Statistical Science
1988, Vol. 3, No. 2, 239–271

Rank-Based Robust Analysis of Linear Models. I. Exposition and Review

David Draper

Abstract. Linear models are widely used in many branches of empirical 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. Two methods of dealing with this that have gained some acceptance are the *data-analytic* and *model expansion* approaches, in which graphical and numerical methods are employed to detect the ways in which the data do not meet the classical assumptions, and either the data are modified appropriately before the classical techniques are applied (data-analytic) or the model is broadened to take account of the departures discovered (model expansion). Another approach involves the use of *robust* methods, which are intended to be sufficiently insensitive to deviations from the classical assumptions that the data may be analyzed without modification or additional (explicit) modeling. In this article a comparison is made between the data-analytic, model expansion and robust approaches to linear models analysis, and the application of one type of robust methods, those based on *R-estimators* (which use the logic of rank tests to motivate inference on the raw data scale), to problems of estimation, testing and confidence and multiple comparison procedures in the general linear model is reviewed.

Key words and phrases: Robust estimation, general inferential strategies, rank-based linear model, *R*-estimators, Hodges-Lehmann, kernel-type density estimation, Bayesian robustness.

1. INTRODUCTION: THE CONTEXT OF ROBUSTNESS IN GENERAL INFERENTIAL STRATEGIES

The linear model is one of the most widely used tools yet devised by statisticians to aid in empirical inquiry. Applications of linear regression, analysis of variance (ANOVA) and analysis of covariance techniques abound in the biological, social, physical and behavioral sciences, as well as in industrial and other business settings. It is a basic truth in mathematical modeling, however, that powerful inferences are often arrived at only through powerful 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

$$(1.1) \quad Y_i = g(X_{i1}, \cdots, X_{ip}) + e_i, \quad i = 1, \cdots, N.$$

Here $(Y_i; X_{i1}, \cdots, X_{ip})$ is the ith of N total observations on the quantitative dependent variable Y and the p quantitative or qualitative (nominal or ordinal) independent variables $X_1, \cdots, X_p$, which are considered to be either under experimenter control or passively observed 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. $g(\cdot)$ is assumed to be of the known functional form

$$(1.2) \quad g(X_{i1}, \cdots, X_{ip}) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j,$$

in which the β_j are unknown parameters. A number of assumptions are made in the classical analysis about

David Draper is a member of the Statistical Research and Consulting Group in the Department of Economics and Statistics at The RAND Corporation, 1700 Main Street, Santa Monica, California 90406.

the errors e_i, which can be listed roughly in order of increasing technical constraint as follows:

- the errors are assumed to have expectation 0, to have the same variance for all i, to be independent and to be identically distributed with density f;
- the error 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, including linearity of the relationship between Y and the X_i.

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), four basic approaches have arisen for dealing with the issue of violation of these technical assumptions:

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

1) The *data-analytic* approach, in which graphical and numerical tools like residual and adjusted variable plots (Draper and Smith, 1981; Daniel and Wood, 1980; Chambers, Cleveland, Kleiner and Tukey, 1983), and quantitative methods for the identification of influential observations (Weisberg, 1985; Belsley, Kuh and Welsch, 1980; Cook and Weisberg, 1982), are used to detect and characterize the ways in which the data at hand do not fit the linear model assumptions. The data are then altered, by setting aside outliers and/or transforming the observed X and Y values, and the classical analysis is applied to the altered data. This process is often undertaken iteratively as one of the potentially many *sensitivity analyses* conducted in the overall investigation: the effects of the outlying/influential observations and transformations on the final inferences are examined by first excluding and then including unusual data values, varying the chosen transformations and so on, all the while observing the resulting behavior of predicted values, standard errors and so forth.

2) The *model expansion* approach, in which the data are examined as in approach 1, to characterize the ways in which they depart from the standard off-the-shelf model, the difference being that when departures are found they are modeled directly on the raw data scale through a broadening of the parametric model (Cox, 1977; Weisberg, 1984). The class of *generalized linear models* (McCullagh and Nelder, 1983), with inference based on classical large-sample maximum likelihood theory, is a leading example of this approach from the frequentist perspective, and Box and Tiao (1962) provide an interesting Bayesian example of model expansion.

3) The *robust* approach, which uses nonclassical techniques intended to be sufficiently insensitive to deviations from the classical assumptions that the data may be analyzed on the raw scale without modification or additional (explicit) modeling.

In practice there is a fair amount of overlap among these approaches to inference. Practitioners of strategies 1 and 2 may well differ more in style than in content, and many analysts who have no interest in using robust methods inferentially nevertheless find them useful diagnostically. The idea is to carry out the classical inference and one or more robust procedures in parallel, to see if they produce sharply different answers. If not, report the classical findings, because they may be most readily accepted by consumers of the analysis; if there are major differences, proceed as in approaches 1 or 2. This overlap notwithstanding, it is useful to draw overall strategic distinctions like the ones above to help organize thought about the strengths and weaknesses of various approaches to inference, and such distinctions will figure in the discussion that follows.

The do-nothing method is widely accepted even today. This can be seen, for example, by observing that until quite recently the ANOVA programs in SPSS (Hull and Nie, 1981), BMDP (Dixon, 1983) and Minitab (Ryan, Joiner and Ryan, 1985), three of the most widely used statistical computing packages, did not permit the user to examine in any way the residuals from fitted models, and another such package, SAS (SAS Institute Inc., 1985), perhaps the most extensively used of the four, still does not. (It has long been possible in all of these packages to do residual analysis in ANOVA by re-expressing the problem in regression terms, but many users of programs like SAS do not know how to do this or are not in the routine habit of doing so.) It is my experience that many people applying linear models simply are not aware of any need to consider the assumptions built into Ω^*. Among practitioners of method 0 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^* (Rao, 1973) continue to hold when some of its assumptions do not (a reliance on the reasonable sounding but false principle of *continuity of optimality*—"What is optimal at the model should be nearly optimal near the model"), 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 (Bradley (1978) and Scheffé (1959), for instance) 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; Bickel, 1973; among many others) to show that techniques

exist that perform noticeably better than the classical methods when one departs from some of those basic assumptions.

It is easy to see why the do-nothing strategy persists. It is straightforward to carry out, there is much precedent for it and the findings of off-the-shelf analyses are easy to interpret (at least naively) by the ultimate users (not a point to be taken lightly—in practice there is constant tension between the best choice of inferential procedure and ready interpretability of findings by the end-users of the analysis). But it has been repeatedly demonstrated (for instance, Weisberg, 1985; Cook and Weisberg, 1982; Atkinson, 1985) that when classical methods are applied unquestioningly a few strange values or transposed digits can yield a completely misleading analysis, particularly with modest sample sizes. Tukey (see Mosteller and Tukey (1977), for example) usefully decomposes the potential failings of a piece of inferential machinery into two components: he speaks of robustness of *validity* (as measured, for instance, by standard frequentist criteria like confidence interval coverage probability and type I error rate) and robustness of *efficiency* (effective separation of signal from noise, as indexed by things like confidence interval length and type II error rate). Analysts adopting the do-nothing approach are often subject to penalties in efficiency *and* validity relative to the other three strategies listed above. Examples of efficiency losses of this type will be given in Section 2, and examples of validity difficulties (effects of dependence and heteroscedasticity on type I error rates of the usual F tests, for instance) can be found in Scheffé (1959, Chapter 10) and Box (1953, 1954a, b).

