens and Hjort 2003; Dukić and Peña 2005; and references therein). This is an important issue that needs to be addressed, but this article focuses on formally validating the model assumptions.

The article is organized as follows. Section 2 describes and discusses the global and the component statistics. Section 3 presents the theoretical justification of the global procedure and derives it as a Neyman smooth test. The asymptotic normality and the asymptotic independence of the components are established in this section. Section 4 describes deletion statistics, obtained by excluding an observation from the analysis. Section 5 presents simulation studies that examine the properties of the procedures. Section 6 illustrates the applications of these procedures to two real datasets. Section 7 provides some concluding thoughts.

2. VALIDATION PROCEDURES

We first present the tests, then provide the theoretical justification for them. Henceforth, we assume that $\mathbf{X}$ has as its first column the $n \times 1$ vector $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{t}}$, so that we are incorporating an intercept term in model (1). Recalling that the ith component of the residual vector $\mathbf{R}$ is $R_i = (Y_i - \hat{Y}_i)/s$, $i = 1, 2, \ldots, n$, where $\hat{Y}_i = \mathbf{x}_i\mathbf{b}$ is the ith fitted value, the first three component statistics are

$$\hat{S}_1^2 = \left\{\frac{1}{\sqrt{6n}}\sum_{i=1}^{n} R_i^3\right\}^2, \tag{5}$$

$$\hat{S}_2^2 = \left\{\frac{1}{\sqrt{24n}}\sum_{i=1}^{n} (R_i^4 - 3)\right\}^2, \tag{6}$$

and

$$\hat{S}_3^2 = \frac{\{(1/\sqrt{n})\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 R_i\}^2}{(\hat{\Omega} - \mathbf{b}^t \hat{\boldsymbol{\Sigma}}_X \mathbf{b} - \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\Sigma}}_X^{-1} \hat{\boldsymbol{\Gamma}}^t)}, \qquad (7)$$

where, with $\bar{\mathbf{z}} = \frac{1}{n}\mathbf{1}^t\mathbf{Z}$ for an $n \times q$ matrix $\mathbf{Z}$, we define

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^4,$$
$$\hat{\boldsymbol{\Sigma}}_X = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^t(\mathbf{x}_i - \bar{\mathbf{x}}), \qquad (8)$$
$$\hat{\boldsymbol{\Gamma}} = \frac{1}{n}\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2(\mathbf{x}_i - \bar{\mathbf{x}}).$$

The fourth component statistic requires a user-supplied $n \times 1$ vector $\mathbf{V}$, which by default is set to be the standardized time sequence $\mathbf{V} = (1, 2, \ldots, n)^t/n$. This is defined via

$$\hat{S}_4^2 = \left\{\frac{1}{\sqrt{2\hat{\sigma}_V^2 n}}\sum_{i=1}^n (V_i - \bar{V})(R_i^2 - 1)\right\}^2, \qquad (9)$$

with $\hat{\sigma}_V^2 = \frac{1}{n}\sum_{i=1}^n (V_i - \bar{V})^2$. The global test statistic is defined as

$$\hat{G}_4^2 = \hat{S}_1^2 + \hat{S}_2^2 + \hat{S}_3^2 + \hat{S}_4^2. \qquad (10)$$

An appealing feature of this global statistic is that variants of the statistics $\hat{S}_i^2$, $i = 1, 2, 3, 4$, have been considered for significance testing purposes in earlier articles. For instance, statistics related to $\hat{S}_1^2$ and $\hat{S}_2^2$ have been given by Anscombe and Tukey (1963), and a statistic related to $\hat{S}_4^2$ has been considered by Cook and Weisberg (1983), Bickel (1978), and Anscombe (1961) in the context of testing for heteroscedasticity. The statistic $\hat{S}_3^2$ is related to Tukey's (1949) test for additivity. One of the main contributions of this article is combining these different directional statistics in a global statistic and determining its properties. We discuss in ensuing sections that this combined global statistic serves as an omnibus statistic for globally testing all of the assumptions of the linear model.

For large n, which for application purposes here is understood to mean that $n - p \geq 30$, the global test for the hypotheses H_0 versus H_1 in (4) at an asymptotic significance level of α is

$$\textit{Global test}: \quad \text{Reject } H_0 \text{ if } \hat{G}_4^2 > \chi^2_{4;\alpha}, \qquad (11)$$

where $\chi^2_{k;\alpha}$ is the $100(1-\alpha)$th percentile of a central chi-squared distribution with k degrees of freedom (df). If the test in (11) leads to the rejection of H_0, then the component statistics $\hat{S}_1^2$, $\hat{S}_2^2$, $\hat{S}_3^2$, and $\hat{S}_4^2$ can be examined by comparing their values to $\chi^2_{1;\alpha}$, or, perhaps more appropriately, to $\chi^2_{1;\alpha/4}$ [see the test in (19)] or $\chi^2_{1;1-(1-\alpha)^{1/4}}$ [see the test in (20)], to get an indication of which particular assumption or assumptions have been violated. The following are rough guidelines for interpreting the values of these component statistics, with these guidelines suggested by the theoretical considerations presented in Section 3 and the simulation results in Section 5:

1. Skewed error distributions are usually indicated by large values of the statistic $\hat{S}_1^2$.
2. Deviations from the normal distribution kurtosis of the true error distribution generally are revealed by large values of statistic $\hat{S}_2^2$.
3. The use of a misspecified link function, possibly due to the absence of other predictor variables in the model, is detected mostly by large values of the statistic $\hat{S}_3^2$.
4. The presence of heteroscedastic errors and/or dependent errors are typically manifested in large values of the statistic $\hat{S}_4^2$.
5. Simultaneous violations of at least two of assumptions (A1)–(A4) are manifested by large values of several of these component statistics.

3. THEORETICAL DEVELOPMENT OF THE PROCEDURE

From (1), if the true parameter values $\boldsymbol{\beta}$ and σ are known, then we may call (perhaps inappropriately) the error vector $\boldsymbol{\epsilon}$ the vector of "true" residuals $\mathbf{R}^0$. Thus $\mathbf{R}^0 \equiv \mathbf{R}^0(\sigma^2, \boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/\sigma$, which therefore is equal in distribution to the error vector $\boldsymbol{\epsilon}$. If H_0 holds, then the density function of $\mathbf{R}^0$ is

$$f_{\mathbf{R}^0}(\mathbf{r}^0) = \prod_{i=1}^n \phi(r_i^0),$$

where $\phi(z) = \exp\{-z^2/2\}/\sqrt{2\pi}$ is the standard normal density function. Following Neyman's (1937) idea for constructing a "smooth" test (Thomas and Pierce 1979; Rayner and Best 1989), we embed $f_{\mathbf{R}^0}(\mathbf{r}^0)$ into a class of density functions, indexed by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_6)^t$, whose members are of the form

$$f_{\mathbf{R}^0}(\mathbf{r}^0|\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta}) = C(\boldsymbol{\theta}; \sigma^2, \boldsymbol{\beta}) f_{\mathbf{R}^0}(\mathbf{r}^0) \exp\{\boldsymbol{\theta}^t \mathbf{Q}(\mathbf{r}^0; \sigma^2, \boldsymbol{\beta})\}, \qquad (12)$$

where $\mathbf{Q}(\mathbf{r}^0; \sigma^2, \boldsymbol{\beta}) \equiv \sum_{i=1}^n \mathbf{Q}_i(r_i^0; \sigma^2, \boldsymbol{\beta})$ with

$$\mathbf{Q}_i(z; \sigma^2, \boldsymbol{\beta}) = \left(z, z^2 - 1, z^3, z^4 - 3, \{(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\beta}\}^2 z, (V_i - \bar{V})[z^2 - 1]\right)^t.$$

The particular choice of the $Q_i(z; \sigma^2, \boldsymbol{\beta})$ functions is motivated by our desire to recover commonly used directional statistics. Other forms for the $Q_i(z; \sigma^2, \boldsymbol{\beta})$ functions, such as trigonometric or wavelet functions, are certainly possible and may lead to procedures with better properties. The function $C(\boldsymbol{\theta}; \sigma^2, \boldsymbol{\beta})$ in (12) is a proportionality constant that makes $f_{\mathbf{R}^0}(\mathbf{r}^0|\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta})$ a density function. A straightforward calculation shows that this constant satisfies $C(\boldsymbol{\theta}; \sigma^2, \boldsymbol{\beta})^{-1} = \prod_{i=1}^n \mathbf{E}[\exp\{\boldsymbol{\theta}^t \mathbf{Q}_i(Z; \sigma^2, \boldsymbol{\beta})\}]$, where Z is a standard normal random variable. Notice that in the embedding class, the null hypothesis density function obtains when $\boldsymbol{\theta} = \mathbf{0}$. When $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}^{*t})^t$ is fixed, this larger family, which does not depend on β_1, is an exponential family of densities and hence has many of the nice properties intrinsic to exponential families. Furthermore, observe that if we allow for the case where $\boldsymbol{\beta}^* = \mathbf{0}$, then the larger model is not identifiable, because ($\theta_5 = 0$, $\boldsymbol{\beta}^* \neq \mathbf{0}$) and ($\theta_5 \neq 0$, $\boldsymbol{\beta}^* = \mathbf{0}$) both lead to the same distribution. But because this model validation issue becomes practically important only in the presence of a "trend," which is the case where $\boldsymbol{\beta}^* \neq \mathbf{0}$, then in the theoretical development we assume that $\boldsymbol{\beta}^*$ resides in $\Re_{p-1} \setminus \{\mathbf{0}\}$, which does not lead to any technical difficulties, because this is still an open set in $\Re_{p-1}$.

