RANK-BASED ROBUST ANALYSIS OF LINEAR MODELS.

I. EXPOSITION AND BACKGROUND

BY

DAVID C. DRAPER*

TECHNICAL REPORT NO. 17
FEBRUARY 1983

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY

*RESEARCH PARTIALLY SUPPORTED BY
NATIONAL SCIENCE FOUNDATION GRANTS
MCS77-01665 AND MCS79-03716

RANK-BASED ROBUST ANALYSIS OF LINEAR MODELS. I. EXPOSITION AND BACKGROUND[†]

by

David C. Draper[‡]

Department of Statistics
University of California, Berkeley

ABSTRACT

Linear models are widely used in many branches of scientific inquiry. The classical analysis of linear models, however, is based on a number of technical assumptions whose failure to apply to the data at hand can result in poor performance of the classical techniques. One method of dealing with this which has gained some acceptance is the data-analytic approach, in which graphical and numerical methods are employed to detect the ways in which the data do not meet the classical assumptions and the data are modified appropriately before the classical techniques are applied. Another approach involves the use of robust methods which are appropriate under broader assumptions and so may be utilized directly with the original data. The application of one type of robust methods, those based on ranks, to problems of estimation and testing in the general linear model is reviewed here.

AMS Subject Classification: Primary 62G10, 62J10; Secondary 62F35, 62G05

Key words and phrases: robust estimation, rank-based, linear model, Hodges-Lehmann, kernel-type density estimation.

[†]Prepared with partial support of National Science Foundation Grants MCS77-01665 and MCS79-03716.

[‡]Now with the Department of Statistics, University of Chicago.

1. Introduction

The general linear model is one of the most widely used tools devised by statisticians to help guide scientific inquiry. Applications of linear regression, analysis of variance (ANOVA), and analysis of covariance techniques abound in the physical, biological, social, and behavioral sciences, among them economics, physics, psychology, sociology, medicine, and biology. It is a basic truth, however, in mathematical modeling that great depth of inference is usually arrived at only through greatly restrictive assumptions, and linear models provide no counterexample to this statement. It is worthwhile to consider these assumptions and to take up the question of what to do in practice when some or all of them are not reasonable for the data at hand.

The general fixed-effects linear model can be written in the form

$$Y_i = g(x_{i1},\ldots,x_{ip}) + e_i , \qquad i = 1,\ldots,N . \tag{1.1}$$

Here $(Y_i ; x_{i1},\ldots,x_{ip})$ is the ith of N total observations on the quantitative dependent variable Y and the p quantitative or qualitative (nominal or ordinal) independent variables $x_1,\ldots,x_p$, which are considered to be under experimenter control and without random error; the e_i are thought of as stochastic errors or disturbance terms. The Y_i and e_i are taken to be random variables and the x_{ij} to be fixed known constants. The function g is assumed to be known and to be of the form

$$g(x_{i1},\ldots,x_{ip}) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j , \tag{1.2}$$

in which the β_j are unknown parameters. A number of technical assumptions are made in the classical analysis about the errors e_i, which can be listed roughly in order of increasing severity as follows: the errors are assumed to have expectation 0, to be independent, to have the same variance for all i, and to be identically distributed with density f; their density f is assumed to be symmetric; and, finally, f is assumed to be a specific symmetric density, the normal. Denote by Ω^* this full model with all of these assumptions.

Over the past fifty or so years in which the model has evolved in this form (see Seal (1967) and Scheffé (1959) for some of the history), three basic approaches have arisen for dealing with the issue of violation of these technical assumptions:

(1) The data-analytic approach, in which graphical and numerical tools like residual plots (cf. Draper and Smith (1981)), and quantitative methods for the identification of influential observations (Weisberg (1980)), enable one to detect and characterize the various ways in which the data at hand do not fit the linear model assumptions. The data are then transformed in some way, through deletion of outliers and other influential observations and/or functional transformations on the observed x and Y values, and the classical analysis is applied to the transformed data;

(2) The robust approach, in which non-classical techniques are applied which are sufficiently insensitive to deviations from the assumptions that the data may be analyzed without modification; and

(3) The do-nothing approach, in which the issue of possible violation of assumptions is never even raised and the classical analysis is applied without question to the data.

That the acceptance of method (3) is widespread even today can be seen by observing, for example, that the ANOVA program in SPSS (Nie et al (1975)), one of the most widely used statistical computing packages, does not permit the user to examine in any way the residuals from the model fit. It is my experience that many users of linear models simply are not aware of any need to consider the assumptions built into Ω^*. Among practitioners of method (3) there also seems to be a class of users who are acting in the hope that the well-known optimality properties of the classical methods under Ω^* continue to hold when some of its assumptions do not, and in the belief that the classical analysis is robust against significant departures from these assumptions. These hopes and beliefs persist in spite of evidence to the contrary; it has been amply shown (cf. Bradley (1978), Scheffé (1959)) that there are many deviations from the basic assumptions against which the classical analysis is not robust, and much work has been done (Lehmann (1963b), Huber (1981), and Bickel (1973), among many others) to show that techniques exist which perform noticeably better than the classical methods when one departs from some of those basic assumptions.

Among statisticians who see the need for methods to deal with violations of the assumptions in Ω^*, use of data-analytic techniques has been far more prevalent to this date than use of robust methods. One of the reasons most often given for this (see Hettmansperger and McKean (1978)) is that robust methods so far have generally failed to satisfy the following criteria of ready usability by the final consumers of the linear model analysis: Whichever techniques are used

a) should have clear intuitive appeal;

b) should be of a unified nature and of general applicability rather than being put together in patchwork form out of solutions of separate but related problems; and

c) should possess simple, closed-form expressions, where possible.

It can be seen in recent work (Hettmansperger and McKean (1977), Draper (1981)) and will be seen in this paper that for some robust techniques these criticisms are no longer valid.

Progress in robust methods in linear models has been based on generalizing existing robust techniques for the one- and two-sample problems, and on starting with the ideas which lead to the classical methods and replacing them at key points with devices which make the resulting techniques more robust. This work so far has concentrated primarily on the distributional assumptions in Ω^*, not because violations of these assumptions are the most critical (they are not -- departures from the homogeneity of variance and independence assumptions have more serious consequences (see e.g. Scheffé (1959))), but probably because this is the most analytically tractable area in which to begin. Significant effort remains in the development of methods which are robust with respect to violation of the non-distributional assumptions in Ω^*. The considerable progress which <u>has</u> been made in the distributionally robust methods has proceeded roughly along three parallel lines: the maximum-likelihood-type or M-methods (Huber (1973)); the rank-based or R-methods (Lehmann (1963b), Jaeckel (1972), Hettmansperger and McKean (1977)); and methods based on linear combinations of order statistics, the L-methods (Bickel (1973)). Each of these approaches has its advantages and disadvantages in terms of robustness, efficiency, and usability. None of them clearly

dominates the others in efficiency and robustness (Huber (1981)), but with respect to practical considerations in implementation, like the ones above, there are some differences. The L-methods appear to be the most awkward of the three in generalizing to linear models (Huber (1981)), and, while the M- and R-methods are both quite unified in their approach and intuitively appealing, the R-methods (particularly in estimation) often have simple, closed-form expressions while the M-methods do not. Attention is restricted in this paper to rank-based methods. A survey of some existing rank-based techniques which are robust with respect to departures from normality is presented, and techniques are discussed which in many cases in addition permit relaxation of the symmetry assumption. Though the emphasis here is on fixed-effects models, several of the methods described would be expected to work well in certain random-effects and mixed models also; this is discussed briefly in Section 2.