Each of the data-analytic, model expansion and robust approaches possesses strengths and weaknesses as inferential strategies generally and in particular in the linear model. Data-analytic techniques based on transformations can yield quite efficient inference on the transformed scale (Box and Cox, 1964), even after paying the appropriate price for using the data to help determine the transformation (Carroll and Ruppert 1981, 1984). Moreover, this inference will be in the context of the familiar classical methods applied to the transformed data, which aids in interpretation of results; but interpretation is simultaneously made more difficult by having strayed away from the raw scale (for example, in situations where the original scale has direct substantive meaning and effect summaries like (additive) pairwise comparisons are desired—see Tukey, 1977). The model expansion approach can also lead to quite efficient inference (McCullagh and Nelder, 1983) and has the advantage that this inference takes place on the raw scale by construction; but with modest amounts of data the accurate specification of the expanded model can be difficult, leading to problems with stability of inference. With samples of small and moderate size, two models that are close enough to be recognized by the data as about equally plausible can result in sharply different inferences, with standard errors differing by factors of 50% or more (Tukey (1960), in the paper that might be said to have kindled modern interest in robustness). Inference based on robust methods, *when available*, can be successful on all the above grounds—efficiency, raw scale interpretability and stability—but these goals have not yet been fully attained in the general linear model. Despite the all-too-widely-held view among practitioners that phrases like "robust" and "nonparametric" are proxies for "assumption-free inference," the word *robust* is just a shorthand for "insensitive with respect to departures from the following underlying assumptions $\cdots$," and the truth is that most existing "robust" linear models methods only address failures of the explicit distributional assumptions in Ω^* (symmetry, normality). This is certainly not because violations of these assumptions are the most critical (they are not—departures from the homogeneity of variance and independence assumptions typically have more serious consequences: see Scheffé, 1959), but stems rather from the fact that this has been the most analytically tractable area in which to make initial progress. Recent robustness work on less tractable but more important issues, for instance:

- Ruppert and Carroll (1982) on dealing with heteroscedasticity,
- Portnoy (1977, 1979) on coping with dependence in the data, and
- many authors (for example, Cleveland, 1979; Stone, 1977; Hastie and Tibshirani, 1986) on nonparametric regression (which attempts to relax the assumption of known functional form for the relationship (1.2) between Y and the X_i)

is promising, but significant effort remains to be expended toward the practical implementation of inferential methods which are robust with respect to violation of the non- (symmetry, normality) assumptions in Ω^*.

Among statisticians who see the need for methods to deal with violations of the assumptions in Ω^*, use of data-analytic and model expansion techniques has been far more prevalent to 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 criteria of ready usability by the final consumers of the linear model analysis. Such criteria might reasonably require that the techniques used:

- should have clear intuitive appeal;
- should be of a unified nature and of general

applicability rather than being put together in patchwork form out of solutions to separate but related problems; and
- should be based on inferential machinery (estimates, tests and so on) possessing 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 article that for some robust linear model techniques these criticisms are no longer valid.

Progress to date in the development of 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 constructs that lead to the classical methods and replacing them at key points with devices that make the resulting techniques more robust. This work has proceeded roughly along three parallel lines: the generalized maximum likelihood type or M-methods (Huber, 1973; Schrader and Hettmansperger, 1980; Sen, 1982); the rank-based or R-methods (Lehmann, 1963b; Jaeckel, 1972; Hettmansperger and McKean, 1977; Hettmansperger, 1984) and methods based on linear combinations of order statistics, the L-methods (Bickel, 1973; Koenker and Bassett, 1978). Each of these approaches has advantages and disadvantages in robustness, efficiency, applicability and usability, depending on the situation. None of them clearly dominates the others in efficiency and robustness (Huber, 1981), but with respect to practical considerations in implementation, like the ones listed above, there are some differences. The L-methods have historically been the most awkward of the three in generalizing to linear models (Huber, 1981; although recent work on regression quantiles and trimmed-mean analogues by such authors as Bassett and Koenker (1982), Ruppert and Carroll (1980) and Welsh (1987a) may eventually change this appraisal), and, although 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 whereas the M-methods do not. In the rest of this article, attention will be restricted to methods based on R-estimators.

Thirty years ago the phrase "rank-based analysis of linear models" would have conjured up images of methods like the Wilcoxon (1945) rank-sum test in the two-independent samples problem, the Kruskal-Wallis (1952) test in the one-way layout, Spearman's (1904) rank correlation coefficient, the Friedman (1937) procedure for analyzing randomized complete block designs and so on. There are two main drawbacks to the approach to linear models analysis represented by these techniques:

- Most methods of this type simply amount to transforming the raw data to the rank scale and feeding the ranks into the classical normal theory procedures (Conover, 1980). For example, the Wilcoxon rank-sum statistic is in disguise just a monotone function of the usual (pooled-variance) two-independent samples t statistic applied not to the original data but to the ranks of the original observations in a data set formed by combining the two samples into one. These methods can thus be seen as a special case of the data-analytic approach in which the rank transformation is used exclusively. This means that in addition to having the general data-analytic disadvantage of loss of raw scale interpretability, these rank scale methods have the further disadvantage of providing no flexibility in the choice of transformation.
- These rank methods also have the appearance of failing the second implementational criterion above (the desirability of a unified set of techniques rather than a patchwork quilt). To potential users watching the rank-based literature evolve, the resulting package of techniques has had a distinctly piecemeal flavor: with one type of linear model data the method of choice is Friedman, with another it is Kruskal-Wallis and so on. In fact, because all these methods are based on the rank transformation, there is more methodological coherence than appearances might indicate, but the methods still lack complete unity in that they all have different null distributions and different versions of the ranking operation.

The R-estimator methods to be described in this article represent an improvement over these previous rank methods in linear models on both of the above counts: they present a unified approach to the analysis rather than offering a collection of separate but related procedures, and the estimates and quantities from which the test statistics are constructed are on the same cardinal measurement scale as the original data. The basic distinction is that the R-estimator methods are not *rank* methods but *rank-based* methods—they are robust procedures whose motivation *stems* from rank methods.

Two R-estimator approaches to linear model analysis are described in Sections 2 and 3: the analysis of variance techniques of Lehmann (1963b) and the general linear model methods of Hettmansperger and McKean (1976, 1977). In both approaches the analogy with classical methods is quite strong. The results include robust estimates of functions of the parameters β_j and standard errors for those estimates, and a robust version of the analysis of variance table, complete with R-estimator 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 to 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 and McKean approach was originally presented as requiring the assumption of symmetry of the underlying error distribution, and more recent attempts (Aubuchon and Hettmansperger, 1984a,b; Sheather and Hettmansperger, 1985) to relax this assumption, although promising, have not yet been fully validated in the general linear model with sufficiently comprehensive simulation studies. The Lehmann method only applies to ANOVA situations with several observations per cell. Moreover, both approaches to inference yield procedures that are distribution-free only asymptotically, so that a check on their small-sample behavior is needed. This article describes results pertaining to two issues—the implementation of the Lehmann and Hettmansperger-McKean significance testing techniques in a way which dispenses when possible with the assumption of symmetry, and the small-sample empirical properties of these techniques (Section 4). Multiple comparison methods and confidence procedures are addressed only briefly here, in Sections 2 and 3; for further examples of the rank-based robust approach to these inferential tools see Hettmansperger and McKean (1978). Although the emphasis below is on fixed-effects models, several of the methods described would be expected to work well in certain random effects and mixed models also; see Lehmann (1963b).

There is a broad literature on rank transformation and other rank-based approaches to linear model analysis which will not be addressed further here; relevant work includes Adichie (1978), Koul (1969, 1970), Puri and Sen (1969, 1973, 1977), Srivastava (1972) and references cited therein.

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

$$(2.1)\qquad Y_{ij} = \mu_i + e_{ij}, \quad \left\{ \begin{matrix} i = 1, \cdots, I \\ j = 1, \cdots, n_i \end{matrix} \quad \sum_{i=1}^{I} n_i = N \right\},$$

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

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

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 (Lehmann, 1975). The finiteness of θ and σ^2 are needed because division by $1/\theta$ and σ^2 plays a role in what follows; these conditions place little or no 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.