We first consider the case where $\boldsymbol{\beta}$ and σ^2 are known, so that $\mathbf{R}^0 = \mathbf{R}^0(\sigma^2, \boldsymbol{\beta})$ is observable. Within the embedding class of density functions specified by (12), the score test for $H_0^*: \boldsymbol{\theta} = \mathbf{0}$ versus $H_1^*: \boldsymbol{\theta} \neq \mathbf{0}$ is easily developed. The use of score tests in this situation is appealing, because it is known that score tests are endowed with a "robustness of optimality" property (see Chen 1983, 1985 regarding this property; Cox and Hinkley 1974 for a general discussion of score tests). In our setting, it is straightforward to see that the score test statistic at $\boldsymbol{\theta} = \mathbf{0}$ equals

$$\mathbf{U}(\boldsymbol{\theta} = \mathbf{0}, \sigma^2, \boldsymbol{\beta}) = \mathbf{Q}(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta}).$$

Because under H_0, R_i^0, $i = 1, 2, \ldots, n$, are iid standard normal variables, for any positive integer k, $\mathbf{E}\{[R_i^0]^{2k+1}\} = 0$ and $\mathbf{E}\{[R_i^0]^{2k}\} = \prod_{j=1}^{k}(2j-1)$; so the covariance matrix of $\frac{1}{\sqrt{n}}\mathbf{Q}(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta})$ is

$$\boldsymbol{\Sigma}_{11}^{(n)}(\sigma^2, \boldsymbol{\beta}) = \begin{bmatrix} 1 & 0 & 3 & 0 & \frac{1}{n}T_2(\boldsymbol{\beta}) & 0 \\ 0 & 2 & 0 & 12 & 0 & 0 \\ 3 & 0 & 15 & 0 & \frac{3}{n}T_2(\boldsymbol{\beta}) & 0 \\ 0 & 12 & 0 & 96 & 0 & 0 \\ \frac{1}{n}T_2(\boldsymbol{\beta}) & 0 & \frac{3}{n}T_2(\boldsymbol{\beta}) & 0 & \frac{1}{n}T_4(\boldsymbol{\beta}) & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{2}{n}\sum_{i=1}^{n}(V_i - \bar{V})^2 \end{bmatrix},$$

where $T_k(\boldsymbol{\beta}) = \sum_{i=1}^{n}[(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\beta}]^k$, $k = 2, 4$. If, as $n \to \infty$, the following conditions are satisfied:

(a) there exists a nonsingular $p \times p$ matrix $\boldsymbol{\Sigma}_X$ such that $\frac{1}{n}T_2(\boldsymbol{\beta}) \xrightarrow{\text{pr}} \boldsymbol{\beta}^{\text{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}$;

(b) there exists a function $\Omega(\boldsymbol{\beta})$ such that $\frac{1}{n}T_4(\boldsymbol{\beta}) \xrightarrow{\text{pr}} \Omega(\boldsymbol{\beta})$;

(c) there exists a $\sigma_V^2 \in (0, \infty)$ such that $\frac{1}{n}\sum_{i=1}^{n}(V_i - \bar{V})^2 \xrightarrow{\text{pr}} \sigma_V^2$;

(d) $\{\max_{1 \le i \le n}[(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\beta}]^4\}/T_4(\boldsymbol{\beta}) = o_p(1)$; and

(e) $\{\max_{1 \le i \le n}(V_i - \bar{V})^2\}/\{\sum_{i=1}^{n}(V_i - \bar{V})^2\} = o_p(1)$,

then it follows from the Lindeberg–Feller central limit theorem (CLT) that, under H_0,

$$\frac{1}{\sqrt{n}}\mathbf{Q}(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta}) \xrightarrow{\text{d}} \text{N}\big(\mathbf{0}, \boldsymbol{\Sigma}_{11}(\sigma^2, \boldsymbol{\beta})\big),$$

where

$$\boldsymbol{\Sigma}_{11}(\sigma^2, \boldsymbol{\beta}) = \begin{bmatrix} 1 & 0 & 3 & 0 & \boldsymbol{\beta}^{\text{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 \\ 0 & 2 & 0 & 12 & 0 & 0 \\ 3 & 0 & 15 & 0 & 3\boldsymbol{\beta}^{\text{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 \\ 0 & 12 & 0 & 96 & 0 & 0 \\ \boldsymbol{\beta}^{\text{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 & 3\boldsymbol{\beta}^{\text{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 & \Omega(\boldsymbol{\beta}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\sigma_V^2 \end{bmatrix}. \qquad (13)$$

In this situation where $\boldsymbol{\beta}$ and σ^2 are assumed known, notice the asymptotic dependence of the components Q_1, Q_3, and Q_5, as well as Q_2 and Q_4. An asymptotic α-level score test for $H_0^*: \boldsymbol{\theta} = \mathbf{0}$ versus $H_1^*: \boldsymbol{\theta} \neq \mathbf{0}$ rejects H_0^* whenever

$$\frac{1}{n}\mathbf{Q}(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta})^{\text{t}}\big[\boldsymbol{\Sigma}_{11}^{(n)}(\sigma^2, \boldsymbol{\beta})\big]^{-1}\mathbf{Q}(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta}) \ge \chi^2_{6;\alpha}.$$

But because σ^2 and $\boldsymbol{\beta}$ are unknown, neither $\mathbf{R}^0$ nor $\boldsymbol{\Sigma}_{11}^{(n)}$ is observable. Thus there is a need to use estimators for σ^2 and $\boldsymbol{\beta}$ in $\mathbf{R}^0(\sigma^2, \boldsymbol{\beta})$, and by substituting the MLEs s^2 and $\mathbf{b}$ given in (2), we obtain the (estimated) residual vector $\mathbf{R} = \mathbf{R}^0(s^2, \mathbf{b})$ given in (3). To develop a test based on $\mathbf{R}$, we need the asymptotic distribution of $\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b})$ under H_0. Toward this goal, observe that the ML estimating equations for σ^2 and $\boldsymbol{\beta}$ that give rise to s^2 and $\mathbf{b}$ are

$$A\big(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta}\big) \equiv \mathbf{R}^0(\sigma^2, \boldsymbol{\beta})^{\text{t}}\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}) - n = 0 \qquad (14)$$

and

$$\mathbf{B}\big(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta}\big) \equiv \sigma\mathbf{X}^{\text{t}}\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}) = \mathbf{0}. \qquad (15)$$

Augmenting the vector $\mathbf{Q}$ with A and $\mathbf{B}$ and then invoking the Lindeberg–Feller CLT, we find that under H_0, plus the conditions guaranteeing asymptotic normality of $\mathbf{Q}(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta})$ enumerated earlier,

$$\frac{1}{\sqrt{n}}\begin{bmatrix} \mathbf{Q}(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta}) \\ A(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta}) \\ \mathbf{B}(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta}) \end{bmatrix} \xrightarrow{\text{d}} \text{N}\big(\mathbf{0}, \boldsymbol{\Xi}(\sigma^2, \boldsymbol{\beta})\big), \qquad (16)$$

where

$$\boldsymbol{\Xi}(\sigma^2, \boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\Sigma}_{11}(\sigma^2, \boldsymbol{\beta}) & \boldsymbol{\Sigma}_{12}(\sigma^2, \boldsymbol{\beta}) \\ \boldsymbol{\Sigma}_{12}(\sigma^2, \boldsymbol{\beta})^{\text{t}} & \boldsymbol{\Sigma}_{22}(\sigma^2, \boldsymbol{\beta}) \end{bmatrix},$$

with

$$\boldsymbol{\Sigma}_{12}(\sigma^2, \boldsymbol{\beta}) = \begin{bmatrix} 0 & \sigma\boldsymbol{\mu}_X \\ 2 & 0 \\ 0 & 3\sigma\boldsymbol{\mu}_X \\ 12 & 0 \\ 0 & \sigma[\boldsymbol{\Gamma}(\boldsymbol{\beta}) + (\boldsymbol{\beta}^{\text{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta})\boldsymbol{\mu}_X] \\ 0 & 0 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{22}(\sigma^2, \boldsymbol{\beta}) = \begin{bmatrix} 2 & \mathbf{0} \\ \mathbf{0} & \sigma^2(\boldsymbol{\Sigma}_X + \boldsymbol{\mu}_X^{\text{t}}\boldsymbol{\mu}_X) \end{bmatrix},$$

and $\boldsymbol{\mu}_X$ and $\boldsymbol{\Gamma}(\boldsymbol{\beta})$ defined according to

$$\bar{\mathbf{x}} \xrightarrow{\text{pr}} \boldsymbol{\mu}_X \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}[(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\beta}]^2(\mathbf{x}_i - \bar{\mathbf{x}}) \xrightarrow{\text{pr}} \boldsymbol{\Gamma}(\boldsymbol{\beta}).$$

By virtue of (14) and (15), when s^2 and $\mathbf{b}$ are substituted for σ^2 and $\boldsymbol{\beta}$, the last two components in the augmented vector are both equal to 0. Consequently, it follows by multivariate normal theory, or could be established more formally by relying on Pierce's (1982) result, that

$$\frac{1}{\sqrt{n}}\mathbf{Q}\big(\mathbf{R}^0(s^2, \mathbf{b}); s^2, \mathbf{b}\big) = \frac{1}{\sqrt{n}}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}) \xrightarrow{\text{d}} \text{N}\big(\mathbf{0}, \boldsymbol{\Xi}_{11.2}(\sigma^2, \boldsymbol{\beta})\big),$$

where $\boldsymbol{\Xi}_{11.2}(\sigma^2, \boldsymbol{\beta}) = \boldsymbol{\Sigma}_{11}(\sigma^2, \boldsymbol{\beta}) - \boldsymbol{\Sigma}_{12}(\sigma^2, \boldsymbol{\beta})\boldsymbol{\Sigma}_{22}(\sigma^2, \boldsymbol{\beta})^{-1}\boldsymbol{\Sigma}_{12}(\sigma^2, \boldsymbol{\beta})^{\text{t}}$. To provide a simplified form for this limiting covariance matrix, we establish the following intermediate result.

