

# Chapter 1

## Rank-Based Analysis of Linear Models and Beyond: A Review

Joseph W. McKean and Thomas P. Hettmansperger

**Abstract** In the 1940s Wilcoxon, Mann and Whitney, and others began the development of rank based methods for basic one and two sample models. Over the years a multitude of papers have been written extending the use of ranks to more and more complex models. In the late 60s and early 70s Jurečková and Jaeckel along with others provided the necessary asymptotic machinery to develop rank based estimates in the linear model. Geometrically Jaeckel's fit of linear model is the minimization of the distance between the vector of responses and the column space of the design matrix where the norm is not the squared-Euclidean norm but a norm that leads to robust fitting. Beginning with his 1975 thesis, Joe McKean has worked with many students and coauthors to develop a unified approach to data analysis (model fitting, inference, diagnostics, and computing) based on ranks. This approach includes the linear model and various extensions, for example multivariate models and models with dependent error structure such as mixed models, time series models, and longitudinal data models. Moreover, McKean and Kloke have developed R libraries to implement this methodology. This paper reviews the development of this methodology. Along the way we will illustrate the surprising ubiquity of ranks throughout statistics.

**Keywords** Efficiency • Diagnostics • High breakdown fits • Mixed models • Nonlinear models • Nonparametric methods • Optimal scores • Rank scores • Rfit • Robust

---

J.W. McKean (✉)

Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA  
e-mail: [joseph.mckean@wmich.edu](mailto:joseph.mckean@wmich.edu)

T.P. Hettmansperger

Department of Statistics, Penn State University, University Park, PA 16002, USA  
e-mail: [tph@stat.psu.edu](mailto:tph@stat.psu.edu)

## 1.1 Introduction

Our intention in writing the following historical development is to provide our perspective on the evolution of nonparametric methodology (both finite and asymptotic). We will focus on a particular development based on ranks. We will show how beginning with simple rank tests in the 1940s, the area has grown into a coherent group of contemporary statistical procedures that can handle data from increasingly complex experimental designs. Two factors have been essential: theoretical developments especially in asymptotic theory, see Hettmansperger and McKean (2011), and in computational developments, see Kloke and McKean (2014). Statistical inference based on ranks of the data has been shown to be both statistically efficient relative to least squares methods as well as robust. Any history is bound to be selective. We have chosen a line of development that is consistent with the theme of this conference. There is a rich and extensive literature on nonparametric methods. We will confine ourselves to references that directly relate to the history as related to the topics of the conference.

When constructing tests for the median of a continuous population, the simplest nonparametric test is the sign test which counts the number of observations greater than the null hypothesized value of the median. The null and alternative distributions of the sign test statistic are both binomial. In the case of the null hypothesis, the binomial parameter is 0.5, and hence, the null distribution of the sign test statistic does not depend on the population distribution. We call such a test nonparametric or distribution free. The use of the sign test for dichotomous data was first proposed by Arbuthnott (1710).

The modern era for nonparametric or distribution free tests began with the work of Wilcoxon (1945) and Mann and Whitney (1947). Wilcoxon proposed the nonparametric Wilcoxon signed rank test for the median of a symmetric population, and the nonparametric Wilcoxon rank sum test for the difference in population medians. Mann and Whitney (1947) showed that the rank sum test is equivalent to the sign test applied to the pairwise differences across the two samples. Tukey (1949) showed the signed rank test is equivalent to the sign test applied to the pairwise averages from the sample (called the Walsh averages by Tukey). Hence, from the earliest time, we have a connection between rank based methods and the  $L_1$  norm expressed through its derivative, the sign statistic. In what follows we will exploit this connection by considering a rank based norm and its relationship to the  $L_1$  norm. In addition, we will need to include the  $L_2$  norm and least squares for comparison in our discussion.

Noether (1955), based on earlier unpublished work by Pitman (1948), introduced Pitman efficiency for hypothesis tests. Then Hodges and Lehmann (1956, 1960) analyzed the efficiency of various rank tests relative to least squares tests ( $t$ - and  $F$ -tests) and proved the surprising result that the efficiency of the Wilcoxon tests relative to the  $t$ -tests is never less than 0.864, is 0.955 at the normal model, and can be arbitrarily large for heavy tailed model distributions. No longer was a rank test considered quick and dirty with low power. Rank tests now provided a serious alternative to least squares  $t$  tests.

Hodges and Lehmann (1963) next developed estimators based on rank test statistics (R-estimates) and showed that they inherit the efficiency of the rank tests that they were derived from. Because of the connection between the Wilcoxon test statistics and the  $L_1$  norm, the Hodges-Lehmann estimate of location is the median of the pairwise averages and the estimate for the difference in locations is the median of the pairwise differences across the two samples. By the mid-sixties rank tests and estimates for location models, including the one-way layout, were available, and they share the excellent efficiency properties. Robustness was introduced during this time by Huber (1964) followed by the work of Hampel (1974). The basic tools of robustness include the influence function and break down point. Ideally we would like to have estimates that have bounded influence and positive breakdown. Indeed, Wilcoxon R-estimates enjoy precisely these good robustness properties in addition to the excellent efficiency properties mentioned above. For example the breakdown value for the Hodges-Lehmann estimate of location, the median of the pairwise averages, is 0.293 while the breakdown of the sample mean is 0.

Hájek and Šidák (1967) published a seminal work on the rigorous development of rank tests. This was followed many years later by a second edition, Hájek et al. (1999) which extends much of the theory and includes material on R-estimates.

Hence, during the 1960s nonparametric and distribution free rank tests and rank-based estimates for location models were well understood and provided excellent alternatives to least squares methods (means,  $t$ - and  $F$ -tests) from the point of view of both efficiency and robustness. Unfortunately the rank methods did not extend in a straight forward way to the two-way layout with interaction terms. For example, a quick check of text books on nonparametrics written before the mid-seventies did not reference a test for interaction in a two-way layout.

The next step involved the extension of rank methods to linear regression where the two-way layout could be formulated in regression terms and natural rank tests for regression parameters were easy to construct. The rank based statistical methods which require the estimation of nuisance parameters will then be asymptotically distribution free but no longer distribution free for finite samples. The tools for the development of rank regression were provided by Jurečková (1969, 1971) and Jaeckel (1972). Jurečková, in particular, provided the asymptotic theory and Jaeckel provided a rank based dispersion function that when minimized produced R-estimates. McKean (1975) developed corresponding rank tests along with the necessary asymptotic distribution theory for the linear model. In the next several sections we explicitly introduce the linear model and discuss the development of rank based methods and their efficiency and robustness properties. In subsequent sections, we discuss extensions of rank-based analyses to nonlinear models and models with dependent errors.

There is R software available to compute these rank-based analyses. In the examples presented, we discuss some of the R code for the computation of these analyses. The rank-based package for linear models, `Rfit`, (see Kloke and McKean 2012), can be downloaded at CRAN (<http://cran.us.r-project.org/>). Supplemental packages for the additional models discussed in the examples can be downloaded at the site <https://github.com/kloke/>.

## 1.2 Rank-Based Fit and Inference for Linear Models

In this section we will review the univariate linear model and present the rank based norm used to derive the rank based statistical methods along with the basic asymptotic tools. Then we will present the efficiency and robustness results that we mentioned in the introduction, but in more detail. We will also describe some of the rank based methods for residual analysis. For details of this development see Chaps. 3–5 of Hettmansperger and McKean (2011).

Let  $\mathbf{Y}$  denote the  $n \times 1$  vector of observations and assume that it follows the linear model

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (1.1)$$

where  $\mathbf{X}$  is an  $n \times p$  full column rank matrix of explanatory variables,  $\mathbf{1}$  is an  $n \times 1$  vector of ones,  $\boldsymbol{\beta}$  is  $p \times 1$  vector of regression coefficients,  $\alpha$  is the intercept parameter, and  $\mathbf{e}$  is the  $n \times 1$  vector of random errors. Letting  $\mathbf{x}'_i$  denote the  $i$ th row of  $\mathbf{X}$ , we have  $y_i = \alpha + \mathbf{x}'_i\boldsymbol{\beta} + e_i$ . For the theory cited in this section, assume that the random errors are iid with pdf  $f(x)$  and cdf  $F(x)$ , respectively.