The genesis of Lehmann's method was as follows. Hodges and Lehmann showed in 1963 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 at about the same time Lehmann began looking for a way to apply these methods to other linear models. For widest applicability of the results it was preferable to adapt the two-sample version of the Hodges-Lehmann estimation technique, because there is no assumption of symmetry implicit in its derivation (as there is in that of the

one-sample estimates), so this suggested trying to estimate $\mu_i - \mu_j$ 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, because most inference in ANOVA (linear hypothesis testing, multiple comparisons and so on) is based on contrasts in the cells means, and any contrast

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

is expressible in terms of differences in the cell means:

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

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 $\binom{I}{2}$ two-sample shift models embedded within it. (With a different approach Lehmann (1963a, 1964) also extended the Hodges-Lehmann one-sample estimates to linear models with several observations per cell and developed rank-based methods for 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 estimate of $\mu_i - \mu_j$,

$$T_{ij} = \text{med}\{Y_{ik} - Y_{jl}: \quad k = 1, \cdots, n_i; \; l = 1, \cdots, n_j\}, \tag{2.6}$$

the median of the set of all pairwise differences among the observations in cells i and j, was unsatisfactory, because the T_{ij} do not satisfy the linearity constraints which the $\mu_i - \mu_j$ themselves do:

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

but

$$T_{ij} + T_{jk} \neq T_{ik}, \tag{2.8}$$

because the operations of subtraction and taking a median do not commute. This makes the raw Hodges-Lehmann estimates (2.6) unsuitable as a basis for tests of linear hypotheses about the μ_i, because in small samples an arbitrary renumbering of the cells would yield somewhat different estimates and test statistics. Lehmann proposed instead adjusting the T_{ij} and estimating $\mu_i - \mu_j$ by

$$W'_{ij} = \bar{T}'_i - \bar{T}'_j, \tag{2.9}$$

where

$$\bar{T}'_i = I^{-1} \sum_{k=1}^{I} T_{ik}. \tag{2.10}$$

(Note that $T_{ii} = 0$ for all i.) The linearity problem was thus removed, at the cost of offending (frequentist) 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 $\bar{T}'_i$, which is inefficient with unbalanced data. Spjøtvoll suggested several ways of remedying the situation, the simplest of which was to work with weighted averages:

$$W^*_{ij} = \bar{T}_i - \bar{T}_j, \quad \bar{T}_i = N^{-1} \cdot \sum_{k=1}^{I} n_k T_{ik}. \tag{2.11}$$

This is the form of Hodges-Lehmann estimation that is used in what follows. Note that $\bar{T}_i$ 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, because $T_{ji} = -T_{ij}$,

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

In effect, the Lehmann-Spjøtvoll adjustment process linearizes the estimates by making all comparisons relative to the grand mean: $\mu_i - \mu_j$ is expressed as $(\mu_i - \bar{\mu}) - (\mu_j - \bar{\mu})$ and estimated by $\bar{T}_i - \bar{T}_j$.

A natural estimate of the contrast (2.4) is then

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

It seems on the face of it that the resulting estimate will not be unique, because the b_{ij} are not; but in fact 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 express the classical test statistics in terms of contrasts and to replace the classical estimates of those contrasts 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 $(\mu_1, \cdots, \mu_I)$ 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

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

where

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

and $(\hat{\mu}_1, \cdots, \hat{\mu}_I)$ 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, follows a χ^2 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/σ^2 converges in distribution under H to χ_q^2 as $n_1, \ldots, n_I \to \infty$ even if f is not normal. If σ^2 is not known, it is 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 the testing ratio, the asymptotic distribution will still be χ_q^2. The classical estimate of σ^2 is

$$\hat{\sigma}^2 = (N - I)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\cdot})^2, \tag{2.17}$$

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

$$F_C = \frac{(C/\sigma^2)/q}{[(N - I)\hat{\sigma}^2/\sigma^2]/(N - I)} = \frac{C/q}{\hat{\sigma}^2} \tag{2.18}$$

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

$$C = \sum_{i=1}^{I-1} n_i \left[\sum_{j=i+1}^{I} a_{ij}(Y_{i\cdot} - Y_{j\cdot}) \right]^2, \tag{2.19}$$

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

$$L = \sum_{i=1}^{I-1} n_i \left(\sum_{j=i+1}^{I} a_{ij} W_{ij} \right)^2. \tag{2.20}$$

Another way to put it, considering (2.12), is that $\bar{T}_i$ 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

$$C = \sum_{i=1}^{I-1} n_i \left\{ \sum_{j=i+1}^{I} a_{ij}[(Y_{i\cdot} - \bar{Y}) - (Y_{j\cdot} - \bar{Y})] \right\}^2 \tag{2.21}$$

is

$$L = \sum_{i=1}^{I-1} n_i \left[\sum_{j=i+1}^{I} a_{ij}(\bar{T}_i - \bar{T}_j) \right]^2. \tag{2.22}$$

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 $\bar{T}_i$ (or, equivalently and even more simply, in view of (2.13) one can replace just $Y_{i\cdot}$ by $\bar{T}_i$).

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

$$H_A\colon \mu_1 = \cdots = \mu_I \tag{2.23}$$

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

$$C_A = \sum_{i=1}^{I} n_i(Y_{i\cdot} - \bar{Y})^2, \tag{2.24}$$

so the robust numerator is simply

$$L_A = \sum_{i=1}^{I} n_i \bar{T}_i^2. \tag{2.25}$$

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

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \begin{Bmatrix} i = 1, \cdots, r \\ j = 1, \cdots, c \\ k = 1, \cdots, n \end{Bmatrix}, \tag{2.26}$$

subject to the side conditions

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{c} \beta_j = \sum_{i=1}^{r} \gamma_{il} = \sum_{j=1}^{c} \gamma_{mj} = 0 \tag{2.27}$$

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

$$\begin{aligned} H_A&\colon \alpha_1 = \cdots = \alpha_r = 0, \\ H_B&\colon \beta_1 = \cdots = \beta_c = 0, \\ H_{AB}&\colon \gamma_{11} = \cdots = \gamma_{rc} = 0. \end{aligned} \tag{2.28}$$

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

$$C_A = nc \sum_{i=1}^{r} (Y_{i..} - Y_{...})^2,$$

$$C_B = nr \sum_{j=1}^{c} (Y_{.j.} - Y_{...})^2, \tag{2.29}$$

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

where as in (2.16) the dot notation indicates averaging over the indicated subscript(s). Here $Y_{ij.}$ and $Y_{...}$ play the roles of $Y_{i.}$ and $\bar{Y}$ in (2.24), and $Y_{i..} - Y_{...}$ is given by

$$Y_{i..} - Y_{...} = c^{-1} \sum_{j=1}^{c} (Y_{ij.} - Y_{...}) \tag{2.30}$$

and similarly for $Y_{.j.} - Y_{...}$, so numbering the $I = rc$ cells as in Table 1 the robust numerators, in the notation of model (2.1), come out

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

$$L_B = nr \sum_{j=1}^{c} \left(r^{-1} \sum_{i=1}^{r} \bar{T}_{j+(i-1)c} \right)^2, \tag{2.31}$$

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

This approach of replacing classical estimates by their robust counterparts in the numerators of the classical test statistics to obtain the robust numerators works 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 $Y_{i.}$ (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 $Y_{i.}$ by the $\bar{T}_i$ in the system of equations whose solution determines the classical numerator and proceeding as in the classical case.

TABLE 1
Notational conversion of the two-way layout into a one-way layout

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

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

$$L_A = \sum_{i=1}^{r} \left[\frac{(\sum_{j=1}^{c} \bar{T}_{j+(i-1)c})^2}{\sum_{j=1}^{c} n_{j+(i-1)c}^{-1}} \right] - \frac{[\sum_{i=1}^{r} (\sum_{j=1}^{c} n_{j+(i-1)c}^{-1})^{-1} \sum_{j=1}^{c} \bar{T}_{j+(i-1)c}]^2}{\sum_{i=1}^{r} (\sum_{j=1}^{c} n_{j+(i-1)c}^{-1})^{-1}}, \tag{2.32}$$

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 (1963b) showed that L/σ_R^2 converges in distribution under H to χ_q^2 as the $n_i \to \infty$, where

$$\sigma_R^2 = (12\theta^2)^{-1} = [12(\int f^2)^2]^{-1} \tag{2.33}$$

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

$$e_{L,C}(f) = \sigma^2/\sigma_R^2 = 12\sigma^2(\int f^2)^2 \tag{2.34}$$

is the asymptotic relative efficiency of the robust procedure to the classical; this is just the familiar expression (Lehmann, 1975) for the efficiency of the Wilcoxon one- and two-sample procedures relative to the corresponding classical t methods. Table 2 gives some values of this efficiency as well as values of σ_R^2 and $1/\theta$ for various distributions. The skewed mixed normal distribution referred to in Table 2 has cumulative distribution function (cdf)

$$F(x) = \lambda\Phi[(x - \mu_1)/\sigma_1] + (1 - \lambda)\Phi[(x - \mu_2)/\sigma_2], \tag{2.35}$$

where Φ is the standard normal cdf.

It can be seen from the table 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 asymptotic results, but, as documented in Draper (1981), empirical small-sample efficiencies are similar to the asymptotic values in designs with as few as $N = 10$ total observations and $n_i = 3$ observations per cell.