Lemma 1. If the first column of $\mathbf{X}$ is $\mathbf{1}$, then $\boldsymbol{\mu}_X(\boldsymbol{\Sigma}_X + \boldsymbol{\mu}_X^{\mathrm{t}}\boldsymbol{\mu}_X)^{-1}\boldsymbol{\mu}_X^{\mathrm{t}} = 1$.

Proof. Write $\mathbf{X} = [\mathbf{1} \quad \mathbf{W}]$ so

$$\boldsymbol{\Sigma}_X + \boldsymbol{\mu}_X^{\mathrm{t}}\boldsymbol{\mu}_X = \begin{bmatrix} 1 & \boldsymbol{\mu}_W \\ \boldsymbol{\mu}_W^{\mathrm{t}} & \boldsymbol{\Sigma}_W + \boldsymbol{\mu}_W^{\mathrm{t}}\boldsymbol{\mu}_W \end{bmatrix}.$$

Applying the partitioned matrix inverse theorem (Anderson 1984, thm. A.3.3), we obtain

$$[\boldsymbol{\Sigma}_X + \boldsymbol{\mu}_X^{\mathrm{t}}\boldsymbol{\mu}_X]^{-1} = \begin{bmatrix} 1 + \boldsymbol{\mu}_W\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\mu}_W^{\mathrm{t}} & -\boldsymbol{\mu}_W\boldsymbol{\Sigma}_W^{-1} \\ -\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\mu}_W^{\mathrm{t}} & \boldsymbol{\Sigma}_W^{-1} \end{bmatrix}.$$

Because $\boldsymbol{\mu}_X = (1 \ \boldsymbol{\mu}_W)$, the assertion immediately follows by matrix multiplication.

By straightforward multiplication and applying Lemma 1, we obtain

$$\begin{aligned} &\boldsymbol{\Delta}(\sigma^2, \boldsymbol{\beta}) \\ &\quad \equiv \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\mathrm{t}} \\ &\quad = \begin{bmatrix} 1 & 0 & 3 & 0 & \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 \\ 0 & 2 & 0 & 12 & 0 & 0 \\ 3 & 0 & 9 & 0 & 3\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 \\ 0 & 12 & 0 & 72 & 0 & 0 \\ \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 & 3\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} & 0 & \eta(\boldsymbol{\beta}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \end{aligned} \tag{17}$$

with $\eta(\boldsymbol{\beta}) = (\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta})^2 + \boldsymbol{\Gamma}(\boldsymbol{\beta})\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Gamma}(\boldsymbol{\beta})^{\mathrm{t}}$. The matrix $\boldsymbol{\Delta}(\sigma^2, \boldsymbol{\beta})$ is the correction factor in the limiting covariance matrix arising from plugging in s^2 and $\mathbf{b}$ for σ^2 and $\boldsymbol{\beta}$. This factor is clearly nonnegligible. Finally, from (13) and (17), a simplified form of $\Xi_{11.2}$ is

$$\Xi_{11.2}(\sigma^2, \boldsymbol{\beta}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 24 & 0 & 0 \\ 0 & 0 & 0 & 0 & \xi(\sigma^2, \boldsymbol{\beta}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\sigma_V^2 \end{bmatrix}, \tag{18}$$

where $\xi(\sigma^2, \boldsymbol{\beta}) = \Omega(\boldsymbol{\beta}) - (\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{\Sigma}_X\boldsymbol{\beta})^2 - \boldsymbol{\Gamma}(\boldsymbol{\beta})\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Gamma}(\boldsymbol{\beta})^{\mathrm{t}}$. We formally state this asymptotic result as a theorem.

Theorem 1. If assumptions (A1)–(A4) hold for the linear model in (1) with $\mathbf{X}$ having as its first column the vector $\mathbf{1}$, and if conditions 1–5 enumerated earlier hold, then $n^{-1/2}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b})$ converges in distribution to a mean-0 normal distribution with covariance matrix $\Xi_{11.2}$ given in (18).

Note the invariance of this asymptotic result to rescaling; that is, the result is independent of σ. This is a consequence of the fact that the model is scale-invariant and the residual vector is scale-equivariant. The theorem also indicates that $Q_1(\mathbf{R}; s^2, \mathbf{b})$ and $Q_2(\mathbf{R}; s^2, \mathbf{b})$ are degenerate at 0—hardly a surprise, because these quantities are the estimating functions for σ^2 and $\boldsymbol{\beta}$. What is surprising, however, is the asymptotic independence of $Q_3(\mathbf{R}; s^2, \mathbf{b})$ and $Q_5(\mathbf{R}; s^2, \mathbf{b})$, because, as noted earlier, $Q_3(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta})$ and $Q_5(\mathbf{R}^0; \sigma^2, \boldsymbol{\beta})$ are *not* asymptotically independent. Thus, interestingly and unexpectedly, replacing the unknown parameters by their MLEs in the quantities $\mathbf{Q}(\mathbf{R}^0(\sigma^2, \boldsymbol{\beta}); \sigma^2, \boldsymbol{\beta})$ made all of the components asymptotically independent!

The quantities $\Omega(\boldsymbol{\beta})$, $\boldsymbol{\Sigma}_X$, and $\boldsymbol{\Gamma}(\boldsymbol{\beta})$ can be consistently estimated by their empirical counterparts and with $\boldsymbol{\beta}$ replaced by $\mathbf{b}$. Their respective estimators are those given in (8), and so we are able to obtain a consistent estimator $\hat{\Xi}_{11.2}$ of $\Xi_{11.2}$. The score statistic for testing $H_0^*: \boldsymbol{\theta} = \mathbf{0}$ versus $H_1^*: \boldsymbol{\theta} \neq \mathbf{0}$, with σ^2 and $\boldsymbol{\beta}$ considered nuisance parameters, is the quadratic form of $\frac{1}{\sqrt{n}}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b})$ with quadratic matrix $\hat{\Xi}_{11.2}^{-}$, where for a matrix $\mathbf{M}$, $\mathbf{M}^{-}$ denotes a generalized inverse. It is immediate to see that this statistic is

$$\frac{1}{n}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b})^{\mathrm{t}}\hat{\Xi}_{11.2}^{-}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}) = \hat{S}_1^2 + \hat{S}_2^2 + \hat{S}_3^2 + \hat{S}_4^2 = \hat{G}_4^2,$$

where $\hat{S}_k^2$, $k = 1, 2, 3, 4$, and $\hat{G}_4^2$ are as defined in (5), (6), (7), (9), and (10). Theorem 1 therefore justifies the use of the chi-squared distribution with 4 df for assessing the magnitude of $\hat{G}_4^2$, as well as the 1-df chi-squared distributions for each of the component statistics.

Before proceeding, we mention three possible competing test procedures to the $\hat{G}_4^2$-based test. These competing tests are also included in the simulation studies. The first procedure is to perform simultaneous testing using the test statistics $\hat{S}_i^2$, $i = 1, 2, 3, 4$, but incorporating a Bonferroni adjustment. By virtue of the foregoing asymptotic results, this test is as follows:

BonfTest: Reject H_0 if $\mathrm{Gmax} \equiv \max_{1 \le i \le 4}\hat{S}_i^2 > \chi^2_{1;\alpha/4}$. (19)

This amounts to rejecting H_0 if at least one of the unidirectional tests rejects H_0 at a level of significance $\alpha/4$.

The second competing test arises by recognizing that under H_0, by invoking the asymptotic independence of the component statistics, the asymptotic distribution of the test statistic Gmax in (19) is $\Pr\{\mathrm{Gmax} \le w | H_0\} \overset{n \to \infty}{\longrightarrow} [P\{\chi_1^2 \le w\}]^4$. As a consequence, an asymptotic α-level test of H_0 is provided by

MaxTest: Reject H_0 if $\mathrm{Gmax} > \chi^2_{1;1-(1-\alpha)^{1/4}}$. (20)

When $\alpha = .05$, the critical values of the tests in (19) and (20) equal 6.239 and 6.205. This explains the almost-identical behavior of these two tests observed in the simulation studies (see Sec. 5).

The third competitor, referred in Section 5 as BoxCox, is the use of the Box and Cox (1964) power transformation. The idea is to fit the linear model on the transformed responses $y_i^*(\hat{\gamma})$, $i = 1, 2, \dots, n$, with the transformation

$$y \mapsto y^*(\gamma) = \begin{cases} (y^\gamma - 1)/\gamma & \text{if } \gamma \neq 0, \\ \log(y) & \text{if } \gamma = 0 \end{cases} \tag{21}$$

where γ is the transformation parameter and $\hat{\gamma}$ is its MLE. The null hypothesis H_0 is then rejected if the null hypothesis $H_0^*: \gamma = 1$ is rejected. The test for H_0^* used a likelihood ratio test, with the numerical implementation relying on the R language (Ihaka and Gentleman 1996) object BoxCox found in the MASS package of Venables and Ripley.

4. DELETION STATISTICS

It is important to accompany assessment of model assumptions with investigation for unusual observations (either outlying or influential), because such observations could affect inferences regarding model validity. Unusual observations may arise either as a consequence of model violations or as rare outcomes when the data in fact adhere to the model. In the first case, exclusion of the unusual observations from the analysis may have little impact, because the global test remains sensitive to violations in the remaining data, or may lead to a nonsignificant global test, because the excluded observations aid detection of violations substantially. In the second case, when the data meet model assumptions apart from rare exceptions, unusual observations may cause the global test to indicate violations, so that their deletion would then permit the procedure to reflect the adherence of the remaining data to the model. In any case, unusual observations should be handled with caution, and solid justification is required for their exclusion, as in the examples in Section 6.