A score generating function is a nondecreasing square-integrable function  $\varphi(u)$  defined on the interval  $(0, 1)$  which, without loss of generality, satisfies the standardizing conditions

$$\int_0^1 \varphi(u) du = 0 \quad \text{and} \quad \int_0^1 [\varphi(u)]^2 du = 1. \quad (1.2)$$

We denote the scores by  $a(i) = \varphi[i/(n+1)]$ .

The basis of traditional analysis of most models in practice is the least squares (LS) fit of the model. This fit minimizes the squared-Euclidean distance between the vector of responses and the estimating region, (subspace if it is a linear model). In the same way, the basis for a rank-based analysis is the fit of the model except that a different norm is used other than the Euclidean norm. This norm leads to a robust fit. For a given score function  $\varphi(u)$ , the norm is defined by

$$\|\mathbf{v}\|_\varphi = \sum_{i=1}^n a_\varphi[R(v_i)]v_i, \quad \mathbf{v} \in \mathbb{R}^n. \quad (1.3)$$

Note that this is a pseudo-norm; i.e., it satisfies all properties of the norm except it is invariant to constant shifts, i.e.,  $\|\mathbf{v} + a\mathbf{1}\|_\varphi = \|\mathbf{v}\|_\varphi$  for all  $a$ , where  $\mathbf{1}$  is a vector of  $n$  ones. The counterpart in LS is the squared-Euclidean pseudo-norm  $\sum_{i=1}^n (v_i - \bar{v})^2$ .

For convenience, we define the dispersion function  $D(\boldsymbol{\beta})$  in terms of the pseudo norm  $\|\cdot\|_\varphi$  as

$$D(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_\varphi = \sum_{i=1}^n a[R(y_i - \mathbf{x}'_i\boldsymbol{\beta})](y_i - \mathbf{x}'_i\boldsymbol{\beta}) = \mathbf{a}'[R(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})](\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.4)$$

where  $R(y_i - \mathbf{x}'_i \boldsymbol{\beta})$  denotes the rank of  $y_i - \mathbf{x}'_i \boldsymbol{\beta}$  among  $y_1 - \mathbf{x}'_1 \boldsymbol{\beta}, \dots, y_n - \mathbf{x}'_n \boldsymbol{\beta}$  and  $\mathbf{a}[R(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]$  is the vector with  $i$ th component  $a[R(y_i - \mathbf{x}'_i \boldsymbol{\beta})]$ . Note that ranks are invariant to constant shifts such as an intercept parameter. The rank-based estimator of  $\boldsymbol{\beta}$  is the minimizer

$$\hat{\boldsymbol{\beta}} = \text{Argmin } D(\boldsymbol{\beta}). \quad (1.5)$$

Let  $V_f$  denote the full model subspace of  $R^n$ ; i.e.,  $V_f$  is the range (column space) of  $\mathbf{X}$ . Then  $D(\hat{\boldsymbol{\beta}})$  is the minimum distance between the vector of responses  $\mathbf{Y}$  and the subspace  $V_f$  in terms of the norm  $\|\cdot\|_\varphi$ . For reference, we have

$$D(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\eta} \in V_f} \|\mathbf{Y} - \boldsymbol{\eta}\|_\varphi. \quad (1.6)$$

Note that this minimum distance between  $\mathbf{Y}$  and  $V_f$  is unique; i.e., the minimum distance does not depend on the basis matrix of  $V_f$ .

Denote the negative of the gradient of  $D(\boldsymbol{\beta})$  by

$$\mathbf{S}(\boldsymbol{\beta}) = -\nabla D(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{a}[R(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]. \quad (1.7)$$

Then the estimator also satisfies  $\mathbf{S}(\hat{\boldsymbol{\beta}}) \doteq \mathbf{0}$ . Generally, the intercept parameter is estimated by the median of the residuals; i.e.,

$$\hat{\alpha} = \text{med}_i \{y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}\}. \quad (1.8)$$

Examples of scores functions include:  $\varphi(u) = \sqrt{12}[u - (1/2)]$ , for Wilcoxon rank-based methods;  $\varphi(u) = \text{sgn}[u - (1/2)]$ , for  $L_1$  methods; and  $\varphi(u) = \Phi^{-1}(u)$ , where  $\Phi(t)$  is the standard normal cdf, for normal scores methods. In Sect. 1.1, we pointed out that the nonparametric sign and Wilcoxon location estimators are based on minimizers of  $L_1$ -norms. This is true also in the regression case for the Wilcoxon and sign scores. First, if sign scores are used then the rank-based estimator of  $\boldsymbol{\beta}$  and  $\alpha$ , as estimated by the median of the residuals, are the  $L_1$  (least absolute deviations) estimators of  $\alpha$  and  $\boldsymbol{\beta}$ ; see page 212 of Hettmansperger and McKean (2011). Secondly, for Wilcoxon scores we have the identity

$$\frac{4(n+1)}{\sqrt{12}} \sum_{i=1}^n \sqrt{12} \left( \frac{R(u_i)}{n+1} - \frac{1}{2} \right) u_i = \sum_{i=1}^n \sum_{j=1}^n |u_i - u_j|, \quad \mathbf{u} \in R^n. \quad (1.9)$$

That is, the Wilcoxon estimator of the regression coefficients minimizes the absolute deviations of the differences of the residuals.

In addition to the above defined rank-based estimator of  $\boldsymbol{\beta}$ , we can also construct an hypothesis test of the general linear hypotheses

$$H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{0} \text{ versus } H_A : \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}, \quad (1.10)$$

where  $\mathbf{M}$  is a  $q \times p$  matrix of full row rank. Let  $V_r$  denote the reduced model subspace; i.e., the subspace of  $V_f$  subject to the null hypothesis. Let  $W$  be a  $n \times (p-q)$  basis matrix of  $V_r$ . Then we write the reduced model as  $\mathbf{Y} = \alpha \mathbf{1} + \mathbf{W}\boldsymbol{\theta} + \mathbf{e}$ . Let  $\hat{\boldsymbol{\theta}}$  denote the rank-based estimator of this reduced model. Then the distance between  $\mathbf{Y}$  and the subspace  $V_r$  is  $D(\hat{\boldsymbol{\theta}})$ , which is the same for any basis matrix of  $V_r$ .

The test statistic of the hypotheses (1.10) is the normalized version of the reduction in distance,  $D(\hat{\boldsymbol{\theta}}) - D(\hat{\boldsymbol{\beta}})$ , given by:

$$F_\varphi = \frac{[D(\hat{\boldsymbol{\theta}}) - D(\hat{\boldsymbol{\beta}})]/q}{\hat{\tau}/2}, \quad (1.11)$$

where  $\hat{\tau}$  is an estimator of the scale parameter

$$\tau^{-1} = \int \varphi'(F(t))f^2(t) dt = \int f(t) d(\varphi(F(t))). \quad (1.12)$$

Koul et al. (1987) developed a consistent estimator of  $\tau$ . Note that the reduction in distance (dispersion) parallels the least squares reduction in sums of squares.

The approximating distributions of the estimator and the test statistic are determined by a linear approximation of the negative gradient of the dispersion and a quadratic approximation of the dispersion. Let  $\boldsymbol{\beta}_0$  denote the true parameter. Then the following approximations can be made asymptotically rigorous under mild regularity conditions:

$$\frac{1}{\sqrt{n}}\mathbf{S}(\boldsymbol{\beta}) \asymp \frac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varphi}(F(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)) - \tau^{-1}\frac{1}{n}\mathbf{X}'\mathbf{X}\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \quad (1.13)$$

$$D(\boldsymbol{\beta}) \asymp D(\boldsymbol{\beta}_0) - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'S(\boldsymbol{\beta}_0) + \frac{1}{2\tau}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\frac{1}{n}\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Based on these results the following asymptotic distributions can be obtained:

$\mathbf{S}(\boldsymbol{\beta}_0)$  is approximately  $MVN(\mathbf{0}, \mathbf{X}'\mathbf{X})$

$$\hat{\boldsymbol{\beta}} \text{ is approximately } MVN(\boldsymbol{\beta}_0, \tau^2(\mathbf{X}'\mathbf{X})^{-1}) \quad (1.14)$$

$$F_\varphi \text{ is approximately } F(q, n-p-1), \text{ under } H_0; \quad (1.15)$$

( $qF_\varphi \rightarrow \chi^2$ -distribution with  $q$  degrees of freedom, under  $H_0$ ).