Just as in the classical case, σ_R^2, the denominator of the Lehmann test statistic, will not be known in

TABLE 2
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	$e = \sigma^2/\sigma_R^2$
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				
($\lambda = 0.75$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 1.9$, $\sigma_2 = 2$)	2.426	4.670	1.817	1.335
($\lambda = 0.82$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 1.9$, $\sigma_2 = 3.5$)	3.558	4.535	1.714	2.076
t with 3 degrees of freedom	3.0	4.353	1.579	1.900
Any f				≥ 0.864

practice and it is necessary to estimate it to obtain usable test statistics. 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

$$L/\hat{\sigma}_R^2 \tag{2.36}$$

from this χ_q^2 distribution, but, as is discussed below in Section 4 and shown in Draper (1981), the distribution of

$$\nu\hat{\sigma}_R^2/\sigma_R^2 \tag{2.37}$$

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

$$F_L = \frac{(L/\sigma_R^2)/q}{(\nu\hat{\sigma}_R^2/\sigma_R^2)/\nu} = \frac{L/q}{\hat{\sigma}_R^2} \tag{2.38}$$

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

Rank-based robust confidence and multiple comparison procedures using the Lehmann approach are straightforward to construct through the same analogies to the classical techniques that gave rise to Lehmann's robust testing ratios. The method (Lehmann, 1963b) consists simply of writing down the classical confidence and multiple comparisons region of interest, substituting the robust estimates $\overline{T}_i$ for the classical $Y_{i\cdot}$, and using an estimate $\hat{\sigma}_R^2$ for the rank analogue of the underlying error variance in place of the usual $\hat{\sigma}^2$ (2.17). As a simple illustration, continuing Example 1 above, the normal theory $100(1 - \alpha)\%$ confidence interval for a contrast $\phi = \sum_{i=1}^I c_i\mu_i = \sum_{i=1}^{I-1}\sum_{j=i+1}^I b_{ij}(\mu_i - \mu_j)$ in the cell means in the one-way layout has the form

$$\hat{\phi}_c \pm t_{N-I}^{-1}(1 - \alpha/2)\hat{\sigma}\left(\sum_{i=1}^I c_i^2/n_i\right)^{1/2}, \tag{2.39}$$

where $\hat{\phi}_c = \sum_{i=1}^{I-1}\sum_{j=i+1}^I b_{ij}(Y_{i\cdot} - Y_{j\cdot})$ is the classical contrast estimate and t_ν is the cdf of the t distribution with ν degrees of freedom. The Lehmann-type rank-based robust alternative to this is simply

$$\hat{\phi} \pm t_{N-I}^{-1}(1 - \alpha/2)\hat{\sigma}_R\left(\sum_{i=1}^I c_i^2/n_i\right)^{1/2}, \tag{2.40}$$

with $\hat{\phi}$ given by (2.14), and using one of the estimates of σ_R^2 provided in Section 4. Scheffé- and Tukey-type multiple comparisons regions are constructed similarly. Lehmann (1963b) demonstrates the large sample validity of this technique and gives an example of its use.

A FORTRAN 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 David Draper at The RAND Corporation.

3. THE TECHNIQUES OF JAECKEL AND HETTMANSPERGER-McKEAN: THE 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 and confidence procedures, is the general fixed effects linear model (Ω) of Section 1,

$$Y_i = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j + e_i, \quad i = 1, \cdots, N, \tag{3.1}$$

or, in matrix form,

$$\mathbf{Y} = \boldsymbol{\beta} + X\boldsymbol{\beta} + \mathbf{e}, \tag{3.2}$$

in which as in model (2.1) the e_i are iid continuous random variables with density f such that both $\theta = \int f^2$ and $\sigma^2 = \mathrm{Var}(e_i)$ are finite, and X is an N by p matrix of known constants. Jaeckel's work simplifies and makes more usable an approach to robust estimation in the linear model due to Jurečková (1971), who generalized the work of Hodges and Lehmann (1963) described above on inverting rank tests to

obtain estimates. Jaeckel's starting point, like Lehmann's in Section 2, is 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 $\boldsymbol{\beta}' = (\beta_0, \cdots, \beta_p)$,

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

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,

$$\begin{aligned} D_C[\mathbf{e}(\boldsymbol{\beta}')] &= \| \mathbf{e}(\boldsymbol{\beta}') \|^2 = \sum_{i=1}^{N} e_i^2(\boldsymbol{\beta}') \\ &= \sum_{i=1}^{N} e_{(i)}(\boldsymbol{\beta}') \cdot e_{(i)}(\boldsymbol{\beta}'), \end{aligned} \tag{3.4}$$

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

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

where $\bar{\mathbf{z}}$ 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, $\| \mathbf{e} - \bar{\mathbf{e}} \|^2 = N \operatorname{Var}(\mathbf{e})$. This yields the classical estimates $\hat{\beta}_j$, not of all of the $\boldsymbol{\beta}$ but only of $(\beta_1, \cdots, \beta_p)$, because this dispersion measure is translation-invariant and β_0 drops out. Then the centering part $\| \bar{\mathbf{e}} \|^2$ of (3.5) with the previously found $(\hat{\beta}_1, \cdots, \hat{\beta}_p)$ 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 $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is estimated by $\hat{\boldsymbol{\beta}}$; and finally, β_0 is estimated as the center of the residuals $Y_i - (X\hat{\boldsymbol{\beta}})_i$. Note that in the usual method of arriving at the classical estimates, in which $\| \mathbf{e} \|^2$ is minimized as a function of β_0 as well as of $(\beta_1, \cdots, \beta_p)$, the minimizing condition which specifies $\hat{\beta}_0$ in terms of $(\hat{\beta}_1, \cdots, \hat{\beta}_p)$ is $\bar{\mathbf{e}} = 0$, 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.

From a robustness point of view, the trouble with the usual normal theory 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:

$$\begin{aligned} D_J[\mathbf{e}(\boldsymbol{\beta}')] &= \sum_{i=1}^{N} a(i) e_{(i)}(\boldsymbol{\beta}') \\ &= \sum_{i=1}^{N} a[R_i'(\boldsymbol{\beta}')] e_i(\boldsymbol{\beta}), \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} D_J[\mathbf{Y} - \beta_0 - X\boldsymbol{\beta}] &= D_J[\mathbf{Y} - X\boldsymbol{\beta}] \\ &= \sum_{i=1}^{N} a[R_i(\boldsymbol{\beta})][Y_i - (X\boldsymbol{\beta})_i], \end{aligned} \tag{3.7}$$

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

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

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 $(\beta_1, \cdots, \beta_p)$). See Koul, Sievers and McKean (1987) for an interesting numerical study of the performance of the Jaeckel estimation procedure with asymmetric scores.

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

$$a_s(i) = \operatorname{sign}[i/(N + 1) - \tfrac{1}{2}] \tag{3.10}$$

and the linear Wilcoxon scores

$$a_W'(i) = i/(N + 1) - \tfrac{1}{2}. \tag{3.11}$$

Other possibilities include van der Waerden-type normal scores,

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

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, which lead to results similar to those from least absolute deviation (L^1) regression (Laplace (1793) (see Stigler, 1986); Kennedy and Gentle, 1980), 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 the Hettmansperger-McKean 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

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

The resulting Wilcoxon-type dispersion measure is

$$\begin{aligned} D_{\mathrm{JW}}[\mathbf{Y} - X\boldsymbol{\beta}] &= 12^{1/2}(N + 1)^{-1} \\ &\quad \cdot \sum_{i=1}^{N} [R_i(\boldsymbol{\beta}) - (N + 1)/2][Y_i - (X\boldsymbol{\beta})_i]. \end{aligned} \tag{3.14}$$

Solving for the $\hat{\boldsymbol{\beta}}_{\mathrm{JW}}$ which minimizes $D_{\mathrm{JW}}[\mathbf{Y} - X\boldsymbol{\beta}]$ yields the rank analogue of the normal equations:

$$\sum_{i=1}^{N} (X_{ij} - \mathbf{X}_j)[R_i(\hat{\boldsymbol{\beta}}_{\mathrm{JW}}) - (N + 1)/2] \doteq 0, \quad j = 1, \cdots, p, \tag{3.15}$$

where

$$\mathbf{X}_j = N^{-1} \sum_{i=1}^{N} X_{ij}. \tag{3.16}$$

The "equations" (3.15) are solved in the sense that a value of $\hat{\boldsymbol{\beta}}_{\mathrm{JW}}$ is sought which makes the lefthand side of (3.15) as a 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. A potentially more serious practical drawback is that 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 $D_{\mathrm{JW}}[\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{\mathrm{JW}}]$, as defined by (3.14), over all $\hat{\beta}_{\mathrm{JW}}$ in the solution set to (3.15). 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 enough. Note however that, due to differences in computer hardware, the same computer program to perform the iterative search for $\hat{\beta}_{\mathrm{JW}}$ may yield somewhat different estimates when run on different computers.