A natural $\hat{G}_4^2$-based procedure for detecting unusual observations arises from the well-known idea of deletion statistics, which reflect the change in values of statistics after the deletion of an observation. For a statistic T, denote by $T[i]$ the value of the statistic after the ith observation is deleted. We are interested in the quantities

$$\Delta\hat{G}_4^2[i] = \left[\frac{\hat{G}_4^2[i] - \hat{G}_4^2}{\hat{G}_4^2}\right] \times 100, \qquad i = 1, 2, \ldots, n, \qquad (22)$$

which represent the percent relative change in the value of the global statistic $\hat{G}_4^2$ after deletion of the ith observation. The idea is that an observation with a large absolute value of $\Delta\hat{G}_4^2[i]$ either is an outlier or has large influence. The sign of this global deletion statistic is also informative, because a positive (negative) value indicates that the deleted observation makes the assumptions more (less) plausible.

Related to the statistic in (22) is the p value after the deletion of the ith observation, that is,

$$p[i] = \Pr\{\hat{G}_4^2[i] > \hat{g}_4^2[i] | H_0\}, \qquad i = 1, 2, \ldots, n,$$

where $\hat{g}_4^2[i]$ is the observed value of the global statistic after deletion of the ith observation. The evaluation of this probability could be performed using the (approximate) chi-squared distribution with 4 df. The idea is that if $p[i]$ is quite different from the other $p[j]$'s, then this will indicate that the ith observation is either an outlier or an influential observation.

A potentially useful and interesting plot is the scatterplot of $\mathbf{p}[\cdot] = (p[1], \ldots, p[n])^t$ versus $\Delta\hat{G}_4^2[\cdot] = (\Delta\hat{G}_4^2[1], \ldots, \Delta\hat{G}_4^2[n])^t$. Following Tukey's (1977) idea, in our plots we indicate the observation labels of those points beyond the outer fences of either $\Delta\hat{G}_4^2[\cdot]$ or $\mathbf{p}[\cdot]$. Such observations are unusual in that they either have a large influence on the value or the p value of the global statistic. This plotting idea is demonstrated in the illustrative examples in Section 6.

5. PROPERTIES OF THE PROCEDURES

We performed simulation studies to assess the achieved levels and powers of the proposed tests for small to moderate sample sizes. The simulation runs for assessing the levels each had 20,000 replications (except for the BoxCox test, which was added later at the suggestion of a reviewer), whereas the runs to determine the powers of the tests had 5,000 replications. The simulation code was in the R language (Ihaka and Gentleman 1996), using the function lm and built-in random-number generators. For each set of simulation runs associated with a particular combination of simulation parameters, we used a common covariate sequence $x_1, x_2, \ldots, x_n$, generated from the standard uniform distribution.

The first set of runs was to determine whether the procedures achieve a prespecified 5% level of significance for the sample sizes considered. The sample size n took values ranging from 5 to 1,200 (Table 1). The response values were generated according to the model

$$Y_i = x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \qquad (23)$$

where the ϵ_i's are generated from a standard normal distribution. The model

$$Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i, \qquad i = 1, 2, \ldots, n, \qquad (24)$$

was fitted, and the resulting residuals, R_i, $i = 1, 2, \ldots, n$, were used in the testing procedures. For $\hat{S}_4^2$, $\mathbf{V}$ was the default standardized time sequence. Table 1 summarizes the observed empirical rejection rates. Note that for small sample sizes ($n \le 30$), the asymptotic approximation is not satisfactory. Except for the test based on $\hat{S}_3^2$, the procedures tend to be conservative. For moderate to large sample sizes, the procedures achieve significance levels close to the nominal 5% level, although the $\hat{S}_2^2$-based statistic has a mild degree of conservatism even when the sample size is large. The rate of convergence to the χ_1^2 distribution for this statistic is rather slow, as has been noted in earlier articles (see, e.g., Doornik and Hansen 1994).

Table 1. Simulated Levels (%) of the Asymptotic 5%-Level Tests Based on 20,000 Replications, Except for the Box–Cox Test Based on 1,000 Replications

RunNum	*n*	$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$	*MaxTest*	*BonfTest*	*BoxCox*
1	5	0	0	12.285	0	0	0	0	4.5
2	15	2.445	.995	6.425	2.545	2.685	2.440	2.370	4.9
3	30	3.660	2.000	5.620	3.910	4.145	3.500	3.460	4.2
4	50	4.060	2.495	5.260	4.355	4.520	3.990	3.945	4.8
5	100	4.680	3.135	4.935	4.760	5.095	4.520	4.445	5.9
6	150	4.645	3.555	5.100	4.995	5.030	4.700	4.620	4.4
7	200	4.800	3.640	5.180	5.100	5.075	4.920	4.825	5.4
8	300	4.845	3.820	4.875	4.920	4.990	4.780	4.675	5.2
9	400	4.805	4.285	5.015	5.205	5.135	5.135	5.040	5.4
10	600	5.055	4.365	4.735	4.795	5.065	5.045	4.980	3.6
11	800	4.940	4.420	5.230	4.885	5.200	5.055	4.975	5.4
12	1,200	5.060	4.700	5.210	5.045	5.185	5.300	5.195	5.0

NOTE: The data were generated according to $Y_i = x_i + \epsilon_i$, where the x_i's are a fixed sequence generated from a standard uniform, and ϵ_i's are iid N(0, 1) variates. The model $Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i$ was fitted.

For the power simulations, n takes values in the set $\{15, 30, 50, 100, 200\}$. We examined the achieved powers of the tests for $n = 15$ and $n = 30$ because, apart from the $\hat{S}_3^2$-based test, the results in Table 1 show that the tests are conservative, which may be acceptable except for a potential decrease in power. We performed power simulations for specific types of departures from the model assumptions and for multiple violations of the assumptions. The generic data-generation model is given by

$$Y_i = x_i + \beta_2 x_i^{\gamma} + \sigma_i^* x_i^{\alpha} \epsilon_i, \qquad i = 1, 2, \ldots, n, \qquad (25)$$

where $\sigma_i^* = 1, i \le n/2$ and for $i > n/2$, $\sigma_i^* = \sigma_2$. In this model all of the assumptions are satisfied whenever $\beta_2 = 0$ or $\gamma = 0$, $\sigma_2 = 1$, $\alpha = 0$, and ϵ_i iid N(0, 1). To induce a dependent error structure, we considered two models. The first model, which gives the error sequence a martingale structure, has

$$\epsilon_i = \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \epsilon_j^* \qquad (26)$$

where the ϵ_j^*'s are iid from N(0, 1). The second model induces an autoregressive [AR(1)] structure via

$$\epsilon_1 = \epsilon_1^* \qquad \text{and} \qquad \epsilon_i = \frac{\rho \epsilon_{i-1} + \epsilon_i^*}{\sqrt{1 + \rho^2}}, \qquad i = 2, \ldots, n, \qquad (27)$$

with ρ a dependence parameter. Extensive summaries of the simulated powers for all of the sample sizes considered in the simulation and for many varieties of departures from model assumptions are summarized in a series of tables in a technical report of the same title as this article, which is available on request from the authors. To conserve space, Table 2 summarizes representative cases for each specific departure from model assumptions. Table 3 contains the results when multiple violations occur. The conclusions obtained from this representative summary in Table 2 coincide with those obtained from the extensive tables in the technical report. In the discussion of the simulation results that follows, we also refer to and make use of the more extensive tables in the technical report.

The first type of violation that we examined was a nonnormal error distribution. We considered several types of error distributions, broadly classified into symmetric and skewed distributions. The first four cases in Table 2 present the simulated powers of the tests when the error distribution is symmetric (t-distributed with 5 df), and is right-skewed (a centered chi-squared distribution with 5 df). The technical report includes the results for other error distributions, including the logistic, double-exponential, t, and chi-squared, with df other than 5. The global test is quite good relative to the best directional test based on the four component statistics, with its power not significantly degraded by combining the four-statistics and sometimes exceeding those based on the best directional test. The best directional test statistic is $\hat{S}_2^2$, which is a kurtosis-type statistic. Notice that the $\hat{S}_3^2$ test does not have any power for detecting this error distribution misspecification. As expected, when the df of the t-distribution increases, the power of the tests decreases. The powers of the MaxTest in (20) and the BonfTest in (19) are almost identical, and for these symmetric distributions are lower than those of the $\hat{G}_4^2$-based test. Additional runs where the error distribution is a normal contaminated with a t_1- or t_3-distribution were also performed. For contaminating proportions of .1 and .3, the results indicate that the $\hat{G}_4^2$-based test has good detection abilities for this violation and slightly higher power than the MaxTest and BonfTest. For the t-distributed error distribution, the performance of the BoxCox test was poor relative to the $\hat{G}_4^2$-test when the sample size is large.