Based on (1.14), an approximate  $(1 - \alpha)100\%$  confidence interval for the linear function  $\mathbf{h}'\boldsymbol{\beta}$  is

$$\mathbf{h}'\hat{\boldsymbol{\beta}} \pm t_{\alpha, n-p}\hat{\tau}\sqrt{\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}}, \quad (1.16)$$

where  $t_{\alpha, n-p}$  is the upper  $\alpha/2$  quantile of a  $t$ -distribution with  $n-p$  degrees of freedom. The use of  $t$ -critical values and  $F$ -critical values for tests and confidence procedures is supported by numerous small sample simulation studies; see McKean and Sheather (1991) for a review of such studies.

### 1.2.1 Diagnostics

After fitting a model, a residual analysis is performed to check for quality of fit and anomalies. A standard diagnostic tool for a LS fit is the scatterplot of residual versus fitted values. A random scatter indicates a good fit, while patterns in the plot indicate a poor fit and, often, lead to the subsequent fitting of more adequate models. For example, suppose the true model is of the form

$$\mathbf{Y} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \quad (1.17)$$

where  $\mathbf{Z}$  is an  $n \times q$  matrix of constants and  $\boldsymbol{\gamma} = \boldsymbol{\theta}/\sqrt{n}$ ,  $\boldsymbol{\theta} \neq \mathbf{0}$ . We fit, though, Model (1.1) using LS; i.e., the model has been misspecified. A straight forward calculation yields

$$\hat{\mathbf{Y}}_{LS} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{e} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma} \quad \text{and} \quad \hat{\mathbf{e}}_{LS} = \mathbf{e} - \mathbf{H}\mathbf{e} + (\mathbf{I} - \mathbf{H})\mathbf{Z}\boldsymbol{\gamma},$$

where  $\mathbf{H}$  is the projection matrix onto the range of  $\mathbf{X}$ . If  $\text{range}(\mathbf{Z}) \not\subset \text{range}(\mathbf{X})^\perp$  then both the fitted values and residuals are functions of  $\mathbf{Z}\boldsymbol{\gamma}$  and, hence, there will be information in the plot concerning the misspecified model. Note that the function  $\mathbf{H}\mathbf{e}$  is unbounded; so, based on this representation, outliers in the random errors are diffused throughout the residuals and fitted values. This leads to distortions in the residual plot which, for example, may even mask outliers.

For the rank-based fit of Model (1.1) when Model (1.17) is the true model, it follows from the above linearity results that

$$\hat{\mathbf{Y}}_{rb} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \tau_\varphi \mathbf{H}\boldsymbol{\varphi}[F(\mathbf{e})] + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma} \quad \text{and} \quad \hat{\mathbf{e}}_{rb} = \mathbf{e} - \tau_\varphi \mathbf{H}\boldsymbol{\varphi}[F(\mathbf{e})] + (\mathbf{I} - \mathbf{H})\mathbf{Z}\boldsymbol{\gamma}.$$

Thus, as with LS, there is information in the residual plot concerning misspecified models. Note from the rank-based representation, the function  $\mathbf{H}\boldsymbol{\varphi}[F(\mathbf{e})]$  is bounded. Hence, the rank-based residual plot is less sensitive to outliers. This is why outliers tend to stand out more in residual plots based on robust fits than in residual plots based on LS fits. Other diagnostic tools for rank-based fits are discussed in Chaps. 3 and 5 of Hettmansperger and McKean (2011). Among them are the Studentized residuals. Recall that the  $i$ th LS Studentized residual is  $\hat{e}_{LS,i}^* = \hat{e}_{LS,i}/[\hat{\sigma} \sqrt{1 - (1/n) - h_i}]$ , where  $h_i$  is the  $i$ th diagonal entry of the projection matrix  $\mathbf{H}$  and  $\hat{\sigma}$  is the square root of  $MSE$ . Note that  $\hat{e}_{LS,i}^*$  is corrected for both scale and location. Rank-based Studentized residuals are discussed in Sect. 3.9.2 of Hettmansperger and McKean (2011). As with LS Studentized residuals, they are corrected for both location and scale. The usual outlier benchmark for Studentized residuals is  $\pm 2$ , which we use in the examples below.

## 1.2.2 Computation

Kloke and McKean (2012, 2014) developed the R package `Rfit` for the rank-based fitting and analysis of linear models. This package along with its auxiliary package `npsm` can be downloaded from the site CRAN; see Sect. 1.1 for the url. The Wilcoxon score function is the default score function of `Rfit` but many other score functions are available in `Rfit` including the normal scores and the simple bent scores (Winsorized Wilcoxon). Furthermore, users can easily implement scores of their choice; see Chap. 3 of Kloke and McKean (2014) for discussion. In subsequent examples we demonstrate how easy `Rfit` is to use.

Analogous to least squares, the rank-based analysis can be used to conduct inference for general linear models, i.e., a robust ANOVA and ANCOVA; see Chaps. 3–5 of Hettmansperger and McKean (2011). As an illustration, we end this section with an example that demonstrates how easy the rank-based analysis can be used to test for interaction in a two-way design.

## 1.2.3 Example

Hollander and Wolfe (1999) provide an example on light involving a  $2 \times 5$  factorial design; see, also, Kloke and McKean (2014) for discussion. The two factors in the design are the light regimes at two levels (constant light and intermittent light) and five different dosage levels of luteinizing release factor (LRF). Sixty rats were put on test under these ten treatments combinations (six repetitions per combination). The measured response is the level of luteinizing hormone (LH), nanograms per ml of serum in the resulting blood samples.

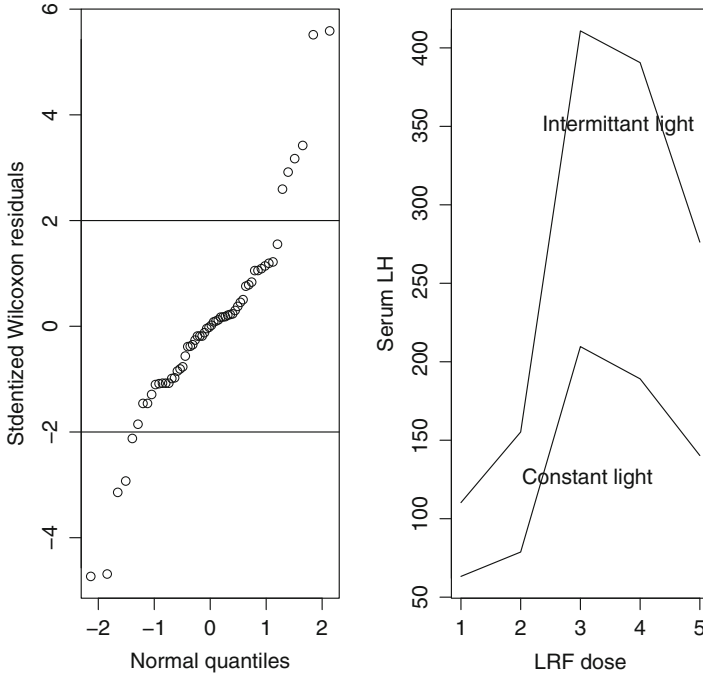
We chose Wilcoxon scores for our analysis. The full model is the usual two-way model with main and interaction effects. The right panel in Fig. 1.1 shows the mean profile plots based on the full model rank-based estimates. The profiles are not parallel indicating that interaction between the factors is present. These data comprise the `serumLH` data set in `Rfit` and hence is loaded with `Rfit`. The `Rfit` function `raov`, (robust ANOVA), obtains the rank-based analysis with one line of code as indicated below. The reduction in dispersion test, (1.11), of each effect is adjusted for all other effects analogous to Type III sums of squares in SAS; see Sect. 5.5 of Kloke and McKean (2014). Further, the design need not be balanced. Here is the code and resulting (with some abbreviation) rank-based ANOVA table:

```
> raov(serum~light.regime+LRF.dose+
      light.regime*LRF.dose, data=serumLH)
```

Robust ANOVA Table

	DF	RD	F	p-value
light.regime	1	1642.3332	58.03844	0.00000
LRF.dose	4	3027.6734	26.74875	0.00000
light.regime:LRF.dose	4	451.4559	3.98850	0.00694





**Fig. 1.1** Plots for the serum LH data

As the figure suggested, the factors light regime and LRF dose interact,  $p = 0.00694$ . In contrast, the LS analysis fails to reject interaction at the 5% level,  $p = 0.0729$ . For this example, at the 5% level of significance, the rank-based and LS analyses would lead to different interpretations. The left panel of Fig. 1.1 displays the  $q-q$  plot of the Wilcoxon Studentized residuals. This plot indicates that the errors are drawn from a heavy tailed distribution with numerous outliers, which impaired the LS analysis. The estimate of the ARE between the rank-based and least squares analyses is the ratio

$$\widehat{ARE} = \frac{\hat{\sigma}^2}{\hat{\tau}_\varphi^2} \tag{1.18}$$

where  $\hat{\sigma}^2$  is the MSE of the full model LS fit. This is often thought of as a measure of precision. For this example, this ratio is 1.88. So, the rank-based analysis cuts the LS precision by a factor of  $1/1.88 = 0.53$ .

