Statistical Science
2004, Vol. 19, No. 4, 562–570
DOI 10.1214/088342304000000549

Robust Analysis of Linear Models

Joseph W. McKean

Abstract. This paper presents three lectures on a robust analysis of linear models. One of the main goals of these lectures is to show that this analysis, similar to the traditional least squares-based analysis, offers the user a unified methodology for inference procedures in general linear models. This discussion is facilitated throughout by the simple geometry underlying the analysis. The traditional analysis is based on the least squares fit which minimizes the Euclidean norm, while the robust analysis is based on a fit which minimizes another norm. Several examples involving real data sets are used in the lectures to help motivate the discussion.

Key words and phrases: Asymptotic relative efficiency, breakdown point, diagnostics, influence function, least squares, linear hypotheses, nonparametrics, norms, rank-based analysis, robustness, Wilcoxon scores.

1. INTRODUCTION

Traditional procedures for the analysis of linear models are the most widely used statistical procedures, because they offer the user a unified methodology with which to attack many diverse problems. For example, the traditional F test of linear hypotheses can be used to test for the effect in the simple two-sample problem as well as to test for interaction between covariates and treatments in a complicated analysis of covariance setting. Traditional methods are based on the least squares (LS) fit of the linear model. The geometry behind this fit is quite simple. The LS fit minimizes the Euclidean distance between the response vector and the space defined by the linear model. The F test is simply the comparison of this minimum distance and the distance to the subspace defined by the null hypothesis.

In this series of three lectures, we discuss a robust analysis of linear models. It is quite analogous to the traditional analysis because we simply replace the Euclidean norm with another norm. Thus the geometry remains the same and, hence, the robust analysis offers the user a unified methodology for linear models similar to LS. Furthermore, the analysis is robust, not sensitive to outliers in the response space and highly efficient compared to the LS analysis.

Joseph W. McKean is Professor, Department of Statistics, Western Michigan University, Kalamazoo, Michigan 49008-5278, USA (e-mail: joe@stat.wmich.edu).

The first lecture, Section 2, presents the fitting procedure and discusses its robustness and efficiency. In the second lecture, Section 3, we discuss the associated F-type test for general linear hypotheses. The practicality of the robust analysis is illustrated by several examples where the robust analysis leads to different interpretations. The last lecture, Section 4, extends the robust analysis to a high breakdown analysis which is robust in both the response and factor spaces, and we discuss diagnostics for investigating differences between fits.

There are several classes of robust estimates from which we can choose. We selected an estimator which has its roots in traditional nonparametric rank procedures for simple location problems. We call this analysis a rank-based analysis or the Wilcoxon analysis (only linear scores are considered in this article). It is defined in terms of a norm, so geometrically it is analogous to LS. Furthermore, it generalizes immediately to high breakdown estimates. However, M estimates can also be used. Huber's M estimate is similar to the Wilcoxon estimates and the GM estimates (see Simpson, Ruppert and Carroll, 1992) are similar to the high breakdown estimates of Section 4.

Because its objective function is convex, the Wilcoxon fit is easy to compute. Computations in this article were obtained at the web sites *www.stat.wmich.edu/slab/RGLM* and *www.stat.wmich.edu/slab/HBR2*, which the reader is invited to use; see Crimin, Abebe and McKean (2003). Minitab also offers the `rregr`

command to obtain the Wilcoxon fit. Recently, Terpstra and McKean (2004) developed packages of R and SPLUS functions which compute these procedures for the Wilcoxon and the high breakdown estimates. These can be downloaded at the site *www.stat.wmich.edu/mckean/HMC/Rcode*. At the web site *www.stat.wmich.edu/mckean/Statsci/Data* the reader can find the data sets used in this article.

These lectures draw on many references. The monographs by Hettmansperger and McKean (1998), Hampel, Ronchetti, Rousseeuw and Stahel (1986), Huber (1981) and Rousseeuw and Leroy (1987) contain much of the material. A recent discussion was presented by Hettmansperger, McKean and Sheather (2000). Chapter 9 of Hollander and Wolfe (1999) presents a recent discussion of the Wilcoxon analysis for linear models. For the most part, we have not included references in the three lectures, trusting that the interested reader will consult these monographs and their included references.

2. ROBUST FIT OF LINEAR MODELS

Suppose the results of an experiment or an observational study consist of p explanatory (predictor) variables, $x_1, \ldots, x_p$, and a response variable Y. Further suppose that we have collected data for n subjects in this study. Denote the data for the ith subject as $(x_{i1}, \ldots, x_{ip}, Y_i)' = (\mathbf{x}_i, Y_i)'$. Let $\mathbf{Y}$ be the $n \times 1$ vector of responses and and let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i$. The predictors may be continuous variables or dummy variables. Assume that $\mathbf{X}$ has full column rank p.

Suppose we have decided to model the response vector $\mathbf{Y}$ as a linear model given by

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, α is an intercept parameter and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. The matrix $\mathbf{X}_c$ is the centered design matrix; that is, $x_{cij} = x_{ij} - \overline{x}_j$, where $\overline{x}_j = n^{-1}\sum_{i=1}^n x_{ij}$. Let V_c be the column space of the matrix $\mathbf{X}_c$. Assume that the errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. with c.d.f. $F(x)$, p.d.f. $f(x)$ and (for LS) variance σ^2.

The vector $\mathbf{X}_c\boldsymbol{\beta}$ lies in V_c. Hence, given a norm on $\mathbb{R}^n$, a simple way to estimate it is to find a vector in V_c which lies closest to $\mathbf{Y}$. Once $\boldsymbol{\beta}$ is estimated, α can be estimated by an appropriate location estimator of the residuals $Y_i - \mathbf{x}_{ci}'\widehat{\boldsymbol{\beta}}$.