When the errors have shifted chi-squared distributions, the global test performs acceptably well relative to the best test among the four directional tests, with the powers slightly degraded because of combining of the four directional tests, some of which do not have good power against this assumptional departure. The best directional test statistic is $\hat{S}_1^2$, the skewness-type statistic. $\hat{S}_3^2$ does not have any detection power for this alternative. When the df increases, the power diminishes, because the chi-squared distribution approaches the normal distribution. The MaxTest and the BonfTest perform just slightly

Table 2. Simulated Powers for Two Sample Sizes and Specific Type of Departure From the Linear Model Assumptions

Type of model violation	*Relevant model parameters*	*n*	$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$	*MaxTest*	*BonfTest*	*BoxCox*
Nonnormal error distribution	t_5	30	20.8	21.06	5.8	10.7	24	21	20	17.2
		100	39.1	60.80	5.3	16.7	60	57	57	28.9
	χ_5^2	30	47.9	19.70	5.4	10.4	34	32	31	77.9
		100	99.0	57.54	5.4	14.7	93	96	96	100.0
Heteroscedastic error	$\alpha = 2$	30	26.7	59.48	13.4	26.4	59	54	54	64.9
		100	40.8	99.20	12.1	44.3	98	98	98	75.0
	$\sigma_2 = 2$	30	12.1	10.66	6.6	40.8	27	25	25	9.7
		100	19.0	39.00	4.5	97.3	91	92	92	12.0
Misspecified link function	$\beta_2 = 3$	30	3.2	1.58	29.3	3.8	11	14	14	8.8
	$\gamma = 2$	100	4.2	2.86	54.7	5.1	31	35	35	9.3
	$\beta_2 = 5$	30	3.5	1.48	32.6	3.8	13	16	16	12.6
	$\gamma = 2$	100	4.4	2.92	95.5	5.4	82	88	88	35.7
Dependent error structure	Martingale type	30	15.6	7.30	2.6	39.3	27	27	26	25.4
		100	56.2	37.84	1.2	73.8	75	72	72	54.1
	Markov type	30	7.4	.86	6.4	22.6	14	13	12	7.3
	($\rho = 5$)	100	28.1	26.58	3.5	51.1	55	50	50	29.2

NOTE: The true model that generated the data is given in (25) with $\beta_2 = 0$, $\gamma = 0$, $\sigma_2 = 1$, $\alpha = 0$, and ϵ_i iid N(0, 1), except with the specific change in the column headed "Relevant model parameters" containing the violation.

Table 3. Simulated Powers (%) of the Tests for Models Where All the Four Assumptions Are Violated

	β_2	γ	ED[a]	α	GR[b]	σ_2	TT[c]	ρ	n	$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$	MaxTest	BonfTest	BoxCox
1	2	2	t_{10}	1	T	1.5	MT[d]	NA	15	10	8.5	11.7	3.4	13	9.7	9.6	37
2	2	2	t_{10}	1	T	1.5	MT	NA	30	21	28.0	8.5	21.1	33	28.4	28.2	42
3	2	2	t_{10}	1	T	1.5	MT	NA	50	38	75.6	23.7	44.0	80	75.4	75.1	68
4	2	2	t_{10}	1	T	1.5	MT	NA	100	45	92.0	25.2	47.4	92	89.8	89.7	68
5	2	2	t_{10}	1	T	1.5	MT	NA	200	52	99.9	42.1	85.4	100	99.9	99.9	79
6	2	.5	χ_{10}^2	2	T	1.5	MT	NA	15	16	14.1	6.8	27.5	25	19.1	19.0	26
7	2	.5	χ_{10}^2	2	T	1.5	MT	NA	30	40	56.9	8.2	17.8	57	52.0	51.9	45
8	2	.5	χ_{10}^2	2	T	1.5	MT	NA	50	55	88.1	17.6	54.3	90	86.0	85.9	49
9	2	.5	χ_{10}^2	2	T	1.5	MT	NA	100	77	99.9	11.9	49.8	100	99.5	99.5	66
10	2	.5	χ_{10}^2	2	T	1.5	MT	NA	200	92	100.0	9.3	69.5	100	100.0	100.0	84
11	1	2	LG[e]	2	T	2	AR[f]	3	15	11	29.6	42.8	3.3	49	36.9	36.5	57
12	1	2	LG	2	T	2	AR	3	30	50	87.4	22.8	34.0	84	81.6	81.5	51
13	1	2	LG	2	T	2	AR	3	50	57	97.8	8.9	75.9	98	97.2	97.1	45
14	1	2	LG	2	T	2	AR	3	100	64	100.0	14.2	90.2	100	100.0	100.0	51
15	1	2	LG	2	T	2	AR	3	200	70	100.0	9.6	98.9	100	100.0	100.0	52
16	−1	3	t_4	1	T	2	AR	−5	15	25	29.0	5.8	19.5	34	25.0	24.8	31
17	−1	3	t_4	1	T	2	AR	−5	30	43	59.6	7.2	29.0	61	55.6	55.4	35
18	−1	3	t_4	1	T	2	AR	−5	50	59	93.0	12.6	51.0	92	90.1	90.0	47
19	−1	3	t_4	1	T	2	AR	−5	100	68	99.7	8.9	89.6	100	99.6	99.6	53
20	−1	3	t_4	1	T	2	AR	−5	200	74	100.0	9.6	95.6	100	100.0	100.0	59

NOTE: The number of replications is 5,000. The true model that generated the data is given in (25), and the model in (24) is fitted to the data.

[a] Error distribution.

[b] Grouping.

[c] Time trend.

[d] Martingale.

[e] Logistic.

[f] AR(1).

better than the $\hat{G}_4^2$-based test for some values of n. In contrast, the BoxCox test performed very well for this right-skewed error distribution. Its power is significantly higher than that of the other tests for sample size $n = 30$. This superior performance of the BoxCox test could be intuitively explained by the fact that the transformation is especially appropriate for nonnormal and nonsymmetric error distributions. Interestingly, this nonsymmetric error distribution is the only instance in Table 2 in which the BoxCox test totally dominated the other tests.

The next set of simulation runs concerns the situation where (A2) is violated, so that the conditional variances of the Y_i's are not equal. Two models were considered for this purpose. The first model has variances that depend on the covariate values. Specifically, the true model is

$$Y_i = x_i + x_i^\alpha \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{28}$$

where the ϵ_i's are iid from N(0, 1). The simulated powers for $\alpha = 2$ are summarized in the fifth and sixth rows of Table 2. The best directional test for this departure is the $\hat{S}_2^2$-test, with the global test performing best among all of the tests. Again, the test based on $\hat{S}_3^2$ has very low power for this heteroscedastic model, though it is not totally devoid of detection power when $n = 200$. The second model for heteroscedastic variances is of form

$$Y_i = \begin{cases} x_i + \sigma_1 \epsilon_i & \text{for } i \le n/2 \\ x_i + \sigma_2 \epsilon_i & \text{for } i > n/2, \end{cases} \tag{29}$$

with the ϵ_i's also iid from N(0, 1). The seventh and eighth rows of Table 2 present the simulated powers for data arising from the model with $\sigma_1 = 1$ and $\sigma_2 = 2$. The best directional test in this situation is based on $\hat{S}_4^2$, followed by the test based on $\hat{S}_2^2$. The global test also has acceptable power, but lower power than $\hat{S}_4^2$. The powers of the MaxTest and BonfTest are just slightly lower than the power of the $\hat{G}_4^2$-test.

The next set of runs was for misspecified link functions; that is, when (A1) is violated. The data analyzed were generated according to the model

$$Y_i = x_i + \beta_2 x_i^\gamma + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{30}$$

with the ϵ_i's iid from N(0, 1). The ninth to twelfth rows of Table 2 provide the simulated powers of the tests for two different sets of (β_2, γ). Interestingly, the directional tests based on $\hat{S}_1^2$, $\hat{S}_2^2$, and $\hat{S}_4^2$ are not at all sensitive to this violation. The best directional test is based on $\hat{S}_3^2$. The global test also has detection power toward this misspecification, although its power is quite degraded relative to that of $\hat{S}_3^2$, possibly because the other three tests have no power against this alternative. Furthermore, the MaxTest and BonfTest have better powers than the $\hat{G}_4^2$-test. When $\beta_2 = 1$ and $\gamma \in \{.5, 2\}$, the powers of the tests are very low. However, this should not be perceived as a defect of the tests, because this is a consequence of the fact that for these parameter sets, the signal-to-noise ratios (SNRs) are very low. This SNR is measured via

$$\mathrm{SNR} = \frac{\mathbf{E}\{\mathrm{MSE(Fitted)}|\mathrm{True}\} - \mathbf{E}\{\mathrm{MSE(True)}|\mathrm{True}\}}{\mathbf{E}\{\mathrm{MSE(True)}|\mathrm{True}\}}, \tag{31}$$

with $\mathbf{E}\{\mathrm{MSE(Model\ A)}|\mathrm{Model\ B}\}$ being the expectation of the mean squared error when model A is fitted with the expectation evaluated with respect to model B. Thus $\mathbf{E}\{\mathrm{MSE(True)}|\mathrm{True}\} = \sigma^2$. It is straightforward to show that for the simulation model in (30), the SNR satisfies, for large n,

$$\mathrm{SNR}(\beta_2, \gamma, \sigma) \approx \left(\frac{\beta_2}{\sigma}\right)^2 \frac{\gamma^2}{(\gamma+1)^2(2\gamma+1)} \left\{1 - \frac{3(2\gamma+1)}{(\gamma+2)^2}\right\}.$$

For the values of $(\beta_2, \gamma, \sigma)$ used in the simulation studies, $\text{SNR}(1, .5, 1) \approx 1/450$, $\text{SNR}(1, 2, 1) \approx 1/180$, $\text{SNR}(3, .5, 1) \approx 1/50$, $\text{SNR}(3, 2, 1) \approx 1/20$, $\text{SNR}(5, .5, 1) \approx 1/18$, and $\text{SNR}(5, 2, 1) \approx 5/36$. These values explain the ordering of the simulated powers for the $\hat{S}_3^2$-based test. Note in particular that $\text{SNR}(5, .5, 1) \approx 1/18$ is only slightly larger than $\text{SNR}(3, 2, 1) \approx 1/20$, as reflected by the small differences in the observed powers for the $\hat{S}_3^2$-based test for these two sets of values of (β_2, γ).