Two rank-based approaches to linear model analysis are described below in Sections 2 and 3: the analysis of variance techniques of Lehmann (1963b) and general linear model methods of Hettmansperger and McKean (1976, 1977). In both approaches the analogy with classical methods is quite strong; one obtains robust estimates of functions of the parameters β_i and standard errors for those estimates, and a robust version of the analysis of variance table, complete with rank analogues of sums of squares, degrees of freedom, and F-ratios (Lehmann (1963b), Schrader and McKean (1977), Draper (1981)). Both approaches use the robust estimation techniques to construct confidence regions and significance tests and carry out multiple comparisons, Lehmann using Hodges-Lehmann two-sample estimates (Lehmann (1963a)) and Hettmansperger

and McKean using the rank-based regression estimates of Jaeckel (1972). Each method has its drawbacks -- the Hettmansperger-McKean approach as developed to date requires the assumption of symmetry of the underlying error distribution, and the Lehmann method only applies to ANOVA situations with several observations per cell. Moreover, both approaches to testing yield tests which are distribution-free only asymptotically, so that a check on their small-sample behavior is needed. This paper describes results pertaining to two issues -- the implementation of the Lehmann and Hettmansperger-McKean hypothesis-testing techniques in a way which dispenses when possible with the assumption of symmetry, and the carrying out of simulation studies to investigate the small-sample properties of these techniques empirically. Multiple comparisons methods and confidence procedures are not considered here; for examples of the rank-based robust approach to these inferential tools see e.g. Hettmansperger and McKean (1978).

It is worth emphasizing that the Lehmann and Jaeckel-Hettmansperger-McKean methods represent an improvement over the previous Kruskal-Wallis- and Friedman-type methods of rank-based linear model analysis in two ways: (1) They address issue (b) above directly in that they present a unified approach to the analysis rather than offering a collection of separate though related procedures; and (2) The estimates and quantities based on estimates from which the test statistics are constructed are on the same cardinal measurement scale as the original data, i.e., these are not methods which involve simply transforming the data to the rank scale. The basic distinction is that the Lehmann and Jaeckel-Hettmansperger-McKean methods are not rank methods -- they are robust procedures whose motivation <u>stems</u> from rank methods.

One advantage they thus possess over the data-analytic methods described above in situations where the data are gathered on a meaningful scale on which it is desired to carry out pairwise comparisons, etc., is that they permit such analyses to occur efficiently on the original scale, whereas the data-analytic methods may only allow such comparisons to be performed efficiently on a transformed scale in such a way that it is difficult or impossible to reinterpret the conclusions back on the original scale. See Draper (1982d) for more on this issue.

The asymptotic distributions of the estimates produced by both of the rank-based estimation techniques examined in Sections 2 and 3 depend on the error density f through $1/\theta$, where

$$\theta = \int_{-\infty}^{\infty} f^2(x)dx \quad , \tag{1.3}$$

so it is necessary to estimate $1/\theta$ to obtain tests and confidence regions. This estimation problem is studied in Draper (1982a); the principal results are outlined in Section 4 below. Several different methods of estimating $1/\theta$ in the context of the linear model were investigated: the techniques suggested originally by Lehmann (1963c) based on the lengths of confidence intervals derived from the Wilcoxon one-and two-sample statistics, and a method due to Schweder (1975) based on window estimation of f. All of these approaches except the method based on the Wilcoxon two-sample confidence intervals depend to at least some extent on symmetry of f about 0. The asymptotic behavior of the methods was worked out and compared; it turns out that asymptotically all methods studied perform equally well. The most interesting consequence of this result is that the method not requiring the

assumption of symmetry is seen to perform as well as those which do need symmetry, so when applicable it is to be preferred. This method, when joined with the Hodges-Lehmann two-sample or Jaeckel estimates, forms the basis of an approach to the analysis of linear models with several observations per cell which assumes of the errors essentially only that they are i.i.d.

A simulation study (Draper (1982b)) was performed to investigate empirically the small-sample properties of the Lehmann and Hettmansperger-McKean testing methods and the associated estimates of $1/\theta$; the major conclusions of this study are given below in Section 4. Bias of the $1/\theta$ estimates, and significance level and power of the tests using both the asymptotic distributions and small-sample improvements on them, were measured in a number of balanced and unbalanced ANOVA models with several observations per cell. The success of the tests in achieving approximately the desired level was seen to depend mostly on the bias of the estimator of $1/\theta$ used and the choice of approximation to the small-sample distributions of the test statistics. The Schweder-type methods studied for estimating $1/\theta$ were shown to be subject to large bias and hence to be unacceptable without modifications which adapt the estimator to the underlying density. Such modifications have not yet been investigated.

The Lehmann-type estimates of $1/\theta$ are also in need of bias corrections, but in their case simpler corrections are available, and several methods were introduced and explored in Draper (1982b). When corrected the Lehmann-type estimates of $1/\theta$ combine with both the Lehmann and Hettmansperger-McKean testing methods to produce tests which under most circumstances have approximately correct level and good to excellent power characteristics relative to the classical methods. By studying the empirical distributions of the test

statistics, deviations of the observed level of the tests from the nominal were seen to be due to nonoptimal choice of the approximations to the null distributions. For best efficiency and type I error performance of the tests it was suggested that the approximations to the null distributions should be chosen adaptively, and several methods of adaptation were proposed; see Draper (1982b) for details. Even without this refinement, however, the tests performed well in most situations studied.

The empirical small-sample efficiencies of the tests relative to the classical methods were seen to agree well with asymptotic predictions and to be quite stable even in the smaller layouts studied. As an alternative to the construction of power tables for the rank-based techniques, this makes feasible methods based on estimated efficiencies for determining sample sizes needed by the rank-based methods to achieve a desired power against an alternative of interest. This issue is also explored in Draper (1982b).

In summary, then, this recent work, together with results of Hettmansperger and McKean (1977) and others, provides rank-based methods which supply the linear models user with a comprehensive analysis package, from estimation and significance testing to confidence and multiple comparisons procedures, which have excellent efficiency and robustness properties relative to the classical methods; and which are sufficiently analogous to the classical techniques that users should have little trouble adapting to them. It is my hope that these and other robust methods of comparable quality will gain increasing acceptance in the near future.

2. The Methods of Lehmann: Analysis of Variance, Several Observations per Cell

The model which Lehmann (1963a) considers for ANOVA with several observations per cell can be written

$$Y_{ij} = \mu_i + e_{ij}, \qquad \left\{ \begin{array}{l} i = 1,\ldots,I \\ j = 1,\ldots,n_i \end{array} \quad \sum_{i=1}^{I} n_i = N \right\}, \tag{2.1}$$

in which the e_{ij} are independent, identically distributed (i.i.d.) continuous random variables with density f satisfying

$$\theta \equiv \int_{-\infty}^{\infty} f^2(x)dx < \infty \tag{2.2}$$

and $\sigma^2 = \mathrm{Var}(e_{ij}) < \infty$. Here μ_i is a measure of centering for the ith of I total cells, Y_{ij} is the jth of the n_i observations in cell i, and N is the total number of observations. For the μ_i to be identifiable an assumption is needed on the manner in which the distribution of the errors e_{ij} is centered at 0; for example, if $E(e_{ij}) = 0$ is assumed then μ_i is the mean of the distribution of the observations in cell i. In much of this section the cell centers μ_i are not as relevant as the differences $\mu_i - \mu_j$ between cell centers, and identifiability of the μ_i is not necessary. In such cases the identical distribution of the e_{ij} is enough for $\mu_i - \mu_j$ to be identifiable as the size of the shift in a two-sample shift model using the observations in cells i and j. In what follows, when an assumption is needed for the identifiability of μ_i the condition $E(e_{ij}) = 0$ will

be understood, in which case μ_i is the ith cell mean; in other cases where the choice of centering is immaterial μ_i will be called the ith cell center.

The above continuity assumption on the e_{ij} and consequently on the Y_{ij} is made to avoid technical complications involving ties in the ranking of the data. When ties are present in linear models data they are often due to the measuring process having made a conceptually continuous variable discrete, and in such situations, provided the size of the roundoff is not large, the methods below may be applied with little harm in acting as if the rounding had not occurred (cf. Lehmann (1975)). The finiteness of θ and σ^2 are needed because division by $1/\theta$ and σ^2 play a role in what follows; these conditions place little practical restriction on the use of the methods. Note that the notation of the model (2.1) is most natural only for the one-way layout, but larger layouts can be accommodated simply by numbering the cells from 1 to I.