2.1 Norms and Estimating Equations

Let us first review the LS estimator of $\boldsymbol{\beta}$ which utilizes the Euclidean norm. The LS estimate is given by $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathrm{Arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}_c\boldsymbol{\beta}\|^2$, where $\|\cdot\|_{\mathrm{LS}}^2$ is the squared Euclidean norm; that is, $\|\mathbf{v}\|_{\mathrm{LS}}^2 = \sum_{i=1}^n v_i^2$, $\mathbf{v} \in \mathbb{R}^n$. Differentiating the right-hand side with respect to $\boldsymbol{\beta}$ and setting the resulting equations to $\mathbf{0}$, we see that the LS estimator solves the estimating equations (normal equations) $\mathbf{X}_c'(\mathbf{Y} - \mathbf{X}_c\boldsymbol{\beta}) = \mathbf{0}$, with the solution $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}$. The estimate of α is the arithmetic average of the residuals which, because the x's are centered, is $\widehat{\alpha}_{\mathrm{LS}} = \overline{Y}$. Under regularity conditions, the large sample distribution for these estimates is given by:

$$\begin{pmatrix} \widehat{\alpha}_{\mathrm{LS}} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \end{pmatrix} \text{ has an approximate } N\left(\begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{bmatrix} 1/n & \mathbf{0}' \\ \mathbf{0} & (\mathbf{X}_c'\mathbf{X}_c)^{-1} \end{bmatrix}\right) \text{ distribution.} \tag{2.2}$$

The Wilcoxon estimator utilizes the norm

$$\begin{aligned} \|\mathbf{v}\|_{\mathrm{W}} &= \sum_{i=1}^n a(R(v_i))v_i \\ &= \sum_{i=1}^n a(i)v_{(i)}, \quad \mathbf{v} \in \mathbb{R}^n, \end{aligned} \tag{2.3}$$

where $R(v_i)$ denotes the rank of v_i among $v_1, \ldots, v_n$, $a(i) = \varphi(i/(n+1))$ and $\varphi(u) = \sqrt{12}\,[u - (1/2)]$. The second representation is based on the relationship between ranks and order statistics. The function $\varphi(u)$ is the score function. Essentially any nondecreasing function on $(0, 1)$ can be used, but in this paper we use only this linear score function. The Wilcoxon estimate of $\boldsymbol{\beta}$ is a vector $\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$ such that $\widehat{\boldsymbol{\beta}}_{\mathrm{W}} = \mathrm{Arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}_c\boldsymbol{\beta}\|_{\mathrm{W}}$. Note that the geometry of the Wilcoxon estimate is similar to the geometry of the LS estimate; only the norm has been changed. To determine the estimating equations of the Wilcoxon estimate, using the second expression in (2.3) it is easy to obtain the partial derivatives and to show that the Wilcoxon estimate solves the estimating equations $\mathbf{X}_c'\mathbf{a}(\mathbf{Y} - \mathbf{X}_c\boldsymbol{\beta}) = \mathbf{0}$, where $\mathbf{a}(\mathbf{Y} - \mathbf{X}_c\boldsymbol{\beta})$ denotes the vector with ith component $a[R(Y_i - \mathbf{x}_{ci}'\boldsymbol{\beta})]$. The solution cannot be obtained in closed form, but there are several algorithms available to obtain the solution, as discussed in Section 1. As our estimate of α, we use the median of the residuals given by $\widehat{\alpha}_S = \mathrm{med}_{1\le i\le n}\{Y_i - \mathbf{x}_{ci}'\widehat{\boldsymbol{\beta}}_{\mathrm{W}}\}$. Under reg-

ularity conditions, the large sample distribution of the Wilcoxon estimator is given by:

$$\begin{pmatrix}\widehat{\alpha}_{\mathrm{W}}\\ \widehat{\boldsymbol{\beta}}_{\mathrm{W}}\end{pmatrix} \text{ has an approximate}$$
$$N\left(\begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix}, \begin{bmatrix}\tau_S^2/n & \mathbf{0}'\\ \mathbf{0} & \tau^2(\mathbf{X}_c'\mathbf{X}_c)^{-1}\end{bmatrix}\right) \text{ distribution,} \tag{2.4}$$

where τ and τ_S are the scale parameters

$$\tau = \left(\sqrt{12}\int f^2(t)\,dt\right)^{-1} \quad \text{and}$$
$$\tau_S = (2f(0))^{-1}. \tag{2.5}$$

Estimates of the scale parameters τ and τ_S were discussed by Hettmansperger and McKean (1998).

Because the variances of the LS and Wilcoxon estimators differ only in the constant of proportionality (σ^2 for LS and τ^2 for the Wilcoxon), it is easy to obtain asymptotic confidence intervals for the regression coefficients based on the Wilcoxon estimator. They are the same as the usual LS confidence intervals except that $\widehat{\tau}$ replaces $\widehat{\sigma}$; that is, an asymptotic $(1-\alpha)100\%$ confidence interval for the parameter β_j, for $j = 1, \ldots, p$, is

$$\widehat{\beta}_{\mathrm{W},j} \pm t_{\alpha/2,n-(p+1)}\widehat{\tau}\sqrt{(\mathbf{X}_c'\mathbf{X}_c)^{-1}_{jj}},$$

where $t_{\alpha/2,n-(p+1)}$ denotes the $\alpha/2$ upper critical point of a t distribution with $n-(p+1)$ degrees of freedom and $(\mathbf{X}_c'\mathbf{X}_c)^{-1}_{jj}$ is the jth diagonal entry of $(\mathbf{X}_c'\mathbf{X}_c)^{-1}$. Finite sample studies show that the use of t-critical values yields empirical confidences close to the nominal $1-\alpha$.

EXAMPLE 2.1 (Water wheel experiment). This data set was discussed by Abebe et al. (2001). For the experiment, mice are placed in a wheel that is partially submerged in water. If they keep the wheel moving, they will avoid the water. The response (Y) is the number of wheel revolutions per minute. There are two groups: a placebo group and a treated group, where the mice are under the influence of a drug. Both groups contain 10 mice each. The predictor in this case is the dummy variable x_i which is 0 if the ith mouse is from the placebo group and is 1 if it is from the treated group. Assume a simple linear model

$$Y_i = \alpha + x_i\beta + \varepsilon_i, \quad i = 1, 2, \ldots, 20.$$

The slope parameter β is the shift in locations between the two groups. The Mann–Whitney–Wilcoxon is the traditional nonparametric analysis of this problem. In this case the estimate of the shift in locations is given by the Hodges–Lehmann estimate, which is the median of all $10 \times 10 = 100$ differences between the treated and placebo responses. By writing out the Wilcoxon estimating equations for this model, we see that the Wilcoxon regression estimate of β is indeed this Hodges–Lehmann estimator. So the Wilcoxon regression analysis generalizes the traditional nonparametrics analysis.