For the simulation runs concerning violations of assumption (A3), we considered the two models described earlier for generating martingale-type and AR-type structures. In the simulation, we performed runs for $\rho \in \{.5, 1, 2, 5, 10\}$. The last four rows of Table 2 present the simulated powers of the tests under these dependent error models, with $\rho = 5$ for the AR-type structure. For the martingale structure, the best directional test is based on $\hat{S}_4^2$, with the global test surpassing the performance of this best directional test for large n and also being slightly better than the MaxTest and BonfTest. For the AR(1) structure, the best is also the $\hat{S}_4^2$-test, with the global test's power also very good, and again the power of the global test is best for large n. The tests based on $\hat{S}_1^2$ and $\hat{S}_2^2$ also have some detection abilities for this violation, but are not competitive with the $\hat{S}_4^2$-based test or the global test. The test based on $\hat{S}_3^2$ has no ability to detect this particular type of violation. The global test performs slightly better than the MaxTest and BonfTest for the AR(1) error structure.

Finally, we consider the situation where several of the assumptions are violated simultaneously. We expect that the global test is ideally suited for this situation. Table 3 presents the achieved powers of the tests for four sets of simulation parameters where all four assumptions (A1)–(A4) are violated as in (25). All four directional tests have detection abilities. The performance of the global test is extremely commendable, because its power is generally higher than any of the directional tests, as well as the MaxTest and the BonfTest. It is interesting to observe that the BoxCox test sometimes has higher power than the $\hat{G}_4^2$-test for small n, but the rate of increase of its power as n increases is relatively slow compared to the latter test. It is conceivable that the BoxCox test has higher power over the $\hat{G}_4^2$-test when n is small simply because the latter test is highly conservative for small n.

6. ILLUSTRATIVE EXAMPLES

Example 1. The first illustration pertains to car mileage data gathered by the first author while commuting from Ann Arbor, Michigan to Bowling Green, Ohio during the period October 20, 1996–January 27, 1999. There were 205 observations corresponding to gas fill-ups for the following variables: Date, the date of the gas fill-up; NumGallons (denoted by Y), the number of gallons of regular unleaded gasoline pumped into the car; MilesLastFill (denoted by X_1), the distance travelled since the last fill-up; NumDaysBetw (denoted by X_2), the number of days since last fill-up; and AveMilesGal, the miles per gallon between gas fill-ups. This dataset is available at *http://www.stat.sc.edu/~pena/DataSets/CarMileage.txt*. We fit the multiple linear regression model, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \sigma\epsilon_i, i = 1, 2, \ldots, 205$, where the ϵ_i's are iid N(0, 1) variates. Scatterplots of Y versus X_1 and X_2 are provided in Figures 1(a) and 1(b). Pearson's correlation coefficient between Y and X_1 is .653, between Y and X_2 is $-.002$, and between X_1 and X_2 is $-.378$. Other summary statistics for this dataset are provided in Table 4.

Table 5 presents the results of the analysis, including the F-value, estimates of regression coefficients, $\hat{\sigma}$, and the coefficient of determination. The row labeled "None" pertains to the analysis where all 205 observations were used. If the model assumptions are satisfied, then the regression coefficients β_1 and β_2 are both found to be significantly different from 0. However, $\hat{G}_4^2 = 24.26$ has a p value of 0, and those associated with $\hat{S}_2^2$ and $\hat{S}_3^2$ are also very small, indicating violation of model assumptions.

As advocated in Section 4, we examined for unusual observations. Figure 1(c), which is a scatterplot of $\Delta\hat{G}_4^2[\cdot]$ versus $\mathbf{p}[\cdot]$, indicates that the 19th, 56th, 67th, 146th, and 200th observations are highly unusual. Details of these observations and others that were excluded in further analyses are given in Table 6. The dates of these observations reveal their unusual nature. The 19th observation was obtained on Christmas Eve just before a long trip. In contrast to usual practice, although the gas tank was still almost half full, a decision was made to fill it completely, thus lowering fuel efficiency; the 146th observation was obtained during a long trip that mostly covered interstate highway driving; and the 200th observation encompassed a period when the car was driven during a blizzard and was stuck in deep snow, explaining the low fuel efficiency. The 56th and 67th observations showed up among these unusual values primarily because of their X_2 values of 26 (on vacation) and 22 (in repair shop) days.

We refitted the linear model with these five observations excluded from the analysis. The results are summarized in the third row in Table 5, which still indicates violations of model assumptions. More important, Figure 1(d) reveals that the 164th observation (in the original dataset) is highly unusual. Similar to the 56th and 67th observations, it has a large value of X_2 (equal to 21 due to vacation). Also observe the sensitivity of the directional statistics to the presence of unusual observations. In the first analysis, the p values of $\hat{S}_1^2$ and $\hat{S}_2^2$ were high and low, respectively, but after the exclusion of the unusual observations, this pattern reversed.

Further excluding this 164th observation (see fifth row of Table 5) now yields a global statistic of $\hat{G}_4^2 = 8.99$ with p value of .06 indicating that model assumptions appear viable, though the statistic $\hat{S}_3^2$ has a p value of .03. A glimpse at the $\Delta\hat{G}_4^2[\cdot]$ versus $\mathbf{p}[\cdot]$ scatterplot in Figure 1(e) reveals that the 58th observation (in the original dataset) is unusual, although there was no obvious explanation for this being so, in contrast to the other six values. Thus, although it may not be fully justifiable, in the final analysis we also excluded the 58th observation. The results, provided in the seventh last row of Table 5, show that all validation test statistics have p values $> .05$, indicating that after the exclusion of the seven unusual observations pinpointed by the deletion statistics, the linear model assumptions appear acceptable. Also observe that the scatterplot of $\Delta\hat{G}_4^2[\cdot]$ versus $\mathbf{p}[\cdot]$ in Figure 1(f) no longer shows any unusual observations. However, note that the p values of $\hat{S}_1^2$ and $\hat{S}_3^2$ are both between .05 and .10, which may be indicating mild violations of the normality and link function assumptions. For these reduced data,

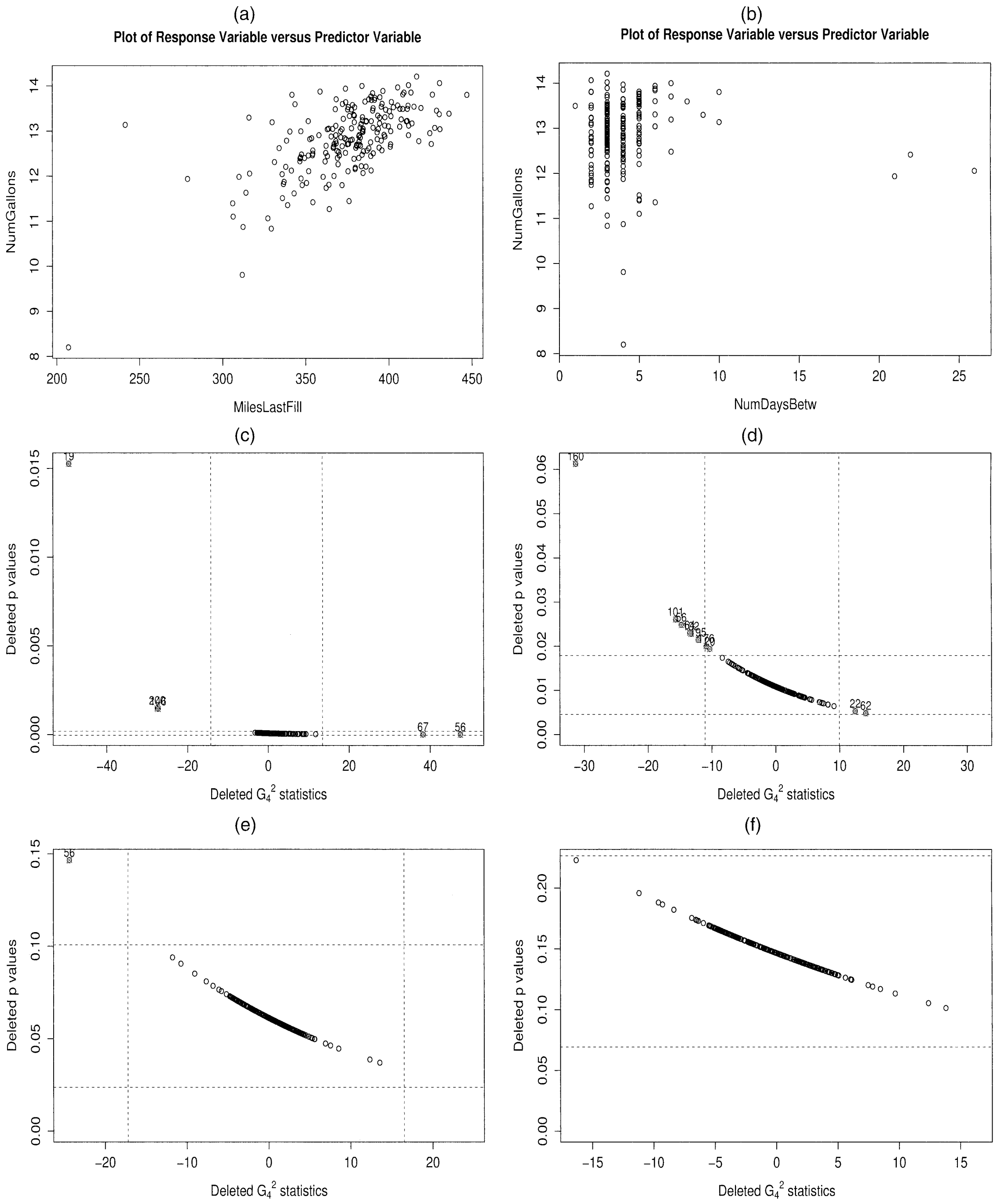

Figure 1. Relevant Plots for the Analysis of the Car Mileage Data. (a) A scatterplot of Y versus X_1. (b) A scatterplot of Y versus X_2. (c)–(f) Scatterplots of $\mathbf{p}[\cdot]$ versus $\Delta\hat{G}_4^2[\cdot]$ for successive reanalyses with some observations excluded.