In a two-way analysis when interaction is present often subsequent inference involves contrasts of interest. To demonstrate how easy this is accomplished using `Rfit`, suppose we consider the contrast between the expected response at the peak

(factor LRF dose at 3, factor light at the intermittent level) minus the expected response at the peak (factor LRF dose at 3, factor light at the constant level). Of course this is after looking at the data, so we are using this in a confirmatory mode. Our confidence interval is of the form (1.16). Using the following code, it computes to  $201.16 \pm 65.63$ . The difference is significant.

```
# full model fit
fitmod <- rfit(serum~factor(light.regime)+
  factor(LRF.dose)+ factor(light.regime)*
  factor(LRF.dose), data=serumLH)
# hvec picks the contrast
hvec <- rep(0,60); hvec[27] <- -1; hvec[57] <- 1
# estimate of contrast
contr <- hvec*%*fitmod$fitted.values
# error in CI
se2 <- t(hvec)*%*mat*%*vc*%*t(mat)*%*hvec
# error term in the confidence interval.
err <- qt(.975,50)*sqrt(se2)
```

### 1.3 Efficiency and Optimality

In general, the relative efficiency of one statistical method to another, in estimating or testing, is the squared ratio of the slopes in their respective linear approximations. Restricting ourselves to the linear model, (1.1), and rank-based procedures, in light of the asymptotic linearity results, (1.14), such ratios involve the scale parameter  $\tau$ . In particular, suppose we consider two rank-based methods using the respective score functions  $\varphi_1(u)$  and  $\varphi_2(u)$ , and, hence, the norms  $\|\cdot\|_{\varphi_1}$  and  $\|\cdot\|_{\varphi_2}$ . Then the asymptotic efficiency of method 1 to method 2 is

$$e(\|\cdot\|_{\varphi_1}, \|\cdot\|_{\varphi_2}) = \frac{\tau_2^2}{\tau_1^2}. \quad (1.19)$$

Note by (1.14) that this is the same as the ratio of the asymptotic variances of the associated estimators of the regression coefficients. Values greater than 1 indicate that methods based on  $\|\cdot\|_{\varphi_1}$  are superior. The larger slope indicates a more sensitive method, where the slope is  $\tau^{-1}$ .

For LS,  $e(\|\cdot\|_{\varphi_1}, \text{LS}) = \sigma^2/\tau_1^2$ , where  $\sigma^2$  is the variance of the random errors. For comparison with Wilcoxon methods, we already mentioned for location models that  $e(\text{Wilcoxon}, \text{LS})$  is 0.955 when the error distribution is normal. Hence,  $e(\text{Wilcoxon}, \text{LS})$  is the same for both linear and location models. Another striking result, due to Hodges and Lehmann (1956), shows that this efficiency is never less than 0.864 and may be arbitrarily large for heavy tailed distributions. In the case of

the normal scores methods, the efficiency relative to least squares is 1 at the normal model and never less than 1 at any other model!

An optimality goal is to select a score function to minimize  $\tau_\varphi$ ; i.e., maximize  $\tau_\varphi^{-1}$ . We can write expression (1.12) as

$$\tau^{-1} = \int_0^1 \varphi(u)\varphi_f(u) du, \quad (1.20)$$

where

$$\varphi_f(u) = \frac{f'[F^{-1}(u)]}{f[F^{-1}(u)]}. \quad (1.21)$$

Recall that the scores have been standardized so that  $\int \varphi^2(u) du = 1$ . Hence  $\tau^{-1}$  can be expressed as

$$\begin{aligned} \tau^{-1} &= \frac{\int_0^1 \varphi(u)\varphi_f(u) du}{\sqrt{\int_0^1 \varphi^2(u) du} \sqrt{\int_0^1 \varphi_f^2(u) du}} \left\{ \sqrt{\int_0^1 \varphi_f^2(u) du} \right\} \\ &= \rho \left\{ \sqrt{\int_0^1 \varphi_f^2(u) du} \right\}. \end{aligned} \quad (1.22)$$

The first factor on the right in the first line is a correlation coefficient which we have indicated by  $\rho$ . Thus  $\tau^{-1}$  is maximized if we select the score function to be  $\varphi_f(u)$ , (standardized form). This makes the correlation coefficient 1 and  $\tau_\varphi^{-1}$  equal to the term in the braces. This term, though, is the square-root of Fisher information. Therefore, by the Rao-Cramér lower bound, the choice of  $\varphi_f(u)$  as the score function leads to an asymptotically efficient (optimal) estimator. For the score functions discussed in earlier sections, it follows that the optimal score function for normally distributed errors is the normal score function, for logistically distributed errors is the Wilcoxon score function, and for Laplace distributed errors is the sign score function.

Of course this optimality only can be accomplished provided that the form of  $f$  is known. Evidently, the closer the chosen score is to  $\varphi_f$ , the more optimal the rank based analysis is. A Hogg-type adaptive scheme where the score function is selected based on initial (Wilcoxon) residuals has proven to be effective; see Sect. 7.6 of Kloke and McKean (2014). McKean and Kloke (2014) successfully modified this scheme for fitting a family of skewed normal distributions.

### 1.3.1 Monte Carlo Study

To illustrate the optimality discussed above, we conducted a small simulation study of a proportional hazards model. Consider a response variable  $T$  with a  $p \times 1$  vector of covariates  $\mathbf{x}$ . Assume  $T$  has a  $\Gamma(1, \zeta)$  distribution where  $\zeta = \exp\{\mathbf{x}'\boldsymbol{\beta}\}$  and  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of parameters. It follows that

$$\log T = \mathbf{x}'\boldsymbol{\beta} + \epsilon, \quad (1.23)$$

where  $\epsilon$  has an extreme-valued distribution; see Chaps. 2 and 3 of Hettmansperger and McKean (2011). The optimal scores for this model are the log-rank scores generated by  $\varphi(u) = -1 - \log(1 - u)$ .

We simulated this model for the following situation: sample size is  $n = 20$ ; the covariates are  $(1, x_i)$ , where  $x_i = i/21$ ; and  $\alpha = -2$  and  $\beta = 5$ . The methods involved are LS and the three rank-based methods: optimal scores (log-rank), Wilcoxon scores, and normal scores. These scores are intrinsic to `Rfit`. The simulation size is 10,000. To show how simple the coding is, here is the gist of the program's loop portion for a simulation:

```
mu <- exp(alpha + beta*x)
y <- rgamma(20, 1, 1/mu); ly <- log(y);
fitw <- rfit(ly~x) # Wil.
fitly <- rfit(ly~x, scores=logrankscores) # opt.
fitns <- rfit(ly~x, scores=nscores) # ns
fitls <- lm(ly~x) # LS
```

Table 1.1 presents the empirical relative efficiencies (ratios of mean square errors) for the parameter  $\beta$ . The efficiencies are relative to the optimal rank-based score procedure. The log-rank score procedure is most efficient followed by the normal scores procedure and then the Wilcoxon. Least squares (LS) performed the worst.