Hodges and Lehmann (1963) had shown earlier how to use rank tests like the one- and two-sample Wilcoxon procedures to construct robust estimates of the center of symmetry of a distribution and the size of the shift in a two-sample shift model, and Lehmann was looking for a way to apply these methods to linear models. For widest applicability of the results it was preferable to adapt the two-sample Hodges-Lehmann estimates, because there is no assumption of symmetry implicit in their derivation (as there is in that of the one-sample estimates), so this suggested trying to estimate $\mu_i - \mu_j$ for $i \neq j$ in a robust fashion. Lehmann reasoned that this would be sufficient as a basis for robust versions of many of the most useful classical techniques, since most inference in ANOVA (linear hypothesis testing, multiple comparisons, etc.)

is based on contrasts in the cells means, and any contrast

$$\phi = \sum_{i=1}^{I} c_i \mu_i \ , \qquad \sum_{i=1}^{I} c_i = 0 \ , \tag{2.3}$$

is expressible in terms of differences in the cell means:

$$\sum_{i=1}^{I} c_i \mu_i = \sum_{i=1}^{I} \sum_{j \neq i} b_{ij} (\mu_i - \mu_j) \ . \tag{2.4}$$

Note that the b_{ij} are not unique. It is not possible with this approach to obtain estimates of the cell means themselves or of the grand mean

$$\bar{\mu} = N^{-1} \sum_{i=1}^{I} n_i \mu_i \ , \tag{2.5}$$

which essentially corresponds to the intercept term β_0 in the model Ω^* of the previous section. In effect, Lehmann was treating the ANOVA setup as an I-sample shift model and obtaining estimates by working separately with the $\binom{I}{2}$ 2-sample shift models embedded within it. (With a different approach Lehmann (1963a) also extended the Hodges-Lehmann one-sample estimates to linear models with several observations per cell, and (1964) applied rank-based methods to some linear models with one observation per cell, but this work is of less generality and is not discussed further here.)

Lehmann found that the simple Hodges-Lehmann estimates of the $\mu_i - \mu_j$,

$$D_{ij} = \text{med} \left\{ Y_{ik} - Y_{j\ell} : k = 1,\ldots,n_i; \ \ell = 1,\ldots,n_j \right\}, \tag{2.6}$$

the median of the set of all pairwise differences among the observations in cells i and j, were unsatisfactory, since these estimates do not satisfy the linearity constraints which the $\mu_i - \mu_j$ themselves do:

$$(\mu_i - \mu_j) + (\mu_j - \mu_k) = (\mu_i - \mu_k) \quad , \tag{2.7}$$

but

$$D_{ij} + D_{jk} \neq D_{ik} \quad . \tag{2.8}$$

This makes them unsuitable as a basis for tests of linear hypotheses about the μ_i. He proposed instead adjusting the Hodges-Lehmann estimates and estimating $\mu_i - \mu_j$ by

$$W'_{ij} = \bar{D}'_i - \bar{D}'_j \quad , \tag{2.9}$$

where

$$\bar{D}'_i = I^{-1} \sum_{k=1}^{I} D_{ik} \quad . \tag{2.10}$$

The linearity problem was thus removed, at the cost of offending intuition by using observations in cells other than i and j to help in the estimation of $\mu_i - \mu_j$. Lehmann pointed out, however, that the size of the influence of cells other than i and j on the estimator of $\mu_i - \mu_j$ tends to 0 in probability as the sample sizes increase. A different drawback of this estimation method was noticed by Spjøtvoll (1968) -- cells with unequal numbers of

observations get equal weight in the calculation of the $\bar{D}_i'$. Spjøtvoll suggested several ways of remedying the situation, the simplest of which was to work with

$$W_{ij} = \bar{D}_i - \bar{D}_j , \qquad \bar{D}_i = N^{-1} \sum_{k=1}^{I} n_k D_{ik} . \tag{2.11}$$

This is the form of Hodges-Lehmann estimation which is used in what follows. Note that $\bar{D}_i$ is an estimate of

$$N^{-1} \sum_{k=1}^{I} n_k(\mu_i - \mu_k) = \mu_i - \bar{\mu} , \tag{2.12}$$

and that, since $D_{ji} = -D_{ij}$,

$$\sum_{i=1}^{I} n_i \bar{D}_i = 0. \tag{2.13}$$

A natural estimate of the contrast (2.4) is then

$$\hat{\phi} = \sum_{i=1}^{I} \sum_{j \neq i} b_{ij} W_{ij} . \tag{2.14}$$

It seems on the face of it that the resulting estimate will not be unique, since the b_{ij} are not; but in fact it is straightforward to show that, by virtue of the above linearization, all choices of b_{ij} lead to the same estimate of ϕ.

Lehmann's method of constructing robust tests of linear hypotheses based on these estimates was to look at how the classical tests worked and to replace the classical estimates by their robust analogues, the W_{ij}. A linear

hypothesis H in the model (2.1) amounts to placing some number q of linearly independent constraints on the vector $(\mu_1,\ldots,\mu_I)$ of cell means, so H can always be expressed in terms of a statement that this vector lies in an $(I-q)$-dimensional subspace of $\mathbb{R}^I$. The classical test statistic for such a hypothesis is based on

$$C = \sum_{i=1}^{I} n_i(Y_{i\cdot} - \hat{\mu}_i)^2 \quad , \tag{2.15}$$

where

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i \tag{2.16}$$

and $(\hat{\mu}_1,\ldots,\hat{\mu}_I)$ is the projection of the cell means vector into the subspace of $\mathbb{R}^I$ specified by H. This statistic, when divided by the variance σ^2 of the error distribution, possesses a chi-square distribution with q degrees of freedom under H when the error density f is $N(0,\sigma^2)$, and under mild regularity conditions (see Huber (1972))

$$C/\sigma^2 \xrightarrow[\text{under } H]{D} \chi_q^2 \tag{2.17}$$

as $n_1,\ldots,n_I \to \infty$ even if f is not normal. If σ^2 is not known, it is of course necessary to estimate it to obtain a working test statistic for H; with any consistent estimate $\hat{\sigma}^2$ used in place of σ^2 in the denominator of (2.17), the asymptotic distribution will still be χ_q^2. The classical estimate of σ^2 is

$$\hat{\sigma}^2 = (N-1)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\cdot})^2 \quad , \tag{2.18}$$

which, when divided by σ^2 and multiplied by (N-I), is χ^2_{N-I} when f is normal, so that since $\hat{\sigma}^2$ and C are independent the ratio

$$F_C = \frac{(C/\sigma^2)/q}{[(N-I)\hat{\sigma}^2/\sigma^2]/(N-I)} = \frac{C/q}{\hat{\sigma}^2} \tag{2.19}$$

has a null F distribution with q and N-I degrees of freedom under normality.

Lehmann's (1963b) approach to obtaining rank-based tests of linear hypotheses involving contrasts was in effect to note that W_{ij} is an estimate of $\mu_i - \mu_j$, with corresponding classical estimate $Y_{i\cdot} - Y_{j\cdot}$, so that rewriting the classical numerator as

$$C = \sum_{i=1}^{I} n_i \left[\sum_{j \neq i} a_{ij} (Y_{i\cdot} - Y_{j\cdot}) \right]^2 \quad , \tag{2.20}$$

which is always possible since the $\hat{\mu}_i$ are linear functions of the $Y_{i\cdot}$, the robust analogue of C becomes clear:

$$L = \sum_{i=1}^{I} n_i \left(\sum_{j \neq i} a_{ij} W_{ij} \right)^2 \quad . \tag{2.21}$$

Another way to put it, considering (2.12), is that $\bar{D}_i$ and $Y_{i\cdot} - \bar{Y}$ (where $\bar{Y} = N^{-1} \sum_{i=1}^{I} n_i Y_{i\cdot}$) are estimating the same thing, $(\mu_i - \bar{\mu})$, so that the analogue of

$$C = \sum_{i=1}^{I} n_i \left\{ \sum_{j \neq i} a_{ij} [(Y_{i\cdot} - \bar{Y}) - (Y_{j\cdot} - \bar{Y})] \right\}^2 \tag{2.22}$$

is

$$L = \sum_{i=1}^{I} n_i \left[\sum_{j \neq i} a_{ij}(\bar{D}_i - \bar{D}_j) \right]^2 . \tag{2.23}$$

It is usually not necessary in practice to determine the a_{ij}; to obtain L for a given problem one simply replaces the quantities $Y_{i\cdot} - \bar{Y}$ in the classical numerator by $\bar{D}_i$ (or, equivalently and even more simply, in view of (2.13) one can replace just $Y_{i\cdot}$ by $\bar{D}_i$).