Table 4. Summary Statistics for the Car Mileage Dataset, Where FQ and TQ Are the First and Third Quartiles

Statistic	X_2 = NumDaysBetw	Y = NumGallons	X_1 = MilesLastFill	AveMilesGal
Min	1	8.199	207.0	18.37
FQ	3	12.459	362.3	28.28
Med	4	12.909	379.3	29.46
TQ	5	13.344	394.5	30.58
Max	26	14.209	447.0	33.47
Mean	4.1	12.823	375.5	29.29
SD	2.7	.778	31.7	1.89

Table 5. Results of Analyses of the Car Mileage Data Using the Complete Data, and With Excluded Observations

Excluded observations	Test statistic/(p value)						Estimates/(SE)				
	$\hat{G}_4^2$	$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{F}$	b_0	b_1	b_2	$\hat{\sigma}$	R^2
None	24.26 ($\approx$ 0)	.03 (.86)	17.24 ($\approx$ 0)	6.90 (.01)	.08 (.78)	99.34 (0)	5.48 (.53)	.019 (.001)	.08 (.02)	.56	50%
19, 56, 67, 146, 200	13.09 (.01)	5.07 (.02)	.08 (.78)	7.81 (.01)	.14 (.71)	105.3 (0)	5.00 (.54)	.019 (.001)	.15 (.02)	.47	52%
19, 56, 67, 146, 164, 200	8.99 (.06)	3.15 (.08)	.09 (.76)	4.47 (.03)	1.28 (.26)	128.9 (0)	4.64 (.52)	.019 (.001)	.24 (.03)	.45	57%
19, 56, 58, 67, 146, 164, 200	6.80 (.15)	2.81 (.09)	.08 (.78)	2.84 (.09)	1.07 (.30)	133.2 (0)	4.48 (.52)	.020 (.001)	.24 (.03)	.44	58%

NOTE: The F-column contains the F-statistic value for testing that $\beta_1 = \beta_2 = 0$, whereas the R^2-column contains the coefficient of determination.

the correlation coefficient between Y and X_1 is .618; that between Y and X_2 is .219, a significant increase from the original correlation of $-.002$ indicating the impact of the unusual observations; and that between X_1 and X_2 is $-.323$.

Example 2. The second example is a multiple regression analysis of Ruppert and Carroll's (1980) water salinity data (see their table 3), which they used to illustrate robust regression techniques, and that was also used for illustrative purposes by Atkinson (1985, pp. 48–52). The data set consisted of 28 observations on the variables Salinity, the water salinity at the specified time period; LagSalinity, the water salinity lagged 2 weeks; Trend, representing one of the six biweekly periods in March–May; and WaterFlow, the river discharge. The response variable is Salinity, and the predictors are LagSalinity, Trend, and WaterFlow. The first part of the analyses fitted the multiple regression model,

$$\text{Salinity} = \beta_0 + \beta_1(\text{LagSalinity}) + \beta_2(\text{Trend}) + \beta_3(\text{WaterFlow}) + \sigma\epsilon. \quad (32)$$

The fitted model had $b_0 = 9.590$, $b_1 = .777$, $b_2 = -.026$, and $b_3 = -.295$. The coefficients β_0, β_1, and β_3 were significantly different from 0. The multiple R^2 was 82.6%. When the model validation procedures were applied, we obtained $\hat{G}_4^2 = .16$ $(p = .997)$, $\hat{S}_1^2 = .02$ $(p = .87)$, $\hat{S}_2^2 = .005$ $(p = .95)$, $\hat{S}_3^2 = 7.63 \times 10^{-6}$ $(p = .998)$, and $\hat{S}_4^2 = .128$ $(p = .72)$. The global test thus indicates that model assumptions are acceptable, although, as noted by a reviewer, the nearness of these p values to 1 also raises a "too good a fit" concern.

An examination of the plot of $\Delta G_4^2[\cdot]$ versus $\mathbf{p}[\cdot]$ in Figure 2(a) reveals that the 16th observation is highly unusual. Atkinson (1985) revealed the unusual nature of this observation was revealed using a half-normal plot of Cook's (1977) statistic. It was also pointed out that the value of WaterFlow for this observation is the reason why it is unusual and highly leveraged. The fact that the global test did not conclude violation of model assumptions, even with this very unusual observation, may cast doubt on the effectiveness of the validation procedure. Further analyses of the data reveal the reason for this behavior, however. LagSalinity is an excellent linear predictor of Salinity, with the correlation coefficient between them equal to .872, whereas WaterFlow is not highly correlated with Salinity, with their correlation coefficient equal to $-.477$. In addition, the correlation coefficient between LagSalinity and WaterFlow is $-.261$. Consequently, when the variables LagSalinity and WaterFlow are included in the regression model, the effect of WaterFlow is diminished by the presence of LagSalinity. When LagSalinity is used in the linear regression model without WaterFlow, then $\hat{G}_4^2 = 2.07$ $(p = .72)$, indicating that model assumptions are not violated. However, when WaterFlow is used as the sole predictor variable for Salinity, then $\hat{G}_4^2 = 10.33$ $(p = .035)$ and $\hat{S}_3^2 = 7.50$ $(p = .006)$, indicating violations of model assumptions and, in addition, the 16th observation is highly unusual. Therefore, when LagSalinity and WaterFlow are both in the linear regression model and the original data are used, then model assumptions are in fact satisfied. Thus the conclusion from the global test that the assumptions are satisfied, even with the unusual observation, is not cause for alarm regarding the effectiveness of the proposed validation procedure.

We follow Atkinson (1985, p. 49), by supposing that the value of 33.443 for WaterFlow for this 16th observation was a misprint of 23.443. We refitted the model in (32) but using 23.443 in place of 33.443. The resulting analysis yielded $b_0 = 18.39$, $b_1 = .70$, $b_2 = -.15$, and $b_3 = -.63$, with β_0, β_1, and β_3 significantly different from 0. The multiple R^2 was 89.28%. Applying the model validation procedures, we obtained $\hat{G}_4^2 = 6.696$ $(p = .15)$, $\hat{S}_1^2 = 1.41$ $(p = .23)$, $\hat{S}_2^2 = .03$

Table 6. Details of the Observations That Were Excluded From the Analyses

Excluded observations	X_2 = NumDaysBetw	Y = NumGallons	X_1 = MilesLastFill	AveMilesGal	Date fill-up
19	4	8.199	207.0	25.25	12/24/96
56	26	12.058	316.2	26.22	6/10/97
58	4	13.043	430.8	33.03	6/19/97
67	22	12.417	346.9	27.94	8/15/97
146	4	9.809	311.8	31.79	5/24/98
164	21	11.937	278.8	23.36	8/17/98
200	10	13.138	241.4	18.37	1/8/99