**Table 1.1** Empirical AREs of estimators of the slope parameter  $\beta$  relative to the rank-based estimator based on the optimal log-rank scores

Method			
Optimal	Wilcoxon	Normal scores	LS
1	1.215	1.175	1.380

The simulation size is 10,000

## 1.4 Influence and High Breakdown

### 1.4.1 Robustness Properties

In the 1960s and 1970s new tools to assess robustness properties of estimators were developed beginning with Huber (1964) and Hampel (1974). The breakdown value of a location estimator is the (limiting) proportion of the data that must be contaminated in order to carry the value of the estimator beyond any finite bound. In the one sample location model with score function  $\varphi^+(u)$ , the breakdown for the rank based estimator is  $\epsilon$  where

$$\int_0^{1-\epsilon} \varphi^+(u) du = \frac{1}{2} \int_0^1 \varphi^+(u) du.$$

A simple computation shows that the least squares estimate, the mean, has 0 breakdown point, worst possible. The median has breakdown 0.5, the best possible. The median of the pairwise averages (Wilcoxon score) has breakdown 0.293, while the normal scores estimate has breakdown 0.239.

Another robustness tool, the influence function, is a measure of how fast the estimator changes when an outlier is moved out beyond the edges of the sample. It is provided by the linear approximation of the negative gradient,  $S(\beta)$ , (1.7). Consider the linear model (1.1). For the rank-based estimator, using the score function  $\varphi(u)$ , the influence function is given by:

$$\Omega(\mathbf{x}, y) = \tau \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \varphi(F(y)) \mathbf{x},$$

where  $(\mathbf{x}, y)$  is the value at which we evaluate the influence. When the score function is bounded the influence is bounded in the  $\mathbf{y}$ -space. However, note that influence is unbounded in factor space. In the case of the location model, the influence functions for the median and the median of the pairwise averages are both bounded. Note also that least squares estimators have unbounded influence functions in both the  $\mathbf{y}$ - and the  $\mathbf{X}$ -spaces.

For most designed experiments and for designs with predictors which are well behaved, the rank-based estimators offer a robust and highly efficient alternative to LS for fitting and analyzing linear models. In the case of messy predictors, though, a robust alternative with bounded influence in factor space and positive breakdown is most useful. In fact a primary use of such fits is to highlight the difference between its fit and that of a highly efficient robust fit and, thus, alerting the user to possible anomalies in factor space. We next discuss a high breakdown rank-based (HBR) fit and its accompanying diagnostics which serve this purpose.

### 1.4.2 High-Breakdown and Bounded Influence Rank-Based Estimates

For the linear model, Chang et al. (1999) developed a rank-based estimator that has bounded influence and can achieve a 50 % breakdown point. It is a weighted version rank-based Wilcoxon fit. By the identity (1.9), the Wilcoxon estimator minimizes the least absolute deviations of the differences of the residuals. Let  $\{b_{ij}\}$  be a set of nonnegative weights. Consider estimators which minimize

$$\hat{\boldsymbol{\beta}} = \text{Argmin} \sum_{i < j} b_{ij} |(y_i - \mathbf{x}'_i \boldsymbol{\beta}) - (y_j - \mathbf{x}'_j \boldsymbol{\beta})|. \quad (1.24)$$

If the weights are all 1, then this is the Wilcoxon estimator.

Chang et al. (1999) proposed weights which are both functions of factor space and residual space. For factor space, it uses robust distances based on the high breakdown minimum covariance determinant (MCD) which is an ellipsoid in  $p$ -space that covers about half the data. For residual space, it uses the high breakdown least trim squares (LTS) fit for an initial fit. See Rousseeuw and Van Driessen (1999).

A brief description of the weights are given next. These are the weights defined for the R function `hbrfit` which are in the R package `npmsReg2` and are discussed in Sects. 7.2 and 7.3 of Kloke and McKean (2014). This package can be downloaded at the github site indicated at the end of Sect. 1.1.

Let  $\hat{\boldsymbol{e}}_0$  denote the residuals from the initial LTS fit. Let  $V$  denote the MCD with center  $\mathbf{v}_c$ . 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  $\sigma$  be estimated by the initial scaling estimate  $MAD = 1.483 \text{ med}_i |\hat{e}_i^{(0)} - \text{med}_j \{\hat{e}_j^{(0)}\}|$ . Letting  $Q_i = (\mathbf{x}_i - \mathbf{v}_c)' V^{-1} (\mathbf{x}_i - \mathbf{v}_c)$ , define

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

Consider the weights

$$\hat{b}_{ij} = \min \left\{ 1, \frac{c\hat{\sigma}}{|\hat{e}_i^{(0)}| |\hat{e}_j^{(0)}|} \min \left\{ 1, \frac{b}{Q_i} \right\} \min \left\{ 1, \frac{b}{Q_j} \right\} \right\}, \quad (1.25)$$

where  $b$  and  $c$  are tuning constants. We set  $b$  at the upper  $\chi_{.05}^2(p)$  quantile and  $c$  is set as

$$c = [\text{med}\{a_i\} + 3MAD\{a_i\}]^2,$$

where  $a_i = \hat{e}_i^{(0)} / (MAD \cdot Q_i)$ . From this point-of-view, it is clear that these weights downweight both outlying points in factor space and outlying responses. Note that

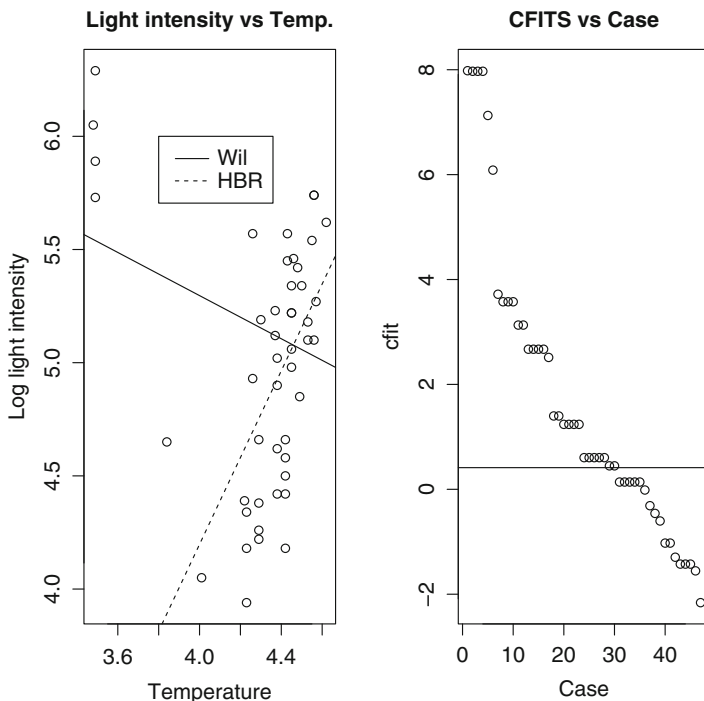
the initial residual information is a multiplicative factor in the weight function. Hence, a good leverage point will generally have a small (in absolute value) initial residual which will offset its distance in factor space.

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  $x$ -spaces, is continuous everywhere, and converges to zero as the point  $(x^*, Y^*)$  gets large in any direction. The asymptotic distribution of  $\hat{\beta}_{HBR}$  is asymptotically normal. As with all high breakdown estimates,  $\hat{\beta}_{HBR}$  is less efficient than the Wilcoxon estimates but it regains some of the efficiency if the weights depend only on factor space.

McKean et al. (1996) developed diagnostics to detect differences in highly efficient and high breakdown robust fits. Their diagnostic TDBETA measures the total difference in fits of the regression coefficients, standardized by the variance-covariance of the Wilcoxon fit. The benchmark is similar to the classic diagnostic DFFITS. A second diagnostic CFITS measures the difference of predicted values at each case. This diagnostic is useful for data sets where TDBETA exceeds its benchmark. Section 7.3 of Kloke and McKean (2014) give a full discussion of these diagnostics with examples. McKean et al. (1999) extended these diagnostics to differences between robust and LS fits. Next, we present an example which illustrates HBR fits and these diagnostics.