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

$$Y_i = \alpha + \beta X_i + e_i, \quad i = 1, \cdots, N, \tag{3.17}$$

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 that implicitly defines Adichie's estimate) is a weighted median of the set of all pairwise slopes

$$\{(Y_j - Y_i)/(x_j - x_i), (i, j) \text{ such that } x_i \neq x_j\} \tag{3.18}$$

in which the weights are proportional to the absolute distance $|x_i - x_j|$ between the independent variable values. (To calculate a weighted median, sort the observations from smallest to largest, carrying their weights along with them, find the overall sum S of the weights, and begin adding the weights from the top or bottom of the sorted list until $S/2$ is reached. The corresponding observation is the weighted median.) This model is examined in more detail in Draper (1981), 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 $\hat{e}_i^* = Y_i - (X\hat{\boldsymbol{\beta}}_J)_i$. 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 $Y_i - (X\boldsymbol{\beta})_i$. 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,

$$\hat{\beta}_0 = \text{med}\{\hat{e}_i^* + \hat{e}_j^*)/2,\ 1 \le i \le j \le N\}, \tag{3.19}$$

the median of the set of all pairwise averages of the residuals $Y_i - (X\hat{\boldsymbol{\beta}}_J)_i$. This is Hettmansperger and McKean's recommended estimate of β_0. In settings where the assumption of symmetry is not tenable, Aubuchon and Hettmansperger (1984b) advocate the use of the ordinary median of the residuals, $\hat{\beta}_0' = \text{med}\{\hat{e}_i^*\}$, although in the absence of symmetry the appropriate measure of residual centering by which the intercept is to be defined may vary from problem to problem.

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 $\boldsymbol{\beta}$ vector, it is convenient to parameterize the model Ω (3.1) in such a way that the design matrix X has full rank p. Denote by ω this model plus the restrictions imposed by H. Expressed in terms of the classical residual dispersion measure, the classical test statistic for H is based on

$$D_C^* = D_C[\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{\omega,C}] - D_C[\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{\Omega,C}], \tag{3.20}$$

the amount of extra lack of fit imposed by accepting the model ω over and above that inherent in Ω. Here $\hat{\boldsymbol{\beta}}_{\omega,C}$ and $\hat{\boldsymbol{\beta}}_{\Omega,C}$ 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_q^2 under H when f is normal, and under the same mild regularity conditions as in Section 2 is asymptotically χ_q^2 even if f is not normal. As before it is typically necessary to estimate σ^2; the classical estimate is

$$\hat{\sigma}^2 = [N - (p+1)]^{-1} \cdot \sum_{i=1}^{N} [Y_i - (X\hat{\boldsymbol{\beta}}_{\Omega,C})_i - \hat{\beta}_{0,C}]^2, \tag{3.21}$$

where

$$\hat{\beta}_{0,C} = N^{-1} \sum_{i=1}^{N} [Y_i - (X\hat{\boldsymbol{\beta}}_{\Omega,C})_i] \tag{3.22}$$

is the sample mean of the full model residuals $\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{\Omega,C}$.

The asymptotic null distribution of $D_C^*/\hat{\sigma}^2$ is still χ_q^2, but because $[N - (p+1)]\hat{\sigma}^2/\sigma^2 \sim \chi_{N-(p+1)}^2$ and D_C^* and $\hat{\sigma}^2$ are independent under normality, a better (Scheffé, 1959) small-sample distribution when f is normal for

$$F_C = \frac{(D_C^*/\sigma^2)/q}{\{[N - (p+1)]\hat{\sigma}^2/\sigma^2\}/[N - (p+1)]} = \frac{D_C^*/q}{\hat{\sigma}^2} \tag{3.23}$$

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

$$D_{\text{JW}}^* = D_{\text{JW}}[\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{\omega,\text{JW}}] - D_{\text{JW}}[\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{\Omega,\text{JW}}], \tag{3.24}$$

where as above $\hat{\boldsymbol{\beta}}_{\omega,\text{JW}}$ and $\hat{\boldsymbol{\beta}}_{\Omega,\text{JW}}$ are the parameter estimates under ω and Ω, respectively, that minimize the Jaeckel dispersion measure. Hettmansperger and McKean found that $2D_{\text{JW}}^*/\sigma_R$ converges in distribution under H to χ_q^2, where as before

$$\sigma_R = 12^{-1/2}\theta^{-1} = 12^{-1/2}(\textstyle\int f^2)^{-1}. \tag{3.25}$$

Note that, unlike in the Lehmann method, θ enters into the asymptotic distribution of the test statistic through $1/\theta$ rather than through $1/\theta^2$. (Intuitively this is because substitution of the scores $a(i)$ into the classical dispersion measure (3.4) causes the residuals to enter into the Jaeckel dispersion measure (3.6) raised only to the first power.) 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 in efficiency terms.

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

$$2D_{\text{JW}}^*/\hat{\sigma}_R \tag{3.26}$$

is still χ_q^2, and this was the distribution originally proposed in practice by Hettmansperger and McKean. They later (1977) found that the χ_q^2 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 $2D_{\text{JW}}^*/\hat{\sigma}_R$, they suggested, without much justification except by analogy with the classical methods, the approximation of the null distribution of

$$(2D_{\text{JW}}^*/q)/\hat{\sigma}_R \tag{3.27}$$

by the $F_{q,N-(p+1)}$ distribution. The success of this and other approximations is discussed in Draper (1981) and in Section 4 below.

As was the case with the Lehmann methods, confidence and multiple comparison procedures using the Hettmansperger-McKean approach derive naturally from the classical techniques, through the replacement of the normal theory parameter estimates $\hat{\boldsymbol{\beta}}_{\Omega,C}$ and estimated error variance $\hat{\sigma}^2$ (3.21) by their robust analogues $\hat{\boldsymbol{\beta}}_{\Omega,\mathrm{JW}}$ and $\hat{\sigma}_R^2$. For an informal argument justifying this and an example illustrating the robust versions of Scheffé and Tukey multiple comparison calculations in a two-way layout with several observations per cell, see Hettmansperger and McKean (1978). Extensive empirical investigation of the small sample properties of both the Hettmansperger-McKean confidence and multiple comparisons procedures and of those of Lehmann outlined in the previous section has not yet been made, but in any given situation one would expect acceptable coverage behavior and interval lengths that compare favorably with those of the classical procedures, based on the good performance (described in the next section) of the testing techniques and the usual connections between significance tests and confidence intervals.

A rank regression command (RREG) implementing the Hettmansperger-McKean approach to robust linear model analysis is now available in some versions of the Minitab statistical computing system (Ryan, Joiner and Ryan 1985), and there are plans to make it available in all versions of Minitab in the next year or two. Alternatively, a stand alone FORTRAN implementation of these methods can be obtained from Joseph McKean at Western Michigan University.

4. ESTIMATION OF σ_R^2; EMPIRICAL RESULTS

The following is a brief description of results pertaining to the estimation of σ_R^2 and to the empirical small sample performance of the rank-based robust testing ratios of Lehmann and Hettmansperger-McKean outlined in Sections 2 and 3 above. For more details see Antille (1976), Draper (1981), Hettmansperger (1984), Jurečková (1973) and the other references cited below.

4.1 Estimation of σ_R^2

Two main approaches to the estimation in linear models of $\sigma_R^2 = 12^{-1}\theta^{-2}$, where $\theta = \int f^2$, have so far been examined in the literature: 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 Schuster (1974) and Schweder (1975, 1981) based on window- or kernel-type density estimation of f. The two methods appear quite different but are in fact closely related—in disguise the Lehmann method is equivalent to a particular version of a density estimate (Draper, 1981; Aubuchon and Hettmansperger, 1984a)—and they have been shown (Draper, 1981) to be asymptotically equally accurate in the estimation of σ_R^2, but (as will be seen below) there are modest differences in small sample performance and substantial differences in ease of implementation. Other approaches to the estimation of σ_R^2 not described here can be found in Antille (1974), Lehmann (1963b), Koul, Sievers and McKean (1987) and Sheather (1985). Data resampling methods such as the bootstrap and jackknife provide a natural alternative to the methods presented here for estimating σ_R^2 and obtaining p-values from the Lehmann and Hettmansperger-McKean testing procedures, but this possibility has not been explicitly investigated to date.