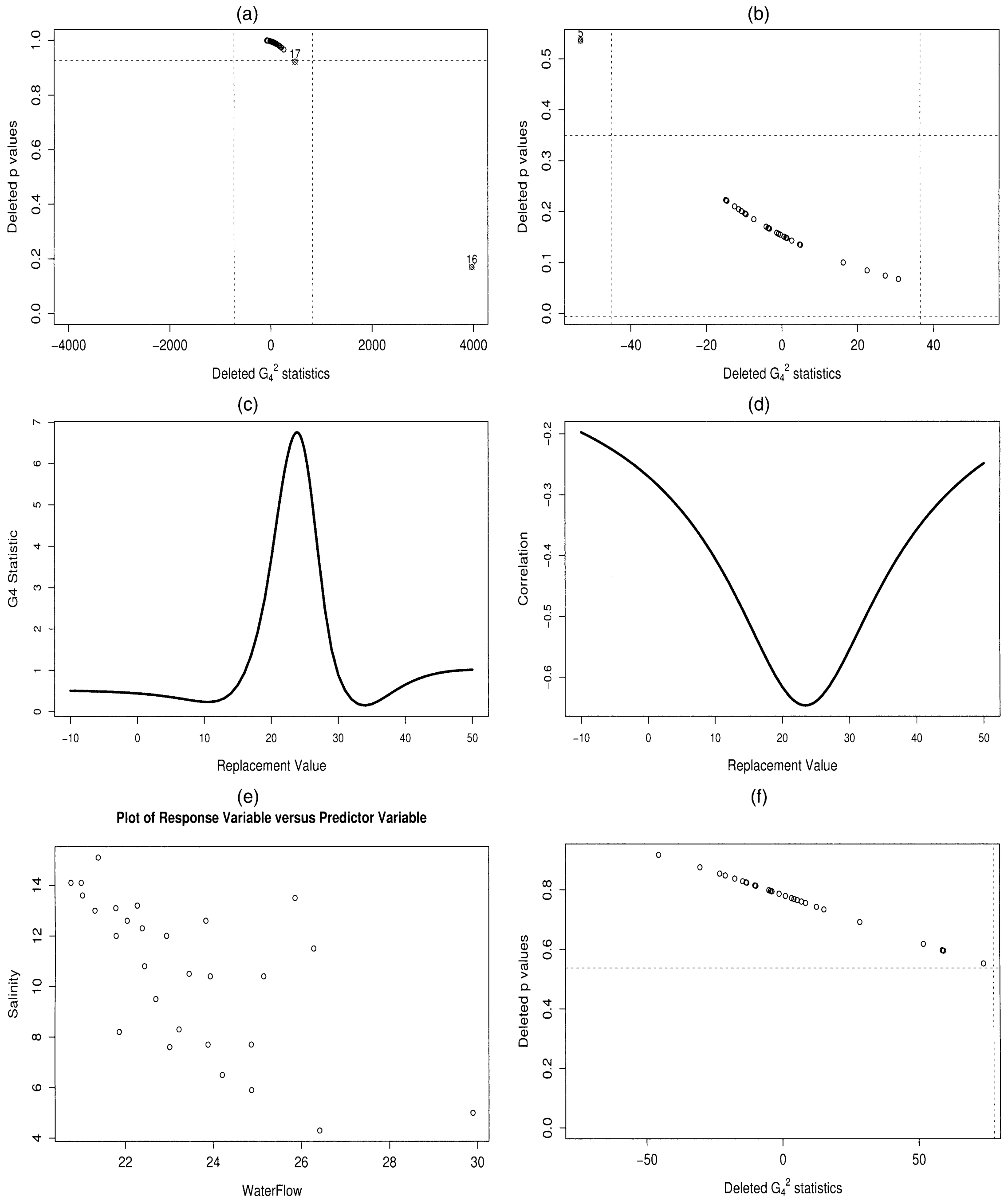

*Figure 2. Relevant Plots for the Salinity Data Example. (a), (b), and (f) Scatterplots of $\Delta\hat{G}_4^2[i]$ versus **p**[i] for the original data, corrected data, and the corrected data with a quadratic WaterFlow term in the model. (c) and (d) The resulting values of $\hat{G}_4^2$ and the correlation between Salinity and WaterFlow for different replacement values for WaterFlow in the 16th observation. (e) A scatterplot of Salinity and WaterFlow using Atkinson's replacement value of 23.443.*

($p = .86$), $\hat{S}_3^2 = 4.21$ ($p = .04$), and $\hat{S}_4^2 = 1.04$ ($p = .31$). Although the global statistic has p value exceeding 10%, the p value for $\hat{S}_3^2$ is .04, which seems to indicate a mild problem in the link function. The scatterplot of $\Delta\hat{G}_4^2[\cdot]$ versus $\mathbf{p}[\cdot]$ in Figure 2(b) indicates no unusual observations, except possibly for the fifth observation.

To gain further insight into these data, we examined the impact of different replacement values for WaterFlow in the 16th observation, specifically on the resulting value of $\hat{G}_4^2$. Figure 2(c) presents the values of $\hat{G}_4^2$ for different replacement values for WaterFlow. Note that the $\hat{G}_4^2$-value is largest—and hence the most indicative of violations of model assumptions—when the replacement value is about 23.9, which is close to the value of 23.443 used by Atkinson. Figure 2(d) presents the values of the correlation coefficients between Salinity and WaterFlow for different replacement values, and from this plot the largest absolute correlation is at a value very close to Atkinson's replacement value of 23.443. Using this value, the correlation coefficient between Salinity and WaterFlow is −.646. Because the correlation coefficients between Salinity and WaterFlow become largest for replacement values in the interval from 23.3 to 23.9, the impact of WaterFlow in the linear regression model when LagSalinity is also included in the model is not easily diminished, in contrast to the situation when using the original data. Consequently, at such replacement values, potential model violations, especially with regard to the linearity assumption for WaterFlow, materialize. Figure 2(e), a scatterplot between Salinity and WaterFlow when using the replacement value of 23.443, partly reveals a curvilinear relationship between Salinity and WaterFlow.

Recognizing the possible problem with the link function, we followed Atkinson's (1985, p. 51) suggestion to incorporate a quadratic term of WaterFlow and fitted the model

$$\begin{aligned}\text{Salinity} = {} & \beta_0 + \beta_1(\text{LagSalinity}) + \beta_2(\text{Trend}) \\ & + \beta_3(\text{WaterFlow}) + \beta_4(\text{WaterFlow})^2 + \sigma\epsilon. \quad (33)\end{aligned}$$

The resulting estimates are $b_0 = 67.49$, $b_1 = .68$, $b_2 = -.25$, $b_3 = -4.57$, and $b_4 = .08$, and the multiple R^2 was 91.65%. Only β_2 did not turn out to be significantly different from 0 with a p value of .053. The model validation statistics are $\hat{G}_4^2 = 1.74$ $(p = .78)$, $\hat{S}_1^2 = 1.20$ $(p = .22)$, $\hat{S}_2^2 = .022$ $(p = .88)$, $\hat{S}_3^2 = .176$ $(p = .67)$, and $\hat{S}_4^2 = .348$ $(p = .55)$. The scatterplot of $\Delta\hat{G}_4^2[\cdot]$ versus $\mathbf{p}[\cdot]$ in Figure 2(f) no longer shows unusual observations.

7. CONCLUDING REMARKS

In this article we have proposed a global test procedure for validating the four assumptions of the linear model. The global test statistic, which is a function of the model residuals, is formed from four asymptotically independent statistics, each with the potential to detect a particular violation. The level and power properties of the tests were examined through simulation studies, which indicate that the global and directional tests have the ability to detect different types of violations of the model assumptions. As such, the tests provide a formal method for globally assessing the validity of model assumptions. Deletion statistics and graphical methods based on the global statistic can be used to identify unusual observations. The proposed formal procedure may help eliminate, or at least reduce, the oftentimes subjective assessment of the validity of model assumptions when using existing graphical techniques.

Other issues remain to be addressed. First, there is the problem of developing an adaptive method. From the simulation results, the power of the global test is generally lower than that of the best directional test when only one assumption is violated. Some of the directional tests have no power for detecting certain types of alternatives, and hence they tend to dilute the power when included in this global statistic. We conjecture that it will be possible to have the data dictate which among the four directional test statistics to combine to form a global test statistic, and by doing so we expect that the resulting adaptive global test may acquire increased power. Several approaches to determining which directional test statistics to combine present themselves, including those using information measures like the Schwartz (1978) Bayesian information criterion or the Akaike information criterion (Akaike 1973). Second, the chi-squared approximation is not satisfactory for small sample sizes, though we have pointed out that except for the $\hat{S}_3^2$-based statistic, the approximation leads to conservative tests. Two possible ways to alleviate this problem are to use empirical estimates of the covariance matrices instead of the theoretical matrices and to use computationally intensive methods to determine the critical regions of the tests. Third, the main motivation for choosing the smoothing functions in the density embedding is to recover some commonly used one-dimensional test statistics, with the aim of formally combining them into one global test statistic. We have achieved this in this article. However, one is not limited in choosing the functions that enter into the embedding. Finally, to make these linear model validation procedures accessible to practitioners, we plan to provide the procedures through a computer package in the R Library.

[Received March 2004. Revised May 2005.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium of Information Theory*, Budapest: Akademiai Kiado, pp. 267–281.

Anderson, T. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: Wiley.

Anscombe, F. (1961), "Examination of Residuals," *Proceedings of the Fourth Berkeley Symposium*, 1, 1–36.

Anscombe, F., and Tukey, J. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141–160.

Atkinson, A. (1985), *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford, U.K.: Clarendon Press.

Bickel, P. (1978), "Using Residuals Robustly I: Tests for Heteroscedasticity, Nonlinearity," *The Annals of Statistics*, 6, 266–291.

Box, G., and Cox, D. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 143, 383–430.

Chen, C. (1983), "Score Tests for Regression Models," *Journal of the American Statistical Association*, 78, 158–161.

——— (1985), "Robustness Aspects of Score Tests for Generalized Linear and Partially Linear Regression Models," *Technometrics*, 27, 277–283.

Claeskens, G., and Hjort, N. (2003), "The Focused Information Criterion" (with discussion), *Journal of the American Statistical Association*, 98, 900–945.

Cook, R. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R., and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman & Hall.

——— (1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, 70, 1–10.

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

Doornik, J., and Hansen, H. (1994), "An Omnibus Test for Univariate and Multivariate Normality," available at *citeseer.nj.nec.com/doornik94omnibu.html.*

Dukić, V., and Peña, E. (2005), "Variance Estimation in a Model With Gaussian Submodels," *Journal of the American Statistical Association*, 100, 296–309.

Durbin, J., and Watson, G. (1950), "Testing for Serial Correlation in Least Squares Regression: I," *Biometrika*, 37, 409–428.

——— (1951), "Testing for Serial Correlation in Least Squares Regression: II," *Biometrika*, 38, 159–178.

Hjort, N., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Kianifard, F., and Swallow, W. (1996), "A Review of the Development and Application of Recursive Residuals in Linear Models," *Journal of the American Statistical Association*, 91, 391–400.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996), *Applied Linear Statistical Models* (4th ed.), New York: Irwin.

Neyman, J. (1937), "'Smooth' Test for Goodness of Fit," *Skandinavsk Aktuarietidskrift*, 20, 150–199.

Pierce, D. (1982), "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics," *The Annals of Statistics*, 10, 475–478.

Rayner, J., and Best, D. (1986), "Neyman Type Smooth Test for Location-Scale Families," *Biometrika*, 73, 437–446.

——— (1989), *Smooth Tests of Goodness of Fit*, New York: Oxford University Press.

Ruppert, D., and Carroll, R. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Theil, H. (1965), "The Analysis of Disturbances in Regression Analysis," *Journal of the American Statistical Association*, 60, 1067–1079.

Theil, H., and Nagar, A. (1961), "Testing the Independence of Regression Disturbances," *Journal of the American Statistical Association*, 56, 793–806.

Thomas, D., and Pierce, D. (1979), "Neyman's Smooth Goodness-of-Fit Test When the Hypotheses Is Composite," *Journal of the American Statistical Association*, 74, 441–445.

Tukey, J. (1949), "One Degree of Freedom for Nonadditivity," *Biometrics*, 5, 232–242.

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.