Example 1. In the one-way layout, for the usual hypothesis

$$H_A : \mu_1 = \ldots = \mu_I \tag{2.24}$$

(here $q = I - 1$), the classical statistic assumes the form

$$C_A = \sum_{i=1}^{I} n_i (Y_{i\cdot} - \bar{Y})^2 , \tag{2.25}$$

so the robust numerator is simply

$$L_A = \sum_{i=1}^{I} n_i \bar{D}_i^2 . \qquad \square \tag{2.26}$$

Example 2. Consider the r be c two-way layout with equal numbers, say n, of observations per cell. The usual notation for this model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad \begin{Bmatrix} i = 1,\ldots,r \\ j = 1,\ldots,c \\ k = 1,\ldots,n \end{Bmatrix} , \tag{2.27}$$

subject to the side conditions

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = \sum_{i=1}^{r} \gamma_{i\ell} = \sum_{j=1}^{c} \gamma_{mj} = 0 \tag{2.28}$$

for all $\ell = 1,\ldots,c$ and $m = 1,\ldots,r$. The three hypotheses addressed by the usual analysis of variance table are

$$H_A : \alpha_1 = \ldots = \alpha_r = 0$$

$$H_B : \beta_1 = \ldots = \beta_c = 0 \tag{2.29}$$

$$H_{AB} : \gamma_{11} = \ldots = \gamma_{rc} = 0 \quad .$$

The classical numerators for H_A, H_B, and H_{AB} can be expressed as

$$C_A = nc \sum_{i=1}^{r} (Y_{i\cdot\cdot} - Y_{\cdot\cdot\cdot})^2 \tag{2.30}$$

$$C_B = nr \sum_{j=1}^{c} (Y_{\cdot j\cdot} - Y_{\cdot\cdot\cdot})^2 \tag{2.31}$$

$$C_{AB} = n \sum_{i=1}^{r} \sum_{j=1}^{c} (Y_{ij\cdot} - Y_{\cdot\cdot\cdot})^2 - C_A - C_B \quad , \tag{2.32}$$

where as in (2.16) the dot notation indicates averaging over the indicated subscript(s). Here $Y_{ij\cdot}$ and $Y_{\cdot\cdot\cdot}$ play the roles of $Y_{i\cdot}$ and $\bar{Y}$ above, and

$$Y_{i\cdot\cdot} - Y_{\cdot\cdot\cdot} = c^{-1} \sum_{j=1}^{c} (Y_{ij\cdot} - Y_{\cdot\cdot\cdot}) \tag{2.33}$$

and similarly for $Y_{\cdot j \cdot} - Y_{\cdots}$, so numbering the $I = r \cdot c$ cells as follows:

	column	1	2	⋯	c
row	1	1	2	⋯	c
	2	c+1	c+2	⋯	2c
	⋮	⋮	⋮	⋮	⋮
	r	(r-1)c+1	(r-1)c+2	⋯	rc = I

(2.34)

the robust numerators, in the notation of model (2.1), come out

$$L_A = nc \sum_{i=1}^{r} \left(c^{-1} \sum_{j=1}^{c} \bar{D}_{j+(i-1)c} \right)^2$$

$$L_B = nr \sum_{j=1}^{c} \left(r^{-1} \sum_{i=1}^{r} \bar{D}_{j+(i-1)c} \right)^2 \qquad (2.35)$$

$$L_{AB} = n \sum_{i=1}^{I} \bar{D}_i^2 - L_A - L_B \,. \qquad \square$$

This approach of replacing classical estimates by their robust counterparts in the classical test statistic numerators to obtain the robust numerators works, of course, only when closed-form expressions exist for the classical statistics. This excludes many situations in unbalanced two- and higher-way layouts. In problems of this type the classical numerator is often found in effect by solving a system of linear equations in the Y_{ij}, or, equivalently by sufficiency, in the $Y_{i \cdot}$ (cf. Scheffé (1959), section 4.4) to determine

the $\hat{\mu}_i$, after which the $\hat{\mu}_i$ are substituted into (2.15); no closed-form expression for the resulting numerator will be possible. In such cases the form of the Lehmann numerator is equally obscure, but its value can be found simply by replacing the $Y_{i\cdot}$ by the $\bar{D}_i$ in the system of equations whose solution determines the classical numerator and proceeding as in the classical case.

Example 3. In the r by c two-way layout of Example 2 with unequal numbers n_i of observations per cell, the robust numerator for H_A is derived by analogy from the classical to be

$$L_A = \sum_{i=1}^{r} \left[\frac{\left(\sum_{j=1}^{c} \bar{D}_{j+(i-1)c} \right)^2}{\sum_{j=1}^{c} n_{j+(i-1)c}^{-1}} \right] - \frac{\left[\sum_{i=1}^{r} \left(\sum_{j=1}^{c} n_{j+(i-1)c}^{-1} \right)^{-1} \sum_{j=1}^{c} \bar{D}_{j+(i-1)c} \right]^2}{\sum_{i=1}^{r} \left(\sum_{j=1}^{c} n_{j+(i-1)c}^{-1} \right)^{-1}}, \quad (2.36)$$

and similarly for L_B; but it is necessary to solve a system of linear equations to obtain the form of the classical numerator for H_{AB}, so that no simple expression exists for either the robust or the classical numerator in that case. □

The analogy between the classical and robust procedures carries over to the asymptotic distributions. Lehmann showed that

$$L/\sigma_R^2 \xrightarrow[\text{under } H]{D} \chi_q^2 , \quad (2.37)$$

where

$$\sigma_R^2 = (12\theta^2)^{-1} = [12(\int f^2)^2]^{-1} \tag{2.38}$$

plays the role for the rank-based numerator which the error variance σ^2 plays for the classical statistic. Their ratio

$$\sigma^2/\sigma_R^2 = 12\sigma^2(\int f^2)^2 \tag{2.39}$$

is the asymptotic relative efficiency of the robust procedure to the classical; this is just the familiar expression for the efficiency of the Wilcoxon one- and two-sample procedures relative to the corresponding classical t-methods. Table 1 gives some values of this efficiency as well as values of σ_R^2 and $1/\theta$ for various distributions.

TABLE 1. σ_R^2, $1/\theta$, and asymptotic relative efficiency of Lehmann's two-sample ANOVA methods to the classical.

Distribution f	σ^2	$1/\theta$	σ_R^2	$e_{L,C}(f) = \sigma^2/\sigma_R^2$
Standard normal	1.0	3.544	1.047	0.9549
Standard logistic	3.290	6.0	3.0	1.097
χ^2 with 8 degrees of freedom	16.0	12.8	13.65	1.172
Skewed mixed normal (λ=0.75, μ_1=0, σ_1=1, μ_2=1.9, σ_2=2)	2.426	4.670	1.817	1.335
(λ=0.82, μ_1=0, σ_1=1, μ_2=1.9, σ_2=3.5)	3.558	4.535	1.714	2.076
t with 3 degrees of freedom	3.0	4.353	1.579	1.900
Any f	--	--	--	$\geq$0.864

The skewed mixed normal distribution referred to in Table 1 has cumulative distribution function (cdf)

$$F(x) = \lambda\Phi[(x-\mu_1)/\sigma_1] + (1-\lambda)\Phi[(x-\mu_2)/\sigma_2] \quad , \tag{2.40}$$

where Φ is the standard normal cdf.

It can be seen that considerable efficiency gains are possible using the rank-based methods on heavy-tailed data, with a loss of only about 4% efficiency for normal data and with a potential loss for any distribution never to exceed about 14%. These are of course asymptotic results, but, as is documented in Draper (1982b), empirical small-sample efficiencies support these figures in designs with as few as $N = 10$ total observations.

Just as in the classical case, σ_R^2, the denominator of the test statistic in (2.37), will not be known in practice and it is necessary to estimate it to obtain usable test statistics, and just as in the classical case replacement of σ_R^2 by any consistent estimate $\hat{\sigma}_R^2$ will result in a statistic whose limiting distribution is still χ_q^2. The estimation of σ_R^2 is described in Section 4 below.