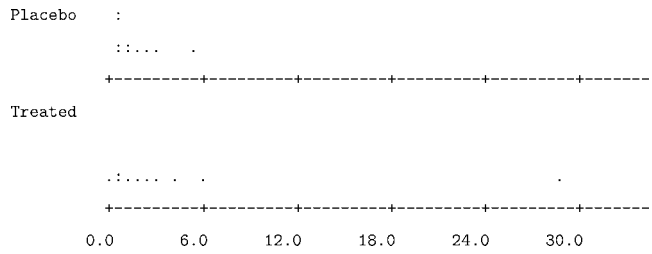

FIG. 1. *Comparison dot plots of the treatment groups in the water wheel experiment.*

A comparison dot plot of the data is given in Figure 1. We see immediately that there is one gross outlier in the treated group; however, there seems to be little or no effect due to the treatment.

Table 1 contains the Wilcoxon and LS estimates of slope and approximate 95% confidence intervals along with t ratios (estimate divided by standard error) and the associated p value for a two-sided test. The Wilcoxon analysis reflects the comparison dot plot. Based on its estimate and confidence interval, there does not appear to be a treatment effect. The LS estimate, on the other hand, has been unduly affected by the outlier. Its estimate of shift is far from reality. At the 5% level it is not significant; however, its p value of 0.28 belies the dot plots. In a discovery setting where large significance levels are often employed, the LS p value might allow further tests on this drug.

TABLE 1
Summaries of the LS and Wilcoxon fits for the water wheel data; the p values correspond to two-sided tests

Analysis	Estimate	Confidence interval	t Ratio	p Value	Scale ($\widehat{\tau}$ or $\widehat{\sigma}$)
Wilcoxon	0.50	$(-1.23, 2.24)$	0.60	0.55	1.85
LS	3.11	$(-2.71, 8.93)$	1.12	0.28	6.19

2.2 Efficiency and Robustness

Both the LS and Wilcoxon estimators have asymptotic distributions centered around the true $\boldsymbol{\beta}$, so we can compare their asymptotic relative efficiencies (AREs) in terms of their asymptotic variances. Hence, for $j = 1, \ldots, p$, the ARE between $\widehat{\beta}_{\mathrm{LS},j}$ and $\widehat{\beta}_{\mathrm{W},j}$ is

$$e(\widehat{\beta}_{\mathrm{W},j}, \widehat{\beta}_{\mathrm{LS},j}) = \frac{\sigma^2}{\tau^2}. \tag{2.6}$$

If the error distribution is normal with variance σ^2, then $\tau^2 = \sigma^2/(3/\pi)$. Hence, the ARE in this case is $3/\pi = 0.955$. Thus under normal errors, the Wilcoxon estimator is 95% as efficient as LS procedures. Thus there is only a 5% loss in efficiency if the Wilcoxon estimator is used and the error distribution is actually normal. On the other hand, if the true distribution has tails heavier than the normal, then this efficiency is usually much larger than 1. As an example of such a distribution, consider the contaminated normal distribution. Suppose $(1-\epsilon)100\%$ of the time we sample from a standard normal distribution, while $\epsilon 100\%$ of the time we sample from the normal distribution with mean 0 and standard deviation σ_c. The scale parameters σ and τ are easily calculated. For $\sigma_c = 3$, Table 2 displays the efficiencies between the Wilcoxon and LS estimators for different values of ϵ. Even at 1% contamination, the Wilcoxon is more efficient than the LS.

For a given data set we call the estimate of the ARE the *empirical measure of efficiency or precision* for that data set, that is,

$$\widehat{e}(\mathrm{W}, \mathrm{LS}) = \frac{\widehat{\sigma}^2}{\widehat{\tau}^2}. \tag{2.7}$$

In Example 2.1, $\widehat{e}(\mathrm{W}, \mathrm{LS}) = (6.19/1.85)^2 = 11.57$, so that the Wilcoxon analysis is estimated to be 11.57 times as efficient as the LS analysis for these data.

2.3 Influence Functions

Briefly, the *influence function* of an estimator measures the change in the estimator when an outlier is added to the data. First consider the LS estimator. Using model (2.1), we can write the LS estimate of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \boldsymbol{\beta} + (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\boldsymbol{\varepsilon}$. If we assume that $n^{-1}\mathbf{X}_c'\mathbf{X}_c \to \boldsymbol{\Sigma}$, then we have the asymptotic representation $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1}(1/\sqrt{n})\mathbf{X}_c'\boldsymbol{\varepsilon} + o_p(1)$ for the LS estimator. For convenience, consider the simple linear model. Then this representation reduces to

$$\sqrt{n}(\widehat{\beta}_{\mathrm{LS}} - \beta) = c_x^{-2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ci}\varepsilon_i + o_p(1), \tag{2.8}$$

where c_x^2 is the scalar (simple linear case) in place of $\boldsymbol{\Sigma}$.

Now suppose we add a point (x^*, ε^*) to the data set. Think of it as an outlier. How does this point affect the LS estimate? Let $\widehat{\beta}_n$ denote the LS fit for the original data and let $\widehat{\beta}_{n+1}$ denote the LS fit when the outlier is added. Using the representation (2.8) and a few steps of algebra, we get

$$\frac{\widehat{\beta}_{n+1} - \widehat{\beta}_n}{1/n} \doteq c_x^{-2}x^*\varepsilon^*, \tag{2.9}$$

which is the relative change in the LS estimator due to the outlier. If ε^* is an outlier, then, based on the model, this leads to an outlier in the Y space. If x^* is an outlier, we say we have an outlier in the x space. Notice that the LS estimator is seriously affected by outliers in either space. In fact, the change in the LS estimator is unbounded in both spaces. Because of this unboundedness, we say that the LS estimator is not *robust*. The LS influence function predicts the poor LS analysis of Example 2.1.