### 1.4.2.1 Stars Data

This data set is drawn from an astronomy study on the star cluster CYG OB1 which contains 47 stars; see Chap. 3 of Hettmansperger and McKean (2011) for discussion. The response is the logarithm of the light intensity of the star while the predictor is the logarithm of the temperature of the star. The scatterplot of the data is in the left panel of Fig. 1.2. Four of the stars, called giants, form a cluster of outliers in factor space while the rest of the stars fall in a point cloud. The panel includes the overlay plot of the Wilcoxon and HBR linear fits. The four giants form a cluster of high leverage points, exerting a strong influence on the Wilcoxon fit while having a minor influence on the HBR fit. The diagnostic TDBETAS between the Wilcoxon and HBR fits has the value 67.92 which exceeds the benchmark of 0.340, indicating a large difference in the fits. The right panel of Fig. 1.2 shows the values of the diagnostic CFITS versus case. The benchmark for this diagnostic is 0.34. The four largest values are the four giant stars. Hence, for this data set, the diagnostics work. The diagnostic TDBETAS alerts the user to the large difference between the fits and CFITS indicates the major points contributing to this difference. The next two largest CFITS values are of interest to astronomers, also. These are stars between the giants and the main sequence stars. Although not shown, the least squares fit is similar to the Wilcoxon fit. The fits and diagnostics can be computed with the following code:



**Fig. 1.2** The *left panel* displays the scatterplot of the log of light intensity versus the temperature of the star, overlaid with the Wilcoxon and HBR fits. The *right panel* displays the values of CFITS versus the case numbers. The *horizontal line* is the benchmark value

```
fitw <- rfit(lintensity ~ temp)
fith <- hbrfit(lintensity ~ temp)
fits <- lm(y ~ x)
fitsdiag <- fitdiag(temp, lintensity, est=c("WIL",
      "HBR"))
```

## 1.5 Extensions to Mixed and Nonlinear Models

In the past 20 years, there have been extensions of the rank-based analysis to many other models. This includes nonlinear models and models with dependency among the responses. In this section, we briefly discuss a few of these models, ending with an example involving a mixed model.

For traditional least squares-based methods for these models, the geometry essentially remains the same in that least squares estimation is based on minimizing the squared-Euclidean distance between the vector of responses and the region of estimation. This is true of the rank-based approach, also, except that the rank-based norm, (1.3), replaces the squared-Euclidean norm.



### 1.5.1 Multivariate Linear Models

Davis and McKean (1993) extended the linear model rank-based procedures for general score functions to the multivariate linear model. They developed asymptotic theory for the estimators and tests of linear hypotheses of the form  $\mathbf{A}\boldsymbol{\beta}\mathbf{K}$  for the matrix of regression coefficients  $\boldsymbol{\beta}$ . See Sect. 6.6 in Hettmansperger and McKean (2011) for discussion and examples. These are component wise estimators, so computations can be based on the package `Rfit`; see the web site indicated at the end of Sect. 1.1 to download a preversion of the package `Rfitmult`. These methods are regression equivariant but they are not affine invariant. Oja (2010) and his collaborators developed affine invariant rank-based procedures for Wilcoxon scores using a transformation retransformation procedure. Nordhausen and Oja (2011) developed the R package `MNM`, downloadable at CRAN, to compute these affine procedures.

### 1.5.2 Nonlinear Linear Models

For responses  $y_i$ , consider a nonlinear model of the form  $y_i = g(\boldsymbol{\theta}, \mathbf{x}_i) + e_i$ ,  $i = 1, \dots, n$ , where  $g$  is a specified nonlinear function,  $\boldsymbol{\theta}$  is a  $k \times 1$  vector of unknown parameters, and  $\mathbf{x}_i$  is a  $p \times 1$  vector of predictors. Let  $\mathbf{y}$  and  $g(\boldsymbol{\theta}, \mathbf{x})$  denote the corresponding  $n \times 1$  vectors. Given a rank score function  $\varphi(u)$ , the associated rank-based estimator of  $\boldsymbol{\theta}$  is

$$\hat{\boldsymbol{\theta}}_\varphi = \text{Argmin} \|\mathbf{y} - g(\boldsymbol{\theta}, \mathbf{x})\|_\varphi,$$

where  $\|\cdot\|_\varphi$  is the norm defined in expression (1.3). Abebe and McKean (2007) obtained asymptotic theory for  $\hat{\boldsymbol{\theta}}_\varphi$  for the case of Wilcoxon scores. The efficiency properties of the Wilcoxon estimator are the same as in the linear model case; so, the estimator is highly efficient for the nonlinear model. Abebe and McKean (2013) extended this development to high breakdown rank-based nonlinear estimators of  $\boldsymbol{\theta}$  which have bounded influence in both the response and factor spaces. The R package `npSmReg2` contains the R function `wiln1` which computes these estimators; see Chap. 7 of Kloke and McKean (2014) for further discussion.

### 1.5.3 Time Series Models

Consider the autoregressive model of order  $p$ ,  $Ar(p)$ :

$$\begin{aligned} X_i &= \phi_0 + \phi_1 X_{i-1} + \phi_2 X_{i-2} + \dots + \phi_p X_{i-p} + e_i \\ &= \phi_0 + \mathbf{Y}'_{i-1} \boldsymbol{\phi} + e_i, \quad i = 1, 2, \dots, n \end{aligned} \tag{1.26}$$

where  $p \geq 1$ ,  $\mathbf{Y}_{i-1} = (X_{i-1}, X_{i-2}, \dots, X_{i-p})'$ ,  $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_p)'$ , and  $\mathbf{Y}_0$  is an observable random vector independent of  $\mathbf{e}$ . Let  $\mathbf{X}$  and  $\mathbf{Y}$  denote the corresponding  $n \times 1$  vector and the  $n \times p$  matrix with components  $X_i$  and  $\mathbf{Y}'_{i-1}$ , respectively. For the score function  $\varphi(u)$ , the rank-based estimator of  $\boldsymbol{\phi}$  is given by

$$\hat{\boldsymbol{\phi}}_\varphi = \text{Argmin} \|\mathbf{X} - \mathbf{Y}'\boldsymbol{\phi}\|_\varphi,$$

where  $\mathbf{Y}$  is the matrix with rows  $\mathbf{Y}'_{i-1}$ . Koul and Saleh (1993) developed the asymptotic theory for these rank-based estimates. Because of the structure of the  $AR(p)$  model, outliers in the random errors become ensuing points of high leverage. As a solution to this problem, Terpstra et al. (2000, 2001) proposed estimating  $\boldsymbol{\phi}$  using the HBR estimators of Sect. 1.4. They obtained the corresponding asymptotic theory for these HBR estimators and showed their validity and empirical efficiency in several large simulation studies. Section 7.8 of Kloke and McKean (2014) discusses the computation of these estimates using the R package `Rfit`.

### 1.5.4 Cluster Correlated Data

Frequently in practice data are collected in clusters. Common examples include: repeated measures on subjects, experimental designs involving blocks, clinical studies over multiple centers, and hierarchical (nested) designs. Generally, the observations within a cluster are dependent.

For discussion, suppose we have  $m$  such clusters. Let  $y_{ki}$  denote the  $i$ th response within the  $k$ th cluster, for  $i = 1, \dots, n_k$  and  $k = 1, \dots, m$ , and let  $\mathbf{x}_{ki}$  denote the corresponding  $p \times 1$  vector of covariates. For cluster  $k$  stack the  $n_k$  responses in the vector  $\mathbf{y}_k$  and let  $\mathbf{X}_k$  denote the  $n_k \times p$  matrix with rows  $\mathbf{x}'_{ki}$ . Assume a linear model of the form

$$\mathbf{y}_k = \beta_0 \mathbf{1}_{n_k} + \mathbf{X}_k \boldsymbol{\beta} + \mathbf{e}_k, \quad k = 1, \dots, m, \quad (1.27)$$

where  $\mathbf{e}_k$  follows a  $n_k$ -multivariate distribution and the vectors  $\mathbf{e}_1, \dots, \mathbf{e}_m$  are independent. We then stack the vectors  $\mathbf{y}_k$  and matrices  $\mathbf{X}_k$  into the vector  $\mathbf{Y}$  and matrix  $\mathbf{X}$ , respectively.