Lehmann's original proposal was to obtain critical values for the usable test statistic

$$L/\hat{\sigma}_R^2 \tag{2.41}$$

from this χ_q^2 distribution, but, as is discussed below and shown in Draper (1982b), the distribution of

$$\nu\hat{\sigma}_R^2/\sigma_R^2 \tag{2.42}$$

is approximately χ_ν^2 for a ν which depends on the method used for estimating σ_R^2, and L and $\hat{\sigma}_R^2$ are approximately independent under H, so that a better small-sample null distribution for the ratio

$$F_L = \frac{(L/\sigma_R^2)/q}{(\nu\hat{\sigma}_R^2/\sigma_R^2)/\nu} = \frac{L/q}{\hat{\sigma}_R^2} \tag{2.43}$$

is the F distribution with q and ν degrees of freedom.

As Lehmann pointed out (1963b), the above rank-based test statistics would be expected to work well in certain random-effects and mixed models as well as in the fixed-effects model (2.1), but that possibility has not yet been investigated.

A computer program to carry out the Lehmann estimation and testing procedures in arbitrary one-way layouts and balanced higher-way layouts with several observations per cell, using a denominator estimate $\hat{\sigma}_R^2$ described below in Section 4, is available from the author.

3. The Techniques of Jaeckel and Hettmansperger-McKean: General Linear Model

The model considered by Jaeckel (1972) in his development of rank-based estimation methods, and by Hettmansperger and McKean (1976, 1977) in their application of Jaeckel's methods to hypothesis testing, is the general fixed-effects linear model of Section 1:

$$(\Omega) \qquad Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \,, \qquad i = 1,\ldots,N \,, \tag{3.1}$$

or, in matrix form,

$$(\Omega) \qquad \underset{N\,1}{Y} = \underset{N\,1}{\beta_0} + \underset{N\,p}{x}\,\underset{p\,1}{\beta} + \underset{N\,1}{e} \,, \tag{3.2}$$

in which as in model (2.1) the e_i are i.i.d. continuous random variables with density f such that both $\theta = \int f^2$ and $\sigma^2 = \mathrm{Var}(e_i)$ are finite; x is a matrix of known constants. Jaeckel's work simplifies and makes more usable in practice an approach to robust estimation in the linear model due to Jurečková (1971), who was generalizing the work of Hodges and Lehmann (1963) described above on inverting rank tests to obtain estimates. Jaeckel's starting point is, like Lehmann's in Section 2, with the classical estimates; but his rank-based modifications are quite different. He considers the errors or residuals e_i as a function of the parameter vector $\underset{\sim}{\beta}' = (\beta_0,\ldots,\beta_p)$,

$$e_i(\underset{\sim}{\beta}') = Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \,, \tag{3.3}$$

and seeks estimates which make the residuals as small as possible. The measure of residual size minimized by the classical estimates is the ordinary Euclidean square norm,

$$D_C[\underline{e}(\underline{\beta}')] = \|\underline{e}(\underline{\beta}')\|^2 = \sum_{i=1}^{N} e_i^2(\underline{\beta}')$$

$$= \sum_{i=1}^{N} e_{(i)}(\underline{\beta}') \cdot e_{(i)}(\underline{\beta}') \quad , \tag{3.4}$$

in which $e_{(i)}(\underline{\beta}')$ is the ith ordered residual. Note however that the size of a vector of observations $\underline{z} = (z_1,\ldots,z_N)$ has both a dispersion component and a centering component; for example, the Euclidean square norm of $\underline{z}$ can be written

$$\|\underline{z}\|^2 = \|\underline{z} - \bar{\underline{z}}\|^2 + \|\bar{\underline{z}}\|^2 \quad , \tag{3.5}$$

where $\bar{\underline{z}}$ is a vector all of whose elements are $\bar{z} = N^{-1} \sum_{i=1}^{N} z_i$.

An alternative to minimizing (3.4) in arriving at the classical estimates involves first minimizing only the dispersion part of (3.5) applied to the residuals, $\|\underline{e} - \bar{\underline{e}}\|^2 = N\,\mathrm{Var}(\underline{e})$. This yields the classical estimates $\hat{\beta}_j$, not of all of the $\underline{\beta}$ but only of $(\beta_1,\ldots,\beta_p)$, since this dispersion measure is translation-invariant and β_0 drops out. Then the centering part $\|\bar{\underline{e}}\|^2$ of (3.5) with the previously found $(\hat{\beta}_1,\ldots,\hat{\beta}_p)$ substituted in is minimized to yield $\hat{\beta}_0$. In effect, first the model (3.1) is recast so that β_0 is regarded as the center of the distribution of the e_i; then $\underline{\beta} = (\beta_1,\ldots,\beta_p)$ is estimated by $\hat{\underline{\beta}}$; and finally β_0 is estimated as the

center of the residuals $Y_i - (x\hat{\underline{\beta}})_i$. Note that in the usual method of arriving at the classical estimates, in which $\|\underline{e}\|^2$ is minimized as a function of β_0 as well as of $(\beta_1,\ldots,\beta_p)$, the minimizing condition which specifies $\hat{\beta}_0$ in terms of $(\hat{\beta}_1,\ldots,\hat{\beta}_p)$ is $\bar{e} = 0$, so the classical measure of residual size (3.4, 3.5) can also be thought of as a dispersion measure.

This alternative approach is the one taken by Jaeckel and Hettmansperger-McKean. As in the classical case Jaeckel also restricts himself to translation-invariant dispersion measures and makes no attempt to estimate β_0; Hettmansperger and McKean later proposed a rank-based estimate of β_0 which is described below.

The trouble with the classical method is that the classical dispersion measure (3.4) places too much weight on the extreme residuals when the data contain gross errors or have a distribution with tails heavier than those of the normal. Jaeckel's dispersion function replaces one of the ordered residuals $e_{(i)}$ in the product in (3.4) by a value or score $a(i)$ based on it which gives less weight to the largest and smallest errors:

$$D_J[\underline{e}(\underline{\beta}')] = \sum_{i=1}^{N} a(i)e_{(i)}(\underline{\beta}') = \sum_{i=1}^{N} a[R_i'(\underline{\beta}')]e_i(\underline{\beta}) \quad , \tag{3.6}$$

where $R_i'(\underline{\beta}')$ is the rank of $e_i(\underline{\beta}')$ among $e_1(\underline{\beta}'),\ldots,e_N(\underline{\beta}')$. To insure the translation invariance of D_J Jaeckel requires of the scores $a(i)$ that they sum to zero; with this condition $D_J[e(\beta')]$ no longer depends on β_0:

$$D_J[\underline{Y} - \beta_0 - x\underline{\beta}] = D_J[\underline{Y} - x\underline{\beta}] = \sum_{i=1}^{N} a[R_i(\underline{\beta})][Y_i - (x\underline{\beta})_i] \quad , \tag{3.7}$$

where $R_i(\underline{\beta})$ is the rank of $Y_i - (x\underline{\beta})_i$ among $Y_1 - (x\underline{\beta})_1, \ldots, Y_N - (x\underline{\beta})_N$. Further, in order that the resulting dispersion measure D_J be convex, and thus readily minimized, the scores must be monotone:

$$a(1) \leq \ldots \leq a(N) \quad . \tag{3.8}$$

Hettmansperger and McKean add to these requirements that of symmetry of the scores,

$$a(i) = -a(N+1-i) \ , \tag{3.9}$$

an assumption which is not necessary in general and which is natural only in the context of the assumption that the e_i are symmetrically distributed (a restriction which is also not needed in Jaeckel's estimation of $(\beta_1, \ldots, \beta_p)$).

Different choices of the scoring function $a(\cdot)$ give rise to estimates with different properties. The simplest are the piecewise constant sign scores

$$a_S(i) = \text{sign}[i/(N+1) - 1/2] \tag{3.10}$$

and the linear Wilcoxon scores

$$a'_W(i) = i/(N+1) - 1/2 \quad . \tag{3.11}$$

Other possibilities include Van der Waerden-type normal scores,

$$a_{VW}(i) = \Phi^{-1}[i/(N+1) - 1/2] \quad , \tag{3.12}$$

and a mixture of sign and Wilcoxon scores proposed by Policello and Hettmansperger (1976), in which a fraction, $\eta/2$ say, of the residuals at each end are given sign scores and the remaining $(1 - \eta)$ in the middle receive Wilcoxon scores.