4.1.1 Lehmann's Method Based on Length of Confidence Intervals

As was the case with Lehmann's robust contrast estimates of Section 2, his idea (Lehmann 1963c) for estimating σ_R^2 arose out of his work with Hodges (1963) on inverting rank tests to obtain estimates of the center of symmetry in the symmetric one-sample model and of the size of the shift in the two-sample shift model. Consider first the usual symmetric one-sample model, in which, say, $Z_1, \cdots, Z_N$ are iid with continuous density f, symmetric about μ. The rank-based one-sample Hodges-Lehmann estimate of μ is obtained by inverting the Wilcoxon signed-rank test, and takes the form

$$\hat{\mu} = \mathrm{med}(S'), \qquad S' = \{(Z_i + Z_j)/2\colon 1 \le i \le j \le N\}, \tag{4.1}$$

the median of the set of all pairwise averages of the observations. Lehmann showed that this set S' of all pairwise averages could serve as the basis of a family of confidence intervals for μ whose confidence levels depend only on the Wilcoxon signed-rank null distribution and not on f, and he then showed how the lengths of these intervals could be used in the estimation of σ_R^2.

Specifically, let $A_{(i)}$ be the ordered elements of S' for $i = 1, \cdots, K \equiv N(N+1)/2$; then for $c'_\alpha = 1, \cdots, [K/2]$

$$(A_{(c'_\alpha)}, A_{(K+1-c'_\alpha)}) \tag{4.2}$$

is a confidence interval for μ, symmetric in the $A_{(i)}$, with confidence level $1 - \alpha = 1 - 2P_0(V \le c'_\alpha - 1)$, where $P_0(V)$ denotes the null distribution of the Wilcoxon signed-rank statistic V and $[x]$ is the integer part of x. Lehmann's result on estimating σ_R^2 was that, as is intuitively reasonable, the lengths $L'(\alpha) = A_{(K+1-c'_\alpha)} - A_{(c'_\alpha)}$ of these distribution-free confidence intervals, when properly normalized, can be used to

estimate the variability of observations with density f. He demonstrated that as $N \to \infty$

$$(4.3) \qquad (3N)^{1/2}L'(\alpha)/\Phi^{-1}(1 - \alpha/2) \xrightarrow{P} 1/\theta$$

for all levels $0 < \alpha < 1$, where Φ^{-1} is the usual inverse normal cdf. Thus,

$$(4.4) \qquad [N^{1/2}L'(\alpha)/2\Phi^{-1}(1 - \alpha/2)]^2$$

is a consistent estimate of σ_R^2 for each α.

Consider now the usual two-sample shift model, in which, say, $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ are independent, the X_i with continuous density $f(x)$ and the Y_j having density $f(y - \Delta)$, so that stochastically $Y =_{st} X + \Delta$. The two-sample Hodges-Lehmann estimate of Δ is obtained by inverting the Wilcoxon rank-sum test and has the form

$$(4.5) \qquad \hat{\Delta} = \text{med}(S), \quad S = \{Y_j - X_i: 1 \le i \le m, 1 \le j \le n\},$$

the median of the set of all pairwise differences of the Y's and X's; this estimator was the basis of Lehmann's rank-based linear model inference described in Section 2. As in the one-sample model, Lehmann showed that this set S of all pairwise differences yields a family of confidence intervals for Δ whose confidence levels are again independent of f, depending in this case only on the Wilcoxon rank-sum null distribution.

More precisely, let $D_{(i)}$ be the ordered elements of S for $i = 1, \cdots, mn$; then for $c_\alpha = 1, \cdots, [mn/2]$

$$(4.6) \qquad (D_{(c_\alpha)}, D_{(mn+1-c_\alpha)})$$

is a confidence interval for Δ, symmetric in the $D_{(i)}$, with confidence level $1 - \alpha = 1 - 2P_0(W \le c_\alpha - 1)$, where $P_0(W)$ is the null distribution of W, the Mann-Whitney form of the Wilcoxon rank-sum statistic. As in the one-sample case, Lehmann showed that the lengths $L(\alpha) = D_{(mn+1-c_\alpha)} - D_{(c_\alpha)}$ of the intervals (4.6), when suitably normalized, can provide an estimate of σ_R^2: His result was that as both m and $n \to \infty$ so that $m/(m + n) \to \rho$, $0 < \rho < 1$,

$$(4.7) \qquad [3mn/(m + n)]^{1/2}L(\alpha)/\Phi^{-1}(1 - \alpha/2) \xrightarrow{P} 1/\theta$$

for all $0 < \alpha < 1$. Thus, in this case for each α

$$(4.8) \qquad \{[mn/(m + n)]^{1/2}L(\alpha)/2\Phi^{-1}(1 - \alpha/2)\}^2$$

is a consistent estimate of σ_R^2.

There are a number of ways to apply these basic results in the linear model. In models with several observations per cell like (2.1), the one-sample method can be applied to the cells separately, with a composite estimator formed from the separate one-sample cell estimates, for example by taking a weighted average; or the two-sample method can be applied separately to all pairs of cells in the layout, with a weighted average composite estimator again constructed. In more general linear models like (3.1) without any replicate structure, the only way to use the Lehmann confidence interval approach is to apply the one-sample method to the residuals of the full linear model fit, treating them as one large sample. The one-sample Lehmann methods have the disadvantage of requiring symmetry of the underlying error distribution, making the two-sample approach more promising in settings for which the Lehmann robust estimation methods of Section 2 were designed, namely models with several observations per cell.

Focusing attention again on the model (2.1) of Section 2,

$$(4.9) \qquad Y_{ij} = \mu_i + e_{ij}, \quad \left\{ \begin{array}{l} i = 1, \cdots, I \\ j = 1, \cdots, n_i \\ \sum_{i=1}^{I} n_i = N, \end{array} \right\},$$

consider choosing two cells $1 \le i < k \le I$ and letting $L_{ik}(\alpha)$ be the length of the $100(1 - \alpha)\%$ confidence interval for $(\mu_i - \mu_k)$ based on the Wilcoxon rank-sum statistic applied to cells i and k. Then set

$$(4.10) \qquad B_{ik}(\alpha) = [3n_in_k/(n_i + n_k)]^{1/2}L_{ik}(\alpha)/\Phi^{-1}(1 - \alpha/2).$$

The $B_{ik}(\alpha)$ form a set of $\binom{I}{2}$ *dependent* estimates of $1/\theta$ because for example B_{ik} and $B_{ik'}$ have the data in cell i in common. A weighted average composite estimator of $1/\theta$ would take the form

$$(4.11) \qquad B(\alpha) = \sum_{i=1}^{I-1} \sum_{k=i+1}^{I} \lambda_{ik}B_{ik}(\alpha).$$

Two practical issues now arise: the choice of the weights λ_{ik} and specification of the confidence level $1 - \alpha$. One natural way to choose the λ_{ik} is to minimize the asymptotic variance of the composite estimator subject to the asymptotic unbiasedness condition $\sum_{i<k} \lambda_{ik} = 1$. Draper (1981) showed that, under the simplest realistic asymptotic specifications, in which the $n_i \to \infty$ in such a way that $n_i/N \to \rho_i$ $(0 < \rho_i < 1)$ for all $i = 1, \cdots, I$, the best choice of the λ_{ik} in this sense is

$$(4.12) \qquad \lambda_{ik} = 2[(I - 1)(\rho_i + \rho_k) - 1]/[(I - 1)(I - 2)]$$

(this result holds only when the number I of cells in the layout is greater than 2). In practice ρ_i is replaced by n_i/N, leading to the estimator

$$(4.13) \qquad B(\alpha) = \sum_{i=1}^{I-1} \sum_{k=i+1}^{I} 2\frac{(I - 1)(n_i + n_k) - N}{N(I - 1)(I - 2)} B_{ik}(\alpha)$$

of $1/\theta$, with corresponding estimate $B^2(\alpha)/12$ for σ_R^2.