The derivation of (2.9) can be made rigorous; see, for example, Hettmansperger and McKean (1998). The resulting analogue is called the *influence function* of the estimator. In terms of the multiple regression model the influence function of the LS estimator at the point $(\mathbf{x}', \varepsilon)$ is given by $\mathrm{IF}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}; \mathbf{x}, \varepsilon) = \boldsymbol{\Sigma}^{-1}\varepsilon\mathbf{x}$.

In the same way, the influence function can be developed for the Wilcoxon estimator based on its asymptotic representation, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{W}} - \boldsymbol{\beta}) = \tau\boldsymbol{\Sigma}^{-1}(1/\sqrt{n}) \cdot \mathbf{X}_c'\boldsymbol{\varphi}[F(\boldsymbol{\varepsilon})] + o_p(1)$, where $\boldsymbol{\varphi}[F(\boldsymbol{\varepsilon})]$ denotes the vector whose ith component is $\varphi[F(\varepsilon_i)]$. Recall that $F(t)$ is the c.d.f. of the random errors and that $\varphi(u)$ is the linear score function $\sqrt{12}\,[u - (1/2)]$ which is defined on $(0, 1)$. Hence, unlike the LS asymptotic representation, the contribution of the random errors to the representation for the Wilcoxon estimator is bounded. This carries over to the influence function of $\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$, which for the multiple regression model is given by $\mathrm{IF}(\widehat{\boldsymbol{\beta}}_{\mathrm{W}}; \mathbf{x}, \varepsilon) = \tau\boldsymbol{\Sigma}^{-1}\boldsymbol{\varphi}[F(\boldsymbol{\varepsilon})]\mathbf{x}$.

Thus the influence function of $\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$ is bounded in the Y space. So the Wilcoxon estimator is less sensitive to outliers in the Y space, as verified in Example 2.1, but

TABLE 2
Efficiencies of the Wilcoxon and LS methods for the contaminated normal distribution

ϵ	0.00	0.01	0.03	0.05	0.10	0.15
$e(\mathrm{W}, \mathrm{LS})$	0.955	1.009	1.108	1.196	1.373	1.497

note that it is not bounded in the x space. In Section 4 we will discuss a generalization of the Wilcoxon estimator which is bounded in both spaces.

2.4 Breakdown Point

An additional robustness concept will prove helpful. Briefly, the proportion of bad data which an estimator can tolerate before becoming completely meaningless is called the *finite sample breakdown point* of the estimator. If this converges to a limit as $n \to \infty$, we call the limit the *breakdown point* of the estimator.

For example, consider a sample of size n for the one-sample location problem. If we move one sample item to infinity, then the sample mean moves toward infinity also, that is, the finite sample breakdown of the sample mean is $1/n$ which converges to 0 as $n \to \infty$. So the sample mean has breakdown point 0. Next consider the sample median. We have to move half of the data to infinity to move the sample median to infinity. Hence, the breakdown point of the sample median is $1/2$. For the location model, this is the highest breakdown possible.

It is clear from their influence functions that the breakdown point of either the LS or the Wilcoxon regression estimator is 0. We need only move one x value to infinity to make both estimators meaningless. In Section 4, we generalize the Wilcoxon estimates to obtain high breakdown estimators.

3. LINEAR HYPOTHESES

In this section, we consider robust tests of general linear hypotheses. As in the previous sections, the Wilcoxon procedure is suggested from a simple geometric point of view. As before, let the responses follow the linear model (2.1). Our hypotheses of interest are collections of independent, linear constraints on the regression parameters. More precisely, a general linear hypothesis and its alternative are given by

$$(3.1) \qquad H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0} \quad \text{versus} \quad H_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{0},$$

where $\mathbf{A}$ is a $q \times p$ specified matrix of full row rank q. The rows of $\mathbf{A}$ form the linear constraints.

3.1 Geometry of the Testing

In hypotheses testing we consider model (2.1) to be the *full model*. Let V_F (F for full) denote the column space of $\mathbf{X}$. For the hypotheses (3.1), the *reduced model* is the full model subject to $H_0 : V_R = \{\mathbf{v} \in V_F : \mathbf{v} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}\}$, where the R stands for reduced. Recall that V_R is a subspace of V_F of dimension $p - q$.

Suppose we have a norm $\|\cdot\|$ for fitting models. Then based on geometry, a simple test procedure can be described. Let $\widehat{\boldsymbol{\eta}}_F$ be the full model fit based on the norm; that is, $\widehat{\boldsymbol{\eta}}_F = \operatorname{Arg\,min} \|\mathbf{Y} - \boldsymbol{\eta}\|$, $\boldsymbol{\eta} \in V_F$. Then the distance between $\mathbf{Y}$ and the subspace V_F is $d(\mathbf{Y}, V_F) = \|\mathbf{Y} - \widehat{\boldsymbol{\eta}}_F\|$. Likewise, we next fit the reduced model and let $d(\mathbf{Y}, V_R)$ denote the distance between $\mathbf{Y}$ and the reduced model space V_R. Because we are minimizing over a smaller subspace, $d(\mathbf{Y}, V_R) \geq d(\mathbf{Y}, V_F)$. An intuitive test statistic is the reduction in distances, passing from the reduced to the full model, that is, $\mathrm{RD}_{\|\cdot\|} = d(\mathbf{Y}, V_R) - d(\mathbf{Y}, V_F)$, where $\mathrm{RD}_{\|\cdot\|}$ denotes reduction in distance. Small values of $\mathrm{RD}_{\|\cdot\|}$ indicate H_0, while large values indicate H_1. Hence, the corresponding test is

$$\text{reject } H_0 \text{ in favor of } H_1 \text{ if } \mathrm{RD}_{\|\cdot\|} \geq c,$$

where c must be determined. The reduction in distance is standardized by an estimate of scale or variance.