Choice of scores is based on a compromise between resistance to outliers and gross errors on the one hand and efficiency considerations on the other. The sign scores have excellent resistance properties but are too inefficient for most data. In the one- and two-sample problems the Wilcoxon scores are known to strike a good balance between efficiency and robustness for many distributions. Hettmansperger and McKean (1977) suggest choosing the scoring function adaptively, by using the data to estimate the optimal fraction of sign and Wilcoxon scores in the Policello mixture. Wilcoxon scores are considered exclusively here; further work is needed to see if optimizing the scores by adapting them to the data at hand significantly improves the performance of this method. (The same investigation could be undertaken for the Lehmann approach of Section 2, in which the Wilcoxon scores were used implicitly; see Hodges and Lehmann (1963).)

It is convenient in what follows to renormalize the Wilcoxon scores a'_W and use instead

$$a_W(i) = 12^{1/2}[i/(N+1) - 1/2] \quad . \tag{3.13}$$

(In general, one way to construct normalized $a(i)$ is to base them on a score-generating function $\phi:[0,1] \to \mathbb{R}$ as follows:

$$a(i) = A^{-1} \phi[i/(N+1)] \quad , \tag{3.14}$$

in which

$$A^2 = \int_0^1 \phi^2(u)du \quad . \tag{3.15}$$

For the Wilcoxon scores $\phi_W(u) = u - 1/2$ and $A_W^2 = 1/12$. See Jurečková (1971) and Hettmansperger and McKean (1977).) The resulting Wilcoxon-type dispersion measure is

$$D_{JW}[\underset{\sim}{Y} - x\underset{\sim}{\beta}] = 12^{1/2}(N+1)^{-1} \sum_{i=1}^{N} [R_i(\underset{\sim}{\beta}) - (N+1)/2][Y_i - (x\underset{\sim}{\beta})_i] \quad . \tag{3.16}$$

Solving for the $\hat{\underset{\sim}{\beta}}_{JW}$ which minimizes $D_{JW}[\underset{\sim}{Y} - x\underset{\sim}{\beta}]$ yields the rank analogue of the normal equations:

$$\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)[R_i(\hat{\underset{\sim}{\beta}}_{JW}) - (N+1)/2] \doteq 0 \quad , \qquad j = 1,\ldots,p \quad , \tag{3.17}$$

where

$$\bar{x}_j = N^{-1} \sum_{i=1}^{N} x_{ij} \quad . \tag{3.18}$$

The "equations" (3.17) are solved in the sense that a value of $\hat{\underset{\sim}{\beta}}_{JW}$ is

sought which makes the left-hand side of (3.17) as close to 0 as possible. The resulting solutions typically do not have closed-form expressions, and iterative computer methods are generally needed to find numerical solutions. Hettmansperger and McKean (1976) have investigated several algorithms, including steepest descent and regula falsi, and report good results with both. Even worse, however, than the solutions to (3.17) not having closed-form expressions, the solutions are not necessarily unique; but Jaeckel (1972) showed that the diameter of the solution set is bounded and goes to 0 in probability as $N \to \infty$. In theoretical situations where this indeterminacy is troublesome, a unique estimator can be identified by taking the centroid of the minimizing set or by minimizing $D_{JW}[\underline{Y} - x\underline{\hat{\beta}}_{JW}]$ (3.16) over all $\hat{\beta}_{JW}$ in the solution set to (3.17). In practice the simpler approach of just being satisfied with whichever point in the minimizing set the iterative convergence has yielded seems to work well. Note however that, due to differences in computer hardware, the same computer program to perform the iterative search for $\hat{\beta}_{JW}$ run on different computers may yield somewhat different estimates.

Example 4. One situation in which the Jaeckel estimates based on Wilcoxon scores do have a closed-form expression is in linear regression with only one independent variable. Rewriting the model (3.1) in this case as

$$Y_i = \alpha + \beta x_i + e_i , \qquad i = 1,\ldots,N , \tag{3.19}$$

the Jaeckel estimate of the slope β, which was first derived by Adichie (1967) (although Adichie did not realize that his estimator had a closed-form expression and Jaeckel did not recognize his estimator to be the solution

of the equation which implicitly defined Adichie's estimate) is a weighted median of the set

$$\{(Y_j - Y_i)/(x_j - x_i),\ (i,j) \ni x_i \neq x_j\} \tag{3.20}$$

of all pairwise slopes, in which the weights are proportional to the absolute distance $|x_i - x_j|$ between the independent variable values. This model is examined in more detail in Draper (1982c), in which a new rank-based robust alternative to the Adichie-Jaeckel estimator is proposed for use in models with several observations per cell. Note that in the usual two-sample shift model, in which n of the x_i are 0 and the remaining $m = N - n$ are 1, the Adichie-Jaeckel estimator is simply the two-sample Hodges-Lehmann estimate of the shift β. □

As described above, estimation of β_0 with the Jaeckel-Hettmansperger-McKean approach involves applying an estimator of location to the residuals $\hat{e}_i^* = Y_i - (x\underline{\hat{\beta}}_J)_i$. As in Section 2, in order that β_0 even be identifiable it is necessary to specify the manner in which the errors e_i are regarded as centered at 0; if for example $E(e_i) = 0$ is assumed then β_0 is the mean of the distribution of the random variables $Y_i - (x\underline{\beta})_i$. If the e_i are further assumed to be symmetrically distributed about 0, then a reasonable choice for an estimate of β_0 is the one-sample Hodges-Lehmann estimator applied to the $\hat{e}_i^*$, namely

$$\hat{\beta}_0 = \text{med}\left\{(\hat{e}_i^* + \hat{e}_j^*)/2,\ 1 \leq i \leq j \leq N\right\}\ , \tag{3.21}$$

the median of the set of all pairwise averages of the residuals $Y_i - (x\hat{\underline{\beta}}_J)_i$. This is Hettmansperger and McKean's recommended estimate of β_o.

How can the Jaeckel estimation technique serve as the basis for tests of linear hypotheses? McKean and Hettmansperger's (1976) approach to constructing tests based on the Jaeckel estimates was, like Lehmann's, to take as a starting point the classical techniques. In testing a hypothesis H which places q linearly independent restrictions on the $\underline{\beta}$ vector, it is convenient to parameterize the model (3.1) in such a way that the design matrix x has full rank p. Denote by ω the model Ω plus the restrictions imposed by H. Expressed in terms of the classical residual dispersion measure, the classical test statistic for H is based on

$$D_C^* = D_C[\underline{Y} - x\hat{\underline{\beta}}_{\omega,C}] - D_C[\underline{Y} - x\hat{\underline{\beta}}_{\Omega,C}] \quad , \tag{3.22}$$

the amount of extra lack of fit imposed by accepting the model ω over and above that inherent in Ω. Here $\hat{\underline{\beta}}_{\omega,C}$ and $\hat{\underline{\beta}}_{\Omega,C}$ are the estimates which minimize the classical dispersion measure under ω and Ω, respectively.

As in Section 2, this statistic, when divided by the variance σ^2 of the error density f, is χ^2_q under H when f is normal, and under the same mild regularity conditions as in Section 2 is asymptotically χ^2_q even if f is not normal. As before it is typically necessary to estimate σ^2; the classical estimate is

$$\hat{\sigma}^2 = [N - (p+1)]^{-1} \sum_{i=1}^{N} [Y_i - (x\hat{\underline{\beta}}_{\Omega,C})_i - \hat{\beta}_{o,C}]^2 \quad , \tag{3.23}$$

where

$$\hat{\beta}_{o,C} = N^{-1} \sum_{i=1}^{N} [Y_i - (x\hat{\underline{\beta}}_{\Omega,C})_i] \tag{3.24}$$

is the sample mean of the full model (Ω) residuals $\underline{Y} - x\hat{\underline{\beta}}_{\Omega,C}$.