Choice of the confidence level in (4.13) corresponds roughly to choosing the bin width when making a histogram; as shown empirically in Draper (1981), as the intervals widen, the bias of $B(\alpha)$ goes up but the standard error goes down. If the central objective were estimation of σ_R^2 itself, it would be reasonable to resolve this tension by minimizing the root mean square error $[(\text{bias})^2 + (\text{standard error})^2]^{1/2}$ as a function of α. But when the main goal is to construct good denominator estimates for use in the Lehmann and Hettmansperger-McKean testing ratios of Sections 2 and 3, it turns out (in order to be able to refer the testing ratios to familiar distributions like the F) that bias considerations dominate (Draper, 1981; Aubuchon and Hettmansperger, 1984a), and to make the bias small in small samples α should be fairly large. Simulations (Draper, 1981) indicate that after a simple bias correction is applied (replacement of $\Phi^{-1}(1 - \alpha/2)$ in $B_{ik}(\alpha)$ (4.10) by $t^{-1}_{n_i+n_k-2}(1 - \alpha/2)$), the two-sample estimator $\hat{\sigma}_R = B(\alpha)/12^{1/2}$ with a confidence level of about $100(1 - \alpha)\% = 50\%$ performs well both as a denominator estimate in the testing ratios of Lehmann and Hettmansperger-McKean and as an approximate standard error in the associated confidence procedures, in models with several observations per cell (in practice this means that most of the cells should have three or more observations).

4.1.2 The Schuster-Schweder Approach Based on Density Estimation

The Schuster-Schweder approach to the estimation of θ, rather than $1/\theta$, is a one-sample technique. Their method, in the one-sample problem with $Z_1, \cdots, Z_N$ iid with cdf F and density f, is based on the observation that

$$\theta = \int f^2(x)\, dx = \int f(x)\, dF(x), \tag{4.14}$$

so that a reasonable estimate of θ can be constructed by using the data twice simultaneously, once through a density estimate $\hat{f}_N$ of f and once with the empirical cdf $\hat{F}_N$:

$$\hat{\theta}_N = \int \hat{f}_N(x)\, d\hat{F}_N(x) = N^{-1} \sum_{i=1}^{N} \hat{f}_N(Z_i). \tag{4.15}$$

The density estimator Schuster and Schweder use is a *window* or *kernel* estimate (Rosenblatt, 1956; Wegman, 1972; Tapia and Thompson, 1978; Cheng and Serfling, 1981),

$$\hat{f}_N(x) = N^{-1} \sum_{i=1}^{N} w_N(x, Z_j), \tag{4.16}$$

in which

$$w_N(x, y) = h_N^{-1} w[(x - y)/h_N] \tag{4.17}$$

is a window or kernel function, where w is any density symmetric about 0, and in which h_N is the *window width*. Putting (4.15)–(4.17) together gives the Schuster-Schweder estimate of θ,

$$\hat{\theta}_N = N^{-2} h_N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} w[(Z_i - Z_j)/h_N], \tag{4.18}$$

which can be rewritten

$$\hat{\theta}_N = w(0)/Nh_N + N^{-2} h_N^{-1} \sum\sum_{i \neq j} w[(Z_i - Z_j)/h_N], \tag{4.19}$$

and the corresponding estimate of σ_R^2 is then $12^{-1}\hat{\theta}_N^{-2}$. Schweder (1975) demonstrated consistency of $\hat{\theta}_N$ for θ under the assumption of symmetry of f (together with some conditions on the window width h_N to be discussed below), but it was later pointed out by Aubuchon and Hettmansperger (1984a) and others that symmetry of f is not needed in the Schuster-Schweder approach.

The principal application of this method in the general linear model setting is, as with the Lehmann one-sample method, to treat the residuals from the full model fit as one large sample and insert them as Z_i's in (4.18). The fact that the residuals are dependent random quantities would be expected intuitively to affect the small sample behavior of the Schuster-Schweder estimator (requiring an upward bias correction to compensate for the underestimation of variability induced by the dependence), but not its large sample performance in an asymptotic setting in which the number of predictors p in the model (3.1) remains fixed and the sample size N grows, and indeed Aubuchon and Hettmansperger (1984b) have shown that consistency of $\hat{\theta}_N$ for θ still holds in that case (under standard regularity conditions on the design matrix X in (3.1)).

In carrying out the Schuster-Schweder idea there are two main choices to be made: the window or kernel function, and the window width. Popular choices for the window density (Aubuchon and Hettmansperger, 1984a; Hettmansperger, 1984) include the normal, rectangular (uniform), and triangular distributions (the triangular being a convolution of the rectangular with itself). Many authors (for example, Bean and Tsokos, 1980) have noted that the choice of window function in density estimation is not nearly as critical as the specification of the window width; in practice the rectangular or triangular densities are often favored on grounds of computational simplicity and speed. The main difficulty with the density estimation approach is the choice of window width, which like the confidence level $(1 - \alpha)$ in the Lehmann methods plays the role of the class interval width in histogram plotting but which, unlike Lehmann's confidence

level, requires considerable delicacy of selection. The same bias/variance tradeoff as in the Lehmann methods is present with the density estimation approach: large window widths lead to density estimates with small variance but large bias. Some guidance comes from asymptotic considerations: to get increasing accuracy as N increases, it turns out that h_N must go to 0, but not too fast. As with the Lehmann methods, the goal in constructing good denominator estimates of σ_R^2 for calibrating the Lehmann and Hettmansperger-McKean testing ratios of Sections 2 and 3 is to choose h_N to minimize the small sample bias of $\hat{\theta}_N$, and to make the bias small h_N should go to 0 as fast as possible. Aubuchon (1982) showed that the fastest possible rate at which h_N can tend to 0 for consistency of $\hat{\theta}_N$ for θ in linear model applications with dependent residuals is $h_N = O(N^{-1/2})$.

Schweder's (1975) original work on bias minimization has been extended by several authors, including Aubuchon and Hettmansperger (1984a) and Sheather and Hettmansperger (1985). Aubuchon and Hettmansperger's idea is first to modify $\hat{\theta}_N$ (4.19) slightly as follows:

$$\hat{\theta}_N = (Nk)^{-1} + [N(N-1)h_N]^{-1} \cdot \sum\sum_{i \neq j} w[(Z_i - Z_j)/h_N]. \tag{4.20}$$

This does not affect consistency and reduces the bias of $\hat{\theta}_N$ from $O(N^{-2/3})$ to $O(N^{-1})$. They then take $h_N = k/N^{1/2}$ and choose k to minimize bias $(\hat{\theta}_N)$, obtaining

$$k = \delta\left[2^{-1}\int_{-\infty}^{\infty}[f_1'(x)]^2\,dx \int_{-\infty}^{\infty} u^2 w(u)\,du\right]^{-1/3}. \tag{4.21}$$

Here $\int u^2 w(u)\,du$ is a known constant determined by the choice of the kernel function, δ is a scale factor obtained by reexpressing the underlying error density f as $f(x) = \delta^{-1} f_1(x/\delta)$, and $\int [f_1'(x)]^2\,dx$ is a shape factor whose appearance in a good choice of window width is sensible on intuitive grounds: $\int (f')^2$ is a global measure of how rapidly f changes its local behavior, and if $\int (f')^2$ is large the window width should be small.

Expression (4.21) summarizes the delicacy of the density estimation approach to estimating $\int f^2$. For good performance in small samples it is necessary to use the data in selecting the window width; in fact, the data must be used twice, to specify both the scale and the shape of the underlying distribution. Scaling is not difficult; Aubuchon and Hettmansperger (1984a) propose using the interquartile range or median absolute deviation from the median. The problem is with shape: one starts out estimating $\int f^2$ and discovers that to do so it is necessary to estimate $\int (f')^2$. Schweder's original idea, echoed by Aubuchon and Hettmansperger, was to choose a distribution like the Gaussian, calculate $\int (f')^2$ for it and hope that this choice works fairly well for all data sets, but simulations (Draper, 1981; Sheather and Hettmansperger, 1985) reveal predictably poor small sample performance of this implementation. Sheather and Hettmansperger (1985), following up on an idea also mentioned by Schweder, propose to carry out a second level of window estimation for $\int (f')^2$. They find, not surprisingly, that sensible choice of the window width at this level depends on the data through $\int (f'')^2$, but they also find through simulations in the one-sample iid case that when a constant is inserted in place of $\int (f'')^2$ at this second level of the density estimation process, much less harm is done than at the first level, and the method actually performs reasonably well in small samples across a variety of distributions in removing most of the bias of $\hat{\theta}_N$. The success of this approach when applied to dependent residuals from the full linear model fit in (3.1) to construct a denominator estimate for the Hettmansperger-McKean testing ratio has not yet been fully explored in simulations or theory.