Suppose the Euclidean norm is used. If the reduction in (squared) Euclidean distances is standardized by $\widehat{\sigma}^2$, the usual estimate of σ^2, then the test statistic for the hypotheses (3.1) is given by $F_{\mathrm{LS}} = [\mathrm{RD}_{\mathrm{LS}}/q]/\widehat{\sigma}^2$. As discussed in Section 3.2, under the null hypothesis F_{LS} has an approximate F distribution with q and $n - (p + 1)$ degrees of freedom. The usual approximate α rejection rule is reject H_0 in favor of H_1 if $F_{\mathrm{LS}} \geq F_{\alpha,q,n-p-1}$. If the underlying error distribution is normal, then F_{LS} is the likelihood ratio test statistic and this rejection rule has exact level α.

Suppose the Wilcoxon norm $\|\cdot\|_{\mathrm{W}}$ [(2.3)] is chosen. Let $\widehat{\mathbf{Y}}_{F,\mathrm{W}}$ denote the Wilcoxon full model fit as discussed in Section 2 and denote the Wilcoxon distance between V_F and $\mathbf{Y}$ by $d_{\mathrm{W}}(\mathbf{Y}, V_F) = \|\mathbf{Y} - \widehat{\mathbf{Y}}_{F,\mathrm{W}}\|_{\mathrm{W}}$. Similarly, the reduced model fit is given by

$$\widehat{\mathbf{Y}}_{\mathrm{W},R} = \operatorname{Arg} \min_{\boldsymbol{\eta} \in V_R} \|\mathbf{Y} - \boldsymbol{\eta}\|_{\mathrm{W}}.$$

Let $d_{\mathrm{W}}(\mathbf{Y}, V_R) = \|\mathbf{Y} - \widehat{\mathbf{Y}}_{R,\mathrm{W}}\|_{\mathrm{W}}$ denote the distance between V_R and $\mathbf{Y}$. The Wilcoxon test statistic (reduction in distance) is $\mathrm{RD}_{\mathrm{W}} = d_{\mathrm{W}}(\mathbf{Y}, V_R) - d_{\mathrm{W}}(\mathbf{Y}, V_F)$. Note that RD_{W} is a reduction in scale, not variance. The appropriate standardization (see Section 3.2) is by an estimate of τ. The usual test statistic is of the form

$$(3.2) \qquad F_{\mathrm{W}} = \frac{\mathrm{RD}_{\mathrm{W}}/q}{\widehat{\tau}/2}.$$

An approximate level α test is reject H_0 in favor of H_1 if $F_{\mathrm{W}} \geq F_{\alpha,q,n-p-1}$.

TABLE 3
Summary of analysis for plank data

Source	Strain	Gender	Age	$S \times G$	$S \times A$	$G \times A$	$S \times G \times A$	Scale
F_{LS}	7.74*	2.63	0.92	1.61	1.28	2.81	0.17	1.10
F_{W}	12.23*	5.27*	1.12	0.11	1.80	3.21*	0.33	0.70

EXAMPLE 3.1 (Plank balance). Abebe et al. (2001) reported the results of a three-way layout obtained from a neurological experiment, where the response of interest is the log time required for a mouse to exit a narrow elevated wooden plank. The experimenters were interested in assessing the effects and associated interactions of three factors: mouse strain (S) with levels Tg+ and Tg−; gender (G) with levels female and male; and age (A) with levels aged, middle and young.

Let Y_{ijkl} denote the response for the lth repetition of the ith level of S, the jth level of G and the kth level of A. As a full model, consider the cell mean model

$$Y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl}, \quad i = 1, 2,\ j = 1, 2, \quad k = 1, 2, 3, l = 1, \ldots, n_{ijk}. \tag{3.3}$$

The design is unbalanced, but there are at least two replications in each cell. The total sample size is $n = 64$. The data can be found at the web site discussed in Section 1.

Denote the $n \times 1$ vector of responses by $\mathbf{Y} = [Y_{ijkl}]$, where the data are entered so that the subscript l runs the fastest, followed by k, and so on. Denote the 12×1 vector of means by $\boldsymbol{\mu} = [\mu_{ijk}]$. Let $\mathbf{W}$ be the 64×12 incidence matrix of cell membership. Then we can write model (3.3) as $\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$.

To illustrate the description of the robust test, we consider the gender main effect hypothesis, that is, the average effect of gender is 0. Because of how the data were arranged in $\mathbf{Y}$, the associated hypothesis matrix is the row vector $\mathbf{A} = [1\,1\,1\,-1\,-1\,-1\,1\,1\,1\,-1\,-1\,-1]$. The null hypothesis can be written as $H_0 : \mathbf{A}\boldsymbol{\mu} = 0$. The Wilcoxon distance between the vector of responses $\mathbf{Y}$ and the full model space V_F is $d_{\mathrm{W}}(\mathbf{Y}, V_F) = 58.28$, while the distance from $\mathbf{Y}$ to the reduced model space is $d_{\mathrm{W}}(\mathbf{Y}, V_R) = 60.13$. The estimate of τ is $\widehat{\tau} = 0.702$. Thus the value of the test statistic (3.2) is $F_{\mathrm{W}} = [(60.13 - 58.28)/1]/(0.702/2) = 5.27$. The approximate p value is 0.026. Thus gender seems to have an effect on log time for the mouse to exit the plank. Of course interaction effects should be considered before main effects.

Table 3 summarizes the LS and Wilcoxon tests for the main effects and interactions hypotheses for this data set. Note that the analyses do differ. The only significant effect of the LS analysis is mouse strain. On the other hand, the Wilcoxon analysis indicates that gender plays a role as does age also, due to its interaction with gender. Abebe et al. (2001) showed that the usual cell mean profile plot clearly indicates the interaction between gender and age, while the Wilcoxon Studentized residual plot clearly shows numerous outliers. Practical interpretations based on the Wilcoxon and LS analysis would be quite different.