The asymptotic null distribution of $D_C^*/\hat{\sigma}^2$ is still χ_q^2, but since $[N-(p+1)]\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-(p+1)}$ and D_C^* and $\hat{\sigma}^2$ are independent under normality, a better small-sample distribution when F is normal for

$$F_C = \frac{(D_C^*/\sigma^2)/q}{\{[N-(p+1)]\hat{\sigma}^2/\sigma^2\}/[N-(p+1)]} = \frac{D_C^*/q}{\hat{\sigma}^2} \tag{3.25}$$

under H is $F_{q,N-(p+1)}$.

Hettmansperger and McKean proposed using the same approach but with the Jaeckel dispersion measure instead of the classical. With the Wilcoxon scores this involves basing a test of H on

$$D_{JW}^* = D_{JW}[\underline{Y} - x\hat{\underline{\beta}}_{\omega,JW}] - D_{JW}[\underline{Y} - x\hat{\underline{\beta}}_{\Omega,JW}] \ , \tag{3.26}$$

where as above $\hat{\underline{\beta}}_{\omega,JW}$ and $\hat{\underline{\beta}}_{\Omega,JW}$ are the parameter estimates under ω and Ω, respectively, which minimize the Jaeckel dispersion measure. They found that

$$2D_{JW}^*/\sigma_R \xrightarrow[\text{under } H]{D} \chi_q^2 \ , \tag{3.27}$$

where as before

$$\sigma_R = 12^{-1/2}\,\theta^{-1} = 12^{-1/2}\left(\int f^2\right)^{-1} . \tag{3.28}$$

Note that, unlike in the Lehmann method, θ enters into the asymptotic distribution of the test statistic through $1/\theta$ rather than through $1/\theta^2$. Even so, Hettmansperger and McKean showed that, as was the case for the Lehmann method, the asymptotic efficiency of the Hettmansperger-McKean approach relative to the classical is $\sigma^2/\sigma_R^2 = 12\sigma^2\theta^2$. Thus the Lehmann and Hettmansperger-McKean methods are asymptotically equally effective.

As in Section 2, with any consistent estimate $\hat{\sigma}_R$ of σ_R the limiting null distribution of

$$2D^*_{JW}/\hat{\sigma}_R \tag{3.29}$$

is still χ^2_q, and this was the distribution originally proposed in practice by Hettmansperger and McKean. They later (1977) found that the χ^2_q distribution is too light-tailed for use in small and moderate-size samples. In searching for a heavier-tailed approximation to the small-sample distribution of $2D^*_{JW}/\hat{\sigma}_R$, they suggested, without much justification except by analogy with the classical methods, the approximation of the null distribution of

$$(2D^*_{JW}/q)/\hat{\sigma}_R \tag{3.30}$$

by the $F_{q,N-(p+1)}$ distribution. The success of this and other approximations is discussed in Draper (1982b) and in Section 4 below.

The distribution of $2D^*_{JW}/\sigma_R$ under alternatives to the null was also

considered by McKean and Hettmansperger (1976). They studied its behavior for a sequence of alternatives contiguous to the null and found that its limiting distribution under such a sequence is noncentral χ^2 with q degrees of freedom and noncentrality parameter

$$\delta_R = \sigma^2 \delta_C / \sigma_R^2 \quad , \tag{3.31}$$

where δ_C is the noncentrality parameter for the noncentral χ^2 distribution of the classical numerator under the same conditions. The same argument applied to the Lehmann numerator L of Section 2 yields the same conclusion about its limiting non-null distribution, raising the possibility of approximating the distribution of $F_L = (L/q)/\hat{\sigma}_R^2$ (2.43) under an alternative by $F_{q,\nu;\delta_R}$, the noncentral F distribution with q and ν degrees of freedom and noncentrality parameter δ_R (3.31). The usefulness of these asymptotic facts as the basis of small-sample power approximations is addressed empirically in Draper (1982b).

Note that the assumption of symmetry of the error density f about 0 is used in the Hettmansperger-McKean approach in three ways: in the estimation of β_0; (as will be seen in Section 4) in the construction of the denominator estimate $\hat{\sigma}_R$; and implicitly in the choice of a symmetric scoring function for the Jaeckel estimates. In the testing of many linear hypotheses it is not necessary to estimate β_0, and in some linear models estimates of σ_R are available which do not require the symmetry assumption (Section 4). Thus, as Draper (1982b) shows, in some contexts the Hettmansperger-McKean method will perform well with asymmetric as well as symmetric f.

4. Empirical results

The following is a brief description of results pertaining to the estimation of σ_R^2 through estimating $\theta = \int f^2$ or its reciprocal, and to the empirical small-sample performance of the rank-based robust testing ratios of Lehmann and Hettmansperger-McKean outlined above. For more detail see Draper (1982a, 1982b).

Two approaches to the estimation of σ_R^2 in linear models were investigated: a method for estimating $1/\theta$ due to Lehmann (1963c) based on the lengths of distribution-free confidence intervals, and an approach to the estimation of θ due to Schweder (1975) based on kernel-type density estimation of f. The Lehmann method has both one- and two-sample versions, which derive from the idea of estimating $1/\theta$ with normalized lengths of confidence intervals based on the Wilcoxon one-sample (signed rank) and two-sample (rank-sum) statistics. Schweder's approach to estimating θ yields a one-sample method based on the notion that $\theta = \int f^2 = \int f \, dF$, where F is the error cdf, can be estimated by $\hat{\theta} = \int \hat{f} \, d\hat{F}$, in which $\hat{F}$ is the empirical cdf and $\hat{f}$ is a kernel- or window-type density estimate of f. In linear models with several observations per cell like (2.1) the Lehmann and Schweder one-sample methods can either be applied to the cells separately, with a composite estimator formed as a weighted average of the separate one-sample cell estimates, or to the residuals of the full linear model fit, treating the residuals as one large sample. (This one-sample residuals idea also of course works in more general linear models like (3.1).) The Lehmann two-sample estimator can be applied separately to all pairs of cells in the layout, with a weighted

average composite estimator again constructed.

The Lehmann one-sample and Schweder methods both have implementational drawbacks. Since the Lehmann one-sample approach is based on the Wilcoxon signed-rank statistic, it requires the assumption of symmetry of the error distribution f for proper functioning, making its performance with asymmetric data problematic. Simulations verify that this method of estimating σ_R^2 is subject to increased bias and standard error when applied to skewed data. The Schweder approach uses window-type density estimation and thus requires specification of the window width, something which in principle needs to be done in a way which adapts the window width to the data at hand both in terms of the shape and the scale of the underlying distribution. Schweder's (1975) implicit position on this issue was that, as far as shape is concerned, the estimator is not too sensitive to imprecise choice of the window width, so that apart from the scaling issue one could get by with a fixed non-adaptive window width which depends only on the sample size. Simulations unfortunately indicate that this is not so; estimates of σ_R^2 based on the non-adaptive Schweder approach behave poorly with mis-specified window width. It may be that the Schweder method will work better when the window width is chosen adaptively, for example using a cross-validation or nearest-neighbor procedure, but this has not yet been pursued. The only method of estimating σ_R^2 which was found empirically to perform well is the Lehmann two-sample approach when modified with a simple bias correction, and it of course only works in models with several observations per cell (in practice this means that most of the cells in the layout should have at least 3 or 4 observations). In more general linear models the Lehmann one-sample estimator applied to the residuals with a different bias correction was also seen to perform well, but only in situations in which the error distribution is fairly symmetric.