There are several rank-based analyses available for these models. Abebe et al. (2016) develop a rank-based analysis for a general estimating equations (GEE) model which includes models of the form (1.27). This allows for very general dependency structure. For a general score function  $\varphi(u)$ , Kloke et al. (2009) developed the asymptotic theory for the rank-based estimator defined in expression (1.5), i.e., the minimizer of the norm  $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_\varphi$ . Their development includes consistent estimators of standard errors and consistent test statistics of general linear hypotheses. The theory requires the additional assumption that the univariate marginal distributions of  $\mathbf{e}_k$  are the same. This is true for many of the usual models in practice such as

simple mixed models (compound symmetry covariance structure) and stationary times series models for the clusters.

The R package `jrfit` computes the analysis developed by Kloke et al. (2009); see Chap. 8 of Kloke and McKean (2014) for discussion and examples. This includes the fit and several options for the estimation of the covariance structure, including compound symmetry and two general estimators, (a sandwich-type estimator and a general nonparametric estimator). We conclude this discussion with an example which illustrates the use of `jrfit` on cluster data.

### 1.5.4.1 Example

For an example, we consider the first base study presented in Hollander and Wolfe (1999). This study investigated three methods of rounding first base for baseball players who are running from home plate to second base. The response is the player's total running time. The three methods are narrow angle (NA), round out (RP), and wide angle (WA); see Hollander and Wolfe for details. Twenty-two ball players participated in the study and each ran six times, two repetitions for each method. The average time of the two repetitions are the response times available. The data can be found in the `firstbase` data set in the R package `npSmReg2`; see Chap. 8 of Kloke and McKean (2014).

Let  $y_{ij}$  denote the running time for the  $j$ th player on method  $i$  and consider the randomized block design given by

$$y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}, \quad (1.28)$$

where  $\alpha_i$  denotes the  $i$ th treatment fixed effect;  $b_j$  denotes the random effect for the  $j$ th player; and  $\epsilon_{ij}$  denotes the random error. We assume that the random errors are iid and the random effects are iid with different distributions. We further assume that the random errors and the random effects are independent.

Although, finite variance is not required for the asymptotic theory, for the discussion, we assume finite variances. The variance-covariance structure of Model (1.28) is compound symmetric. Besides fixed effects analyses, we are interested in the estimation of the variance components given by  $\sigma_b^2$  the variance of  $b_j$ ,  $\sigma_\epsilon^2$  the variance of  $\epsilon_{ij}$ , and the intraclass correlation coefficient  $\rho = \sigma_b^2 / (\sigma_b^2 + \sigma_\epsilon^2)$ . Kloke et al. (2009) provided robust estimates of these components based on the rank-based fit of Model (1.28). These estimates have been incorporated into the package `jrfit`.

The null hypothesis for the fixed effects is  $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ . The traditional nonparametric test of this hypothesis is based on Friedman's test statistic, which for this example results in the value of 11.14 with  $p$ -value 0.003. The comparative rank-based analysis is the Wald-type test on the rank-based fit. As shown below, the value of the test statistic is 19.31 with a  $p$ -value of 0.0001. As with the Friedman test, the Wald-type test is highly significant. An experimenter, though, wants a much more in depth analysis than just this test of the fixed effects. The left

**Table 1.2** Rank-based estimates of fixed effects and variance components, firstbase data

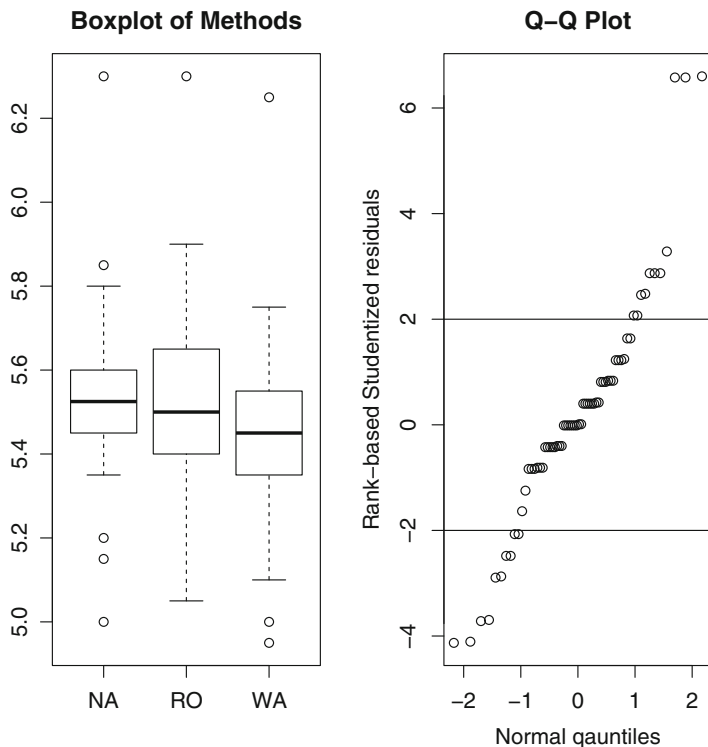
	Fixed effects		Variance components
	Effect	SE	
$\mu_{RO} - \mu_{NA}$	0.000	0.016	$\hat{\sigma}_B^2 = 0.0124$
$\mu_{WA} - \mu_{NA}$	-0.053	0.012	$\hat{\sigma}_\epsilon^2 = 0.0049$
$\mu_{WA} - \mu_{RO}$	-0.057	0.023	$\hat{\rho} = 0.715$

panel of Fig. 1.3 displays the comparative boxplots of the methods. Note that, based on this plot, it appears that the wide-angle method results in the quickest times. Such a judgement is easily confirmed by considering the estimates and confidence intervals for the three pairwise comparisons. These are shown in Table 1.2 based on the rank-based fit. They do confirm that the wide-angle method results in significantly faster times than the other two methods. Furthermore, the estimated effects provide the experimenter with an estimate of how much faster the wide-angle method is than the other two methods.

Table 1.2 also displays the robust estimates of the variance components. Note that the estimate of the intraclass correlation coefficient is 0.715 indicating a strong correlation among the times of a runner. The right panel of Fig. 1.3 shows the normal  $q-q$  plot of the Studentized residuals of the rank-based fit. The horizontal lines at  $\pm 2$  are the usual benchmark for potential outliers. This plot confirms the outliers in the boxplots and indicates a heavy tailed error structure. The three largest positive outliers correspond to Runner 22 who had the slowest times in all three methods.

The results in Table 1.2 and Fig. 1.3 are based on computations using the package `jrfit`. Some of the code for the computations is given by:

```
# The data are in the data set firstbase in the
# package npsmReg2. More discussion of the
# computations can be found in Chapter 8
# of Kloke McKean (2014).
#
# times is the vector of running times; player is
# the indicator of the player; method is the
# indicator of the method.
xmat <- model.matrix(~as.factor(method))[,2:3]
fit <- jrfit(xmat,times,player)
stud <- rstudent(fit)          #Studentized residuals
vee(fit$resid,fit$block,method='mm')    #Var comp.
h1 <-c(0,1,0); h2<-c(0,0,1); hmat<-rbind(h1,h2)
mid <- solve(hmat%*%fit$varhat%*%t(hmat))
tst <- t(hmat%*%fit$coef)%*%mid%*%hmat%*%fit$coef
19.31442
1-pchisq(tst,2)
6.396273e-05
```



**Fig. 1.3** The *left panel* displays the comparative boxplots for the three methods of rounding first base, while the *right panel* shows the *q-q* plot of the rank-based Studentized residuals

## 1.6 Conclusion

As we indicated in Sect. 1.1, the rank tests for simple location problems were initially used because of their quick calculation in the pre computer age. These methods were further found to be highly efficient and robust by Hodges and Lehmann in the mid 1950s. The traditional *t*-tests for these problems, though, are based on least squares (LS) fitting which easily generalizes to much more complicated models, including linear, nonlinear, and models with dependent error structure. For all such models, the LS fitting is based on minimizing the squared-Euclidean distance between the vector (or matrix) of responses and the region (space) of estimation. Further, LS testing of general linear hypotheses is based on a comparison of distances between the vector of responses and full and reduced model subspaces. Also, there are ample diagnostic procedures to check the quality of the LS fit of a model. As in the location problems, though, LS procedures are not robust. Hence, a generalization was needed for the robust nonparametric methods.