3.2 Asymptotic Distribution Theory and Relative Efficiency

Algebraically, the reduction in variance form of the LS test statistic can be written as

$$\frac{\mathrm{RD}_{\mathrm{LS}}}{\sigma^2} = (\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}})'[\sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}.$$

The reduction in distance form of the Wilcoxon test statistic can be written as

$$\frac{\mathrm{RD}_{\mathrm{W}}}{\tau/2} = (\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{W}})'[\tau^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{W}} + o_p(1).$$

It is clear from these representations and the asymptotic distributions of $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$ [(2.2) and (2.4), respectively] that both random variables $\mathrm{RD}_{\mathrm{LS}}/\sigma^2$ and $\mathrm{RD}_{\mathrm{W}}/(\tau/2)$ have, under regularity conditions, an asymptotic χ^2 distribution with q degrees of freedom under the null hypothesis. The approximate distribution theory for the test statistics F_{LS} and F_{W} cited previously is based on this and the fact that the estimators of scale used in the denominators of the test statistics are consistent for their respective parameters. Small sample studies have indicated that the use of F critical values instead of the χ^2 critical values leads to empirical levels closer to nominal α levels.

The influence functions of these test statistics are discussed in Hettmansperger and McKean (1998). Similar to their counterparts in estimation, the influence function for F_{W} is bounded in the Y space, but not the $\mathbf{x}$ space. The influence function for F_{LS} is unbounded

in both spaces. Hence, the influence functions predict the behavior of these test statistics in the last example.

Efficiency results for these tests can easily be formulated based on the noncentral χ^2 asymptotic distributions of the test statistics under a sequence of local alternatives. The ARE is the ratio of the test statistics' noncentrality parameters, which reduces to the ratio (2.6); that is, the ARE for the test statistics is the same as the ARE for the corresponding estimators. In particular, the high efficiency of the Wilcoxon rank tests in the simple location models extends to the the Wilcoxon test statistic F_W for general linear hypotheses. In Example 3.1, the empirical measure of efficiency [(2.7)] is $(1.098/0.7019)^2 = 2.45$. Hence the Wilcoxon analysis is estimated to be 2.5 times more efficient than the LS analysis on this data set.

4. HIGH BREAKDOWN ESTIMATORS

Although the Wilcoxon analysis provides an attractive, robust alternative to the traditional LS analysis, the analysis is not robust in the $\mathbf{x}$ space. In this section, we briefly discuss generalizations of the Wilcoxon analysis which are robust in both the Y and $\mathbf{x}$ spaces and which have positive breakdown.

Consider the linear model (2.1). We begin with the simple identity concerning the Wilcoxon norm [(2.3)]:

$$\|\mathbf{v}\|_W = \sum_{i=1}^{n} a(R(v_i))v_i = \frac{\sqrt{3}(n+1)}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|v_i - v_j|, \quad \mathbf{v}\in\mathbb{R}^n.$$

Recall that the Wilcoxon estimate of $\boldsymbol{\beta}$ minimizes $\|\mathbf{Y} - \mathbf{X}_c\boldsymbol{\beta}\|_W$. Thus all the absolute differences in residuals $|(Y_i - \mathbf{x}'_{ci}\boldsymbol{\beta}) - (Y_j - \mathbf{x}'_{cj}\boldsymbol{\beta})|$ receive the same weight. If we are in a situation, though, where some of the $\mathbf{x}_{ci}$ are more influential than others, we may want to downweight the contribution of such points. That is, select weights w_i, $i = 1, 2, \ldots, n$, and choose the estimator $\widehat{\boldsymbol{\beta}}^*$ given by

$$\widehat{\boldsymbol{\beta}}^* = \operatorname{Arg\,min}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \big|(Y_i - Y_j) - (\mathbf{x}_{ci} - \mathbf{x}_{cj})'\boldsymbol{\beta}\big|. \tag{4.1}$$

Note that the function being minimized is a convex function of $\boldsymbol{\beta}$. Consider the following example of a simple regression data set:

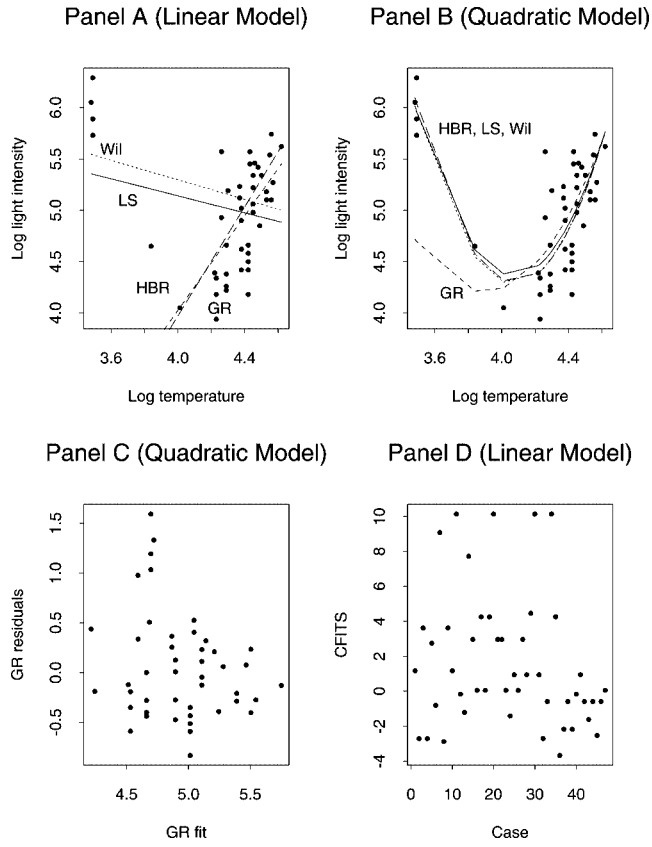

FIG. 2. *Plots for Stars data*: *Panel* A, *fits of linear model*; *Panel* B, *fits of quadratic model*; *Panel* C, *GR residual plot for quadratic model*; *Panel* D, *casewise CFITS for the linear model*.

EXAMPLE 4.1 (Stars data). This data set is drawn from an astronomy study on the star cluster CYG OB1, which contains 47 stars; see Rousseeuw and Leroy (1987) for discussion and the data. The response is the logarithm of the light intensity of the star, while the independent variable is the logarithm of the temperature of the star. The data are shown in Panel A of Figure 2. Note that four of the stars, called giants, are outliers in factor space, while the rest of the stars fall in a point cloud. Panel A shows also the overlay plot of the LS and Wilcoxon fits. Note that the cluster of four outliers in the $\mathbf{x}$ space has exerted such a strong influence on the fits that it has drawn the fits toward the cluster. This behavior is predictable based on the influence functions of these estimates.