The asymptotic χ^2_q distributions for the Lehmann and Hettmansperger-McKean testing ratios $L/\hat{\sigma}^2_R$ (2.41) and $2D^*_{JW}/\hat{\sigma}_R$ (3.29) were found to provide quite poor approximations to the null distributions with even fairly large sample sizes. This is because the extra variability imposed on the ratios by using an estimate of σ^2_R instead of the true value results in distributions with heavier tails than χ^2_q. Huber (1970) conjectured that the small-sample distribution of $\nu\hat{\sigma}^2_R/\sigma^2_R$ might be well approximated by χ^2_ν for a value of ν depending on the error density f; the bias corrected versions of the Lehmann one- and two-sample estimators of σ^2_R described above were seen in simulations to support this conjecture quite well. This encourages the approximation of the null distribution of the Lehmann statistic $(L/q)/\hat{\sigma}^2_R$ by the heavier-tailed $F_{q,\nu}$, but suggests that it might be necessary to estimate the denominator degrees of freedom ν from the data. In practice it was found empirically that the same F distribution which would be used with the classical statistic in the linear model at hand provides a surprisingly good approximation for both the Lehmann and Hettmansperger-McKean ratios $(L/q)/\hat{\sigma}^2_R$ and $(2D^*_{JW}/q)/\hat{\sigma}_R$. This finding is convenient both from the point of view of not having to adapt the null distribution to the data and of preserving the analogy with the classical procedures, thus making the robust methods easier to use by practitioners accustomed to the traditional analysis. The resulting tests not only had approximately correct levels with the wide variety of error distributions listed in Table 1, but also fulfilled their promise in terms of asymptotic efficiency as indicated in that Table by exhibiting excellent power characteristics relative to the classical F-test. Table 2 presents some typical power comparisons, which demonstrate that the

power loss at the normal model for the robust methods is small while the gain with skewed and heavy-tailed distributions can be considerable. In the robust analogues of the estimation and multiple comparisons procedures typical in linear models work this efficiency gain will of course manifest itself in more precise estimates and narrower confidence intervals.

TABLE 2. Typical Monte-Carlo power comparisons between the classical and rank-based robust testing procedures.

Linear Model	Error Density	Testing Method — Numerator	Testing Method — Denominator	Power at Level .10	.05	.01
one-way layout, 6 cells, 10 observations per cell	standard normal	Hettmansperger-McKean	Lehmann two-sample	.94	.87	.70
		Classical		.96	.90	.75
one-way layout, 6 cells, 10 observations per cell	t with 3 degrees of freedom	Lehmann	Lehmann one-sample	.99	.98	.90
		Hettmansperger-McKean	Lehmann two-sample	.99	.97	.88
		Classical		.91	.85	.70
two-way layout, 12 cells, 5 observations per cell	skewed mixed normal (fourth entry in Table 1)	Hettmansperger-McKean	Lehmann two-sample	.72	.61	.34
		Classical		.62	.49	.24

Note: Approximate standard errors for these power estimates $\hat{p}$ based on $n = 1000$ Monte-Carlo replications can be calculated in the usual $[\hat{p}(1-\hat{p})/n]^{1/2}$ binomial manner and range for the given power values from about 0.003 to about 0.016.

In 1959 Henry Scheffé wrote

> "... it appears that there probably exist tests which have the robustness of the [classical] F-tests concerning type I errors, a little less power against normal alternatives, but much greater power against 'most' nonnormal alternatives. At present such tests have not been developed for the relatively complicated hypotheses usually considered in [linear models], and even if they were, the methods of estimation with which one would usually want to follow them up when they rejected, ..., while then possible in principle, would seem hopelessly complicated to carry out in any but the very simplest cases ... ."

Twenty years later such robust testing and multiple-comparison methods have indeed been developed, and the testing methods, at least, are essentially ready for general use.

REFERENCES

Adichie, J. N. (1967). "Estimates of regression parameters based on rank tests," Annals of Mathematical Statistics, Vol. 38, pp. 894-904.

Bickel, P. J. (1973). "On some analogues to linear combinations of order statistics in the linear model," Annals of Statistics, Vol. 1, pp. 597-616.

Bradley, James V. (1978). "Robustness?" British Journal of Mathematical and Statistical Psychology, Vol. 31, pp. 144-152.

Draper, David C. (1981). "Rank-based robust analysis of linear models," unpublished Ph.D. dissertation, Department of Statistics, University of California, Berkeley.

Draper, David C. (1982a). "Rank-based robust analysis of linear models. II. Estimation of the asymptotic variance of rank-based estimators and associated statistics," technical report, Department of Statistics, University of California, Berkeley.

Draper, David C. (1982b). "Rank-based robust analysis of linear models. III. Empirical findings," technical report, Department of Statistics, University of California, Berkeley (in press).

Draper, David C. (1982c). "Rank-based robust inference in regression models with several observations per cell," technical report, Department of Statistics, University of California, Berkeley.

Draper, David C. (1982d). "Robust versus data-analytic methods in linear models analysis," technical report, Department of Statistics, University of Chicago (in press).

Draper, N. R. and Smith, H. (1981). Applied Regression Analysis, 2nd Edition. New York: Wiley.

Hettmansperger, Thomas P. and McKean, Joseph W. (1976). "Computational problems involved in analysis of linear models based on ranks," Proceedings of the Statistical Computing Section of the American Statistical Association, pp. 88-94.

Hettmansperger, Thomas P. and McKean, Joseph W. (1977). "A robust alternative based on ranks to least squares in analyzing linear models," Technometrics, Vol. 19, pp. 275-284.

Hettmansperger, Thomas P. and McKean, Joseph W. (1978). "Statistical inference based on ranks," Psychometrika, Vol. 43, pp. 69-79.

Hodges, Jr., J. L. and Lehmann, E. L. (1963). "Estimates of location based on rank tests," Annals of Mathematical Statistics, Vol. 34, pp. 598-611.

Huber, Peter J. (1970). "Studentizing robust estimates," in Nonparametric Techniques in Statistical Inference, M. L. Puri, ed. Cambridge University Press, pp. 453-463.

Huber, Peter J. (1972). "Robust statistics: a review," Annals of Mathematical Statistics, Vol. 43, pp. 1041-1067.

Huber, Peter J. (1973). "Robust regression: asymptotics, conjectures and Monte Carlo," Annals of Statistics, Vol. 1, pp. 799-821.

Huber, Peter J. (1981). Robust Statistics. New York: Wiley.

Jaeckel, Louis A. (1972). "Estimating regression coefficients by minimizing the dispersion of the residuals," Annals of Mathematical Statistics, Vol. 43, pp. 1449-1458.

Jurečková, Jana (1971). "Nonparametric estimate of regression coefficients," Annals of Mathematical Statistics, Vol. 42, pp. 1328-1338.

Lehmann, E. L. (1963a). "Robust estimation in analysis of variance," Annals of Mathematical Statistics, Vol. 34, pp. 957-966.

Lehmann, E. L. (1963b). "Asymptotically nonparametric inference: an alternative approach to linear models," Annals of Mathematical Statistics, Vol. 34, pp. 1494-1506.

Lehmann, E. L. (1963c). "Nonparametric confidence intervals for a shift parameter," Annals of Mathematical Statistics, Vol. 34, pp. 1507-1512.

Lehmann, E. L. (1964). "Asymptotically nonparametric inference in some linear models with one observation per cell," Annals of Mathematical Statistics, Vol. 35, pp. 726-734.

Lehmann, E. L. (1975). Nonparametrics: Statistical Methods Based on Ranks. San Francisco: Holden-Day.

McKean, Joseph W. and Hettmansperger, Thomas P. (1976). "Tests of hypotheses based on ranks in the general linear model," Communications in Statistics--Theory and Methods, Vol. A5(8), pp. 693-709.

Nie, Norman, H., Hull, C. Hadlai, Jenkins, J. G., Steinbrenner, K., and Bent, D. H. (1975). SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill.

Policello, G. E. II and Hettmansperger, Thomas P. (1976). "Adaptive robust procedures for the one-sample location problem," Journal of the American Statistical Association, Vol. 71, pp. 624-633.

Scheffé, Henry (1959). The Analysis of Variance. New York: Wiley.

Schrader, Ronald M. and McKean, Joseph W. (1977). "Robust analysis of variance," Communications in Statistics -- Theory and Methods, Vol. A6(9), pp. 879-894.

Schweder, Tore (1975). "Window estimation of the asymptotic variance of rank estimators of location," Scandinavian Journal of Statistics, Vol. 2, pp. 113-126.

Seal, H. L. (1967). "The historical development of the Gauss linear model," Biometrika, Vol. 54, pp. 1-24.

Spjøtvoll, Emil (1968). "A note on robust estimation in analysis of variance," Annals of Mathematical Statistics, Vol. 39, pp. 1486-1492.

Weisberg, S. (1980). Applied Linear Regression. New York: Wiley.