As briefly outlined in Sect. 1.2, the extension of nonparametric methods to linear models came about in the late 1960s and early 1970s, with the robust estimation procedures developed by Jurečková and Jaeckel. In particular, as shown by McKean and Schrader (1980), Jaeckel's estimation involves minimizing a distance between the responses and the full model subspace. This distance is based on the norm defined in expression (1.3). For general linear hypotheses, McKean and Hettmansperger (1976) developed an accompanying analysis based on a comparison of distances between the vector of responses and full and reduced model spaces where distance is based on the norm (1.3). Diagnostics for this rank-based analysis were developed by McKean et al. (1990). This rank-based analysis is as general as the traditional LS analysis. As with LS, for any linear model, it offers a complete procedure including fitting, diagnostic checking of the fit, confidence regions, and tests of general linear hypotheses. Details of this analysis are discussed in Chaps. 3–5 of Hettmansperger and McKean (2011).

As discussed in Sect. 1.3, the rank-based analysis is highly efficient. For example, the rank-based procedure based on Wilcoxon scores has efficiency 0.955 relative to LS procedures when the random errors are normally distributed and is much more efficient when the distribution of the random errors has heavy tails. Further, if the form of the error distribution is known, then optimal scores can be used which results in fully asymptotically efficient procedures. Rank-based procedures based on minimizing the norm (1.3) are robust in response space but, similar to LS procedures, are not robust in factor space. A simple weighting scheme, based on robust distances in factor space and residuals from an initial robust fit, leads to a robust rank-based procedure which is robust in both response and factor space as well as having a 50 % breakdown point.

As reviewed in Sect. 1.5, these rank-based procedures have been extended to nonlinear models and models in which the errors have dependencies. For these models, LS fitting is still based on minimizing squared-Euclidean distance between the responses and the space of estimation. In the same way, the rank-based fitting of these models is obtained by minimizing the distance based on the norm (1.3). In recent years, asymptotic theory has been developed for these rank-based procedures. Hence, besides linear models, robust rank-based procedures exist for diverse models, including nonlinear models, autoregressive times series models, multivariate regression models, mixed models, and hierarchical models.

The easy computation of rank-based analyses is performed with R software. For linear models, the package `Rfit` offers a complete computation for the rank-based analysis. A wide variety of scores functions are intrinsic to the package with an option for user-supplied scores. For models other than linear there are accompanying R packages for computations. We have discussed the computation based on these packages throughout this paper. See Kloke and McKean (2014) for discussion of these packages.

## References

- Abebe, A., & McKean, J. W. (2007). Highly efficient nonlinear regression based on the Wilcoxon norm. In D. Umbach (Ed.), *Festschrift in Honor of Mir Masoom Ali on the Occasion of his Retirement* (pp. 340–357).
- Abebe, A., & McKean, J. W. (2013). Weighted Wilcoxon estimators in nonlinear regression. *Australian & New Zealand Journal of Statistics*, *55*, 401–420.
- Abebe, A., McKean, J. W., Kloke, J. D., & Bilgic, Y. (2016). Iterated reweighted rank-based estimates for GEE models. In R. Y. Liu & J. W. McKean (Eds.), *Robust rank-based and nonparametric methods*. New York: Springer
- Arbuthnott, J. (1710). An argument for divine providence taken from the constant regularity observed in the birth of both sexes. *Philosophical Transactions*, *27*, 186–190.
- Chang, W., McKean, J. W., Naranjo, J. D., & Sheather, S. J. (1999). High breakdown rank-based regression. *Journal of the American Statistical Association*, *94*, 205–219.
- Davis, J., & McKean, J. W. (1993). Rank based methods for multivariate linear Models. *The Journal of the American Statistical Association*, *88*, 241–251
- Hájek, J., & Šidák, Z. (1967). *Theory of rank tests*. New York: Academic.
- Hájek, J., Šidák, Z., & Sen, P. K. (1999). *Theory of rank tests* (2nd ed.). New York: Academic.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383–393.
- Hettmansperger, T. P., & McKean, J. W. (2011). *Robust nonparametric statistical methods* (2nd ed.). Boca Raton, FL: Chapman-Hall.
- Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, *27*, 324–335.
- Hodges, J. L., Jr., & Lehmann, E. L. (1960). Comparison of the normal scores and Wilcoxon tests. In *Proceedings 4th Berkeley Symposium* (Vol. 1, pp. 307–317).
- Hodges, J. L., Jr., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics*, *34*, 598–611.
- Hollander, M., & Wolfe, D. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.
- Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics*, *43*, 1449–1458.
- Jurečková, J. (1969). Asymptotic linearity of rank statistics in regression parameters. *Annals of Mathematical Statistics*, *40*, 1449–1458.
- Jurečková, J. (1971). Nonparametric estimate of regression coefficients. *Annals of Mathematical Statistics*, *42*, 1328–1338.
- Kloke, J. D., & McKean, J. W. (2012). Rfit: Rank-based estimation for linear models. *The R Journal*, *4*, 57–64.
- Kloke, J. D., & McKean, J. W. (2014). *Nonparametric statistical methods using R*. Boca Raton, FL: Chapman-Hall.
- Kloke, J., McKean, J. W., & Rashid, M. (2009). Rank-based estimation and associated inferences for linear models with cluster correlated errors. *Journal of the American Statistical Association*, *104*, 384–390.
- Koul, H. L., & Saleh, A. K. M. E. (1993). R-estimation of the parameters of autoregressive [AR(p)] models. *The Annals of Statistics*, *21*, 534–551.
- Koul, H. L., Sievers, G. L., & McKean, J. W. (1987). An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scandinavian Journal of Statistics*, *14*, 131–141.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one or two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50–60.
- McKean, J. W., Jr. (1975). *Tests of Hypotheses Based on Ranks in the General Linear Model*. Ph.D. dissertation, University Park, Penn State University.

- McKean, J. W., & Hettmansperger, T. P. (1976). Tests of hypotheses of the general linear model based on ranks. *Communications in Statistics, Part A-Theory and Methods*, 5, 693–709.
- McKean, J. W., & Kloke, J. D. (2014). Efficient and adaptive rank-based fits for linear models with skewed-normal errors. *Journal of Statistical Distributions and Applications*, 1, 18. <http://www.jsdajournal.com/content/1/1/18>.
- McKean, J. W., Naranjo, J. D., & Sheather, S. J. (1996). Diagnostics to detect differences in robust fits of linear models. *Computational Statistics*, 11, 223–243.
- McKean, J. W., Naranjo, J. D., & Sheather, S. J. (1999). Diagnostics for comparing robust and least squares fits. *Journal of Nonparametric Statistics*, 11, 161–188.
- McKean, J. W., & Schrader, R. (1980). The geometry of robust procedures in linear models. *Journal of the Royal Statistical Society, Series B, Methodological*, 42, 366–371.
- McKean, J. W., & Sheather, S. J. (1991). Small sample properties of robust analyses of linear models based on r-estimates. In W. Stahel & S. Weisberg (Eds.), *Directions in robust statistics and diagnostics, part II* (Vols. 1–20). New York: Springer.
- McKean, J. W., Sheather, S. J., & Hettmansperger, T. P. (1990). Regression diagnostics for rank-based methods. *Journal of the American Statistical Association*, 85, 1018–1028.
- Noether, G. E. (1955). On a theorem of Pitman. *Annals of Mathematical Statistics*, 26, 64–68.
- Nordhausen, K., & Oja, H. (2011). Multivariate L1 methods: The package MNM. *Journal of Statistical Software*, 43(5), 1–28. <http://www.jstatsoft.org/v43/i05/>.
- Oja, H. (2010). *Multivariate nonparametric methods with R*. New York: Springer.
- Pitman, E. J. G. (1948). *Notes on nonparametric statistical inference* (Unpublished notes).
- Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.
- Terpstra, J., McKean, J. W., & Naranjo, J. D. (2000). Highly efficient weighed Wilcoxon estimates for autoregression. *Statistics*, 35, 45–80.
- Terpstra, J., McKean, J. W., & Naranjo, J. D. (2001). GR-estimates for an autoregressive time series. *Statistics and Probability Letters*, 51, 172–180.
- Tukey, J. W. (1949). *The simplest signed rank tests*. Princeton University Stat. Res. Group, Memo Report no. 17.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.