With regard to weights, it seems reasonable to downweight points far from the center of the data. The leverage values $h_i = n^{-1} + \mathbf{x}'_{ci}(\mathbf{X}'_c\mathbf{X}_c)^{-1}\mathbf{x}_{ci}$, for $i = 1, \ldots, n$, measure distance (Mahalanobis) from the center relative to the scatter matrix $\mathbf{X}'_c\mathbf{X}_c$. Leverage values, though, are based on means and the usual (LS)

variance–covariance scatter matrix, which are not robust estimators. There are several robust estimators of location and scatter from which to choose, including the high breakdown *minimum covariance determinant* (MCD), which is an ellipsoid that covers about half of the data and yet has minimum determinant. Although computationally intensive, Rousseeuw and Van Driessen (1999) presented a fast computational algorithm for the MCD. Let $\mathbf{v}_c$ denote the center of the ellipsoid. Letting $\mathbf{V}$ denote the MCD, the robust distances are given by $v_{ni} = (\mathbf{x}_i - \mathbf{v}_c)'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{v}_c)$. We define the associated weights by $w_i = \min\{1, c/v_{ni}\}$, where c is usually set at the 95th percentile of the $\chi^2(p)$ distribution. Note that "good" points generally have weights 1. The estimator $\widehat{\boldsymbol{\beta}}^*$ [(4.1)] of $\boldsymbol{\beta}$ obtained with these weights is called a generalized R (GR) estimator. For the stars data, the resulting GR fit goes through the point cloud as depicted in Panel A of Figure 2. In general, this GR estimate has a bounded influence function in both the Y and the $\mathbf{x}$ spaces, and a positive breakdown.

Due to the downweighting, though, the GR estimator is less efficient than the Wilcoxon estimator. At times, the loss in efficiency can be severe. We next discuss weights which can regain some of this efficiency. To motivate this second choice of weights, suppose we lacked subject knowledge concerning the stars data. Based on the scatter plot, we may decide to fit a quadratic model. The plots of the LS, Wilcoxon and GR fits for the quadratic model are found in Panel B of Figure 2. The quadratic fits based on the LS and Wilcoxon estimates follow the curvature in the data, while the GR fit misses the curvature. For this case, the outliers are good data points, but the GR fit uses weights which downweight this important information. The pattern in the GR residual plot (Panel C) is not random, but the plot does not indicate how to proceed with model selection. This is often true for residual plots based on high breakdown fits; see McKean, Sheather and Hettmansperger (1993).

As do the GR weights, our second class of weights uses the MCD to determine weights in the $\mathbf{x}$ space, but it also uses residual information from the Y space. The residuals are based on a high breakdown initial estimate of the regression coefficients. We have chosen to use the *least trim squares* (LTS) estimate, which is $\operatorname{Arg\,min}\sum_{i=1}^{h}[Y - \alpha - \mathbf{x}'\boldsymbol{\beta}]^2_{(i)}$, where $h = [n/2] + 1$ and (i) denotes the ith ordered residual; see Rousseeuw and Van Driessen (1999). Let $\widehat{\mathbf{e}}_0$ denote the residuals from this initial fit.

Define the function $\psi(t)$ by $\psi(t) = 1$, t or -1 according as $t \geq 1$, $-1 < t < 1$ or $t \leq -1$. Let σ be estimated by the initial scaling estimate $\text{MAD} = 1.483\,\text{med}_i\,|\widehat{e}_i^{(0)} - \text{med}_j\{\widehat{e}_j^{(0)}\}|$. Let $Q_i = (\mathbf{x}_i - \mathbf{v}_c)' \times \mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{v}_c)$ and let

$$m_i = \psi\left(\frac{b}{Q_i}\right) = \min\left\{1, \frac{b}{Q_i}\right\}.$$

Consider the weights

$$\widehat{b}_{ij} = \min\left\{1, \frac{c\widehat{\sigma}}{|\widehat{e}_i|}\frac{\widehat{\sigma}}{|\widehat{e}_j|}\min\left\{1, \frac{b}{\widehat{Q}_i}\right\}\min\left\{1, \frac{b}{\widehat{Q}_j}\right\}\right\},$$

where the tuning constants b and c are both set at 4. From this point of view, it is clear that these weights downweight both outlying points in factor space and outlying responses. Note that the initial residual information is a multiplicative factor in the weight function. Hence, a good leverage point will generally have a small (in absolute value) initial residual which will offset its distance in factor space.

The HBR Wilcoxon estimate is then defined as $\widehat{\boldsymbol{\beta}}_{\text{HBR}} = \operatorname{Arg\,min}\sum_{i,j}\widehat{b}_{ij}|Y_i - Y_j - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}|$. Once the weights are determined, the estimates are obtained by minimizing a convex function. Note that for the stars data, the HBR estimate fits the point cloud for the linear model but fits the quadratic model similarly to the Wilcoxon estimate.

In general, the HBR estimator has a 50% breakdown point, provided the initial estimates used in forming the weights have 50% breakdown. Further, its influence function is a bounded function in both the Y and the $\mathbf{x}$ spaces, is continuous everywhere and converges to zero as $(\mathbf{x}^*, Y^*)$ get large in any direction. The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\text{HBR}}$ is asymptotically normal. As with all high breakdown estimates, $\widehat{\boldsymbol{\beta}}_{\text{HBR}}$ is less efficient than the Wilcoxon estimates, but it regains some of the efficiency loss of the GR estimate.

4.1 Diagnostics to Differentiate between HBR and Wilcoxon Fits

For a given data set, highly efficient robust estimates and high breakdown estimates can produce very different fits. This can be due to influential points in factor space and/or curvature. We present diagnostics which indicate, first, whether the HBR and Wilcoxon fits differ and, second, if they do differ, what cases are involved in the discrepancy.

First, as with the Wilcoxon estimates, estimate the intercept by the median of the HBR residuals. Then the difference in regression estimates between the HBR

and Wilcoxon estimates is the vector $\widehat{\mathbf{b}}_{\mathrm{W}} - \widehat{\mathbf{b}}_{\mathrm{HBR}}$. An effective standardization is the estimate of the variance–covariance of $\widehat{\mathbf{b}}_{\mathrm{W}}$. A statistic which measures the total difference in the fits of $\widehat{\mathbf{b}}_{\mathrm{W}}$ and $\widehat{\mathbf{b}}_{\mathrm{HBR}}$ is $\mathrm{TDBET}AS_{\mathrm{R}} = (\widehat{\mathbf{b}}_{\mathrm{W}} - \widehat{\mathbf{b}}_{\mathrm{HBR}})'\widehat{\mathbf{A}}_{\mathrm{W}}^{-1}(\widehat{\mathbf{b}}_{\mathrm{W}} - \widehat{\mathbf{b}}_{\mathrm{HBR}})$, where $\mathbf{A}_{\mathrm{W}}$ is the limiting Wilcoxon covariance matrix in formula (2.4). Large values of $\mathrm{TDBET}AS_{\mathrm{R}}$ indicate a discrepancy between the fits. A useful cutoff value is $(4(p+1)^2)/n$.

If $\mathrm{TDBET}AS_{\mathrm{R}}$ exceeds its benchmark, then usually we want to determine the individual cases causing this discrepancy in the fits. Let $\widehat{y}_{\mathrm{W},i} = \widehat{\alpha}_{\mathrm{W}} + \mathbf{x}'\widehat{\boldsymbol{\beta}}_{\mathrm{W}}$ and $\widehat{y}_{\mathrm{HBR},i} = \widehat{\alpha}_{\mathrm{HBR}} + \mathbf{x}'\widehat{\boldsymbol{\beta}}_{\mathrm{HBR}}$ denote the respective fitted values for the ith case. A statistic which detects the observations that are fitted differently is $\mathrm{CFIT}S_{\mathrm{R},i} = (\widehat{y}_{\mathrm{R},i} - \widehat{y}_{\mathrm{HBR},i})/\sqrt{n^{-1}\widehat{\tau}_S^2 + h_{c,i}\widehat{\tau}^2}$. An effective *benchmark for* $\mathrm{CFIT}S_{\mathrm{R},i}$ is $2\sqrt{(p+1)/n}$. We should note here that the objective of the diagnostic CFIT S is *not* outlier deletion. Rather the intent is to identify the *critical few* data points for closer study, because these critical few points often largely determine the outcome of the analysis or the direction that further analysis should take. In this regard, the proposed benchmarks are meant as a heuristic aid, not a boundary to some formal critical region.

For the simple linear model of the stars data (Example 4.1), the diagnostic $\mathrm{TDBET}AS_{\mathrm{R}}$ has the value 109.9, which greatly exceeds the benchmark (0.34) and, hence, numerically indicates that the fits differ. Panel D of Figure 2 shows the casewise diagnostic $\mathrm{CFIT}S_{\mathrm{R},i}$ versus case. In this plot, the four giant stars clearly stand out from the rest. The plot shows that the fits for two other stars also differ. These are the stars between the giant stars and the rest of the stars as shown in Panel A. For the quadratic model, $\mathrm{TDBET}AS_{\mathrm{R}} = 0.217$, which is less than the benchmark value of 0.776. McKean, Naranjo and Sheather (1999) extended these diagnostic procedures to investigate differences between LS and robust estimates.

5. CONCLUSION

In these three lectures, we have presented a robust analysis of linear models. As with its counterpart, the traditional analysis based on LS estimates, this robust analysis offers the user a unified methodology for the analysis of linear models. For designs without outliers in the **x** space, this analysis is robust and highly efficient. As with the LS analysis, the geometry of the analysis is based on a norm, so it has the same interpretation as the LS analysis. For designs with outliers in factor space it can be easily generalized to an analysis based on a high breakdown estimator which is robust in both spaces. Further, simple diagnostic procedures exist to explore the differences between the highly efficient and high breakdown estimates. We have presented examples which show that the robust analysis is more effective than the traditional analysis in discovering alternatives and patterns in the presence of outliers or underlying error distributions with thick tails. For such data, practical interpretations based on the Wilcoxon and LS analysis can be quite different.

REFERENCES

ABEBE, A., CRIMIN, K., MCKEAN, J. W., HAAS, J. V. and VIDMAR, T. J. (2001). Rank-based procedures for linear models: Applications to pharmaceutical science data. *Drug Information J.* **35** 947–971.

CRIMIN, K., ABEBE, A. and MCKEAN, J. W. (2003). Robust general linear models and graphics via a user interface. Unpublished manuscript.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. J. (1986). *Robust Statistics. The Approach Based on Influence Functions*. Wiley, New York.

HETTMANSPERGER, T. P. and MCKEAN, J. W. (1998). *Robust Nonparametric Statistical Methods*. Arnold, London.

HETTMANSPERGER, T. P., MCKEAN, J. W. and SHEATHER, S. J. (2000). Robust nonparametric methods. *J. Amer. Statist. Assoc.* **95** 1308–1312.

HOLLANDER, M. and WOLFE, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

MCKEAN, J. W., NARANJO, J. D. and SHEATHER, S. J. (1999). Diagnostics for comparing robust and least squares fits. *J. Nonparametr. Statist.* **11** 161–188.

MCKEAN, J. W., SHEATHER, S. J. and HETTMANSPERGER, T. P. (1993). The use and interpretation of residuals based on robust estimation. *J. Amer. Statist. Assoc.* **88** 1254–1263.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

ROUSSEEUW, P. J. and VAN DRIESSEN, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

SIMPSON, D. G., RUPPERT, D. and CARROLL, R. J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Amer. Statist. Assoc.* **87** 439–450.

TERPSTRA, J. T. and MCKEAN, J. W. (2004). Rank-based analysis of linear models using R. Technical Report 151, Statistical Computation Lab, Western Michigan Univ.