4.2 Empirical Findings: Significance Testing

Once accurate estimates of σ_R^2 are developed, it still remains to arrive at good small sample approximations to the null distributions of the Lehmann and Hettmansperger-McKean testing ratios $L/\hat{\sigma}_R^2$ (2.36) and $2D_{\mathrm{JW}}^*/\hat{\sigma}_R$ (3.26), and to learn about their small sample performance. The asymptotic χ_q^2 distributions discussed in Sections 2 and 3 for both the Lehmann and Hettmansperger-McKean testing ratios have been found (Hettmansperger and McKean, 1977; Draper, 1981) 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_R^2 in the denominator instead of the true value results in distributions with heavier tails than χ_q^2. Huber (1970) conjectured that the small sample distribution of $\nu\hat{\sigma}_R^2/\sigma_R^2$ might be well approximated by χ_ν^2 for a value of ν depending on the error density f; simulations (Draper, 1981) have supported this conjecture quite well for bias-corrected versions of the Lehmann one- and two-sample estimators of σ_R^2 described above. This encourages the approximation of the null distribution of the Lehmann statistic $(L/q)/\hat{\sigma}_R^2$ by the heaver-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 has been found empirically (Draper, 1981; Hettmansperger and McKean, 1977) that the same F distribution that 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 $(L/q)/\hat{\sigma}_R^2$ and $(2D_{\mathrm{JW}}^*/q)/\hat{\sigma}_R$.

TABLE 3
Typical Monte Carlo power comparisons between the classical and rank-based robust testing procedures

Linear model	Error density	Testing method		Power at level		
		Numerator	Denominator	0.10	0.05	0.01
One-way layout, 6 cells, 10 observations per cell	Standard normal	Hettmansperger-McKean	Lehmann two-sample	0.94	0.87	0.70
		Classical		0.96	0.90	0.75
One-way layout, 6 cells, 10 observations per cell	t with 3 degrees of freedom	Lehmann	Lehmann one-sample	0.99	0.98	0.90
		Hettmansperger-McKean	Lehmann two-sample	0.99	0.97	0.88
		Classical		0.91	0.85	0.70
Two-way layout, 12 cells, 5 observations per cell	Skewed mixed normal (fourth entry in Table 2)	Hettmansperger-McKean	Lehmann two-sample	0.72	0.61	0.34
		Classical		0.62	0.49	0.24

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.

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. Simulations (Hettmansperger and McKean, 1977; Draper, 1981) indicate that the resulting tests not only have approximately correct levels with the wide variety of error distributions listed in Table 1 (with the actual level at nominal 0.05 ranging from about 0.04 to about 0.065, for instance, as N ranges from 10–60 in one- and two-way ANOVA layouts using the Lehmann two-sample denominator estimate), but also fulfill their promise in terms of asymptotic efficiency as indicated in that table by exhibiting good power characteristics relative to the classical F test. Table 3 (Draper, 1981) presents some typical power comparisons, which demonstrate that the power loss at the normal model for the robust methods is small, whereas the gain with skewed and heavy-tailed distributions can be considerable. In the robust analogues of the estimation and multiple comparison procedures typical in linear models work this efficiency gain manifests itself in more precise estimates and narrower confidence intervals.

5. CONCLUSIONS

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 simplest cases ... "

Thirty years later such robust testing and estimation methods have indeed been developed and are essentially ready for general use. Recent work of Hettmansperger and McKean (1977), Draper (1981) and others has provided rank-based methods that supply the linear models user, in modeling situations in which the assumption of an iid error structure is at least roughly tenable, with a comprehensive analysis package, from estimation and significance testing to confidence and multiple comparisons procedures, based on methods that have excellent efficiency and robustness properties relative to the standard methods and sufficient similarity in interpretation 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, so that evidence may continue to accumulate about the comparative utility of data-analytic, model expansion and robust approaches to linear model analysis.

ACKNOWLEDGMENTS

This article summarizes part of my thesis work with Erich Lehmann and Juliet Shaffer at the Department of Statistics, University of California, Berkeley, and underwent substantial revision while I was on the

faculty in the Department of Statistics, University of Chicago. I would like to acknowledge gratefully the advice and encouragement given me by many colleagues at Berkeley and Chicago, particularly Erich Lehmann, Juliet Shaffer and David Freedman at the former and David Wallace and Donald Rubin at the latter. Special thanks are due to my RAND colleague James Hodges for perceptive comments that improved the paper generally and the content and presentation of Section 1 in particular. This work was supported in part by National Science Foundation grants and RAND Corporation research funds.

REFERENCES

ADICHIE, J. N. (1967). Estimates of regression parameters based on rank tests. *Ann. Math. Statist.* **38** 894–904.

ADICHIE, J. N. (1978). Rank tests of subhypotheses in the general linear regression. *Ann. Statist.* **6** 1012–1026.

ANTILLE, A. (1974). A linearized version of the Hodges-Lehmann estimator. *Ann. Statist.* **2** 1308–1313.

ANTILLE, A. (1976). Asymptotic linearity of Wilcoxon signed-rank statistics. *Ann. Statist.* **4** 175–186.

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

AUBUCHON, J. C. (1982). Rank tests in the linear model: Asymmetric errors. Ph.D. dissertation, Dept. Statistics, Pennsylvania State Univ.

AUBUCHON, J. C. and HETTMANSPERGER, T. P. (1984a). A note on the estimation of the integral of $f^2(x)$. *J. Statist. Plann. Inference* **9** 321–331.

AUBUCHON, J. C. and HETTMANSPERGER, T. P. (1984b). Rank-based inference for linear models: Asymmetric errors. Unpublished manuscript.

BASSETT, G., JR. and KOENKER, R. (1982). An empirical quantile function for linear models with iid errors. *J. Amer. Statist. Assoc.* **77** 407–415.

BEAN, S. J. and TSOKOS, C. P. (1980). Developments in nonparametric density estimation. *Internat. Statist. Rev.* **48** 267–287.

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

BICKEL, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.* **1** 597–616.

BOX, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335.

BOX, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statist.* **25** 290–302.

BOX, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effect of inequality of variance and of correlation of errors in the two-way classification. *Ann. Math. Statist.* **25** 484–498.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

BOX, G. E. P. and TIAO, G. C. (1962). A further look at robustness via Bayes' theorem. *Biometrika* **49** 419–432.

BRADLEY, J. V. (1978). Robustness? *British J. Math. Statist. Psych.* **31** 144–152.

CARROLL, R. J. and RUPPERT, D. (1981). On prediction and the power transformation family. *Biometrika* **68** 609–616.

CARROLL, R. J. and RUPPERT, D. (1984). Comment on "The analysis of transformed data," by D. V. Hinkley and G. Runger. *J. Amer. Statist. Assoc.* **79** 312–313.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. and TUKEY, P. A. (1983). *Graphical Methods for Data Analysis.* Duxbury, Boston.

CHENG, K. F. and SERFLING, R. J. (1981). On estimation of a class of efficacy-related parameters. *Scand. Actuar. J.* **8** 83–92.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

CONOVER, W. J. (1980). *Practical Nonparametric Statistics*, 2nd ed. Wiley, New York.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

COX, D. R. (1977). Nonlinear models, residuals, and transformations. *Math. Operationsforsch. Statist. Ser. Statist.* **8** 3–22.

DANIEL, C. and WOOD, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd ed. Wiley, New York.

DIXON, W. J. (1983). *BMDP Statistical Software.* Univ. California Press, Berkeley.

DRAPER, D. (1981). Rank-based robust analysis of linear models. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

DRAPER, N. R. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. Wiley, New York.

FRIEDMAN, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32** 675–701.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models (with discussion). *Statist. Sci.* **1** 297–318.

HETTMANSPERGER, T. P. (1984). *Statistical Inference Based on Ranks.* Wiley, New York.

HETTMANSPERGER, T. P. and MCKEAN, J. W. (1976). Computational problems involved in analysis of linear models based on ranks. *Proc. Statist. Comp. Sec. Amer. Statist. Assoc.* 88–94.

HETTMANSPERGER, T. P. and MCKEAN, J. W. (1977). A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* **19** 275–284.

HETTMANSPERGER, T. P. and MCKEAN, J. W. (1978). Statistical inference based on ranks. *Psychometrika* **43** 69–79.

HODGES, J. L., JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

HUBER, P. J. (1970). Studentizing robust estimates. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 453–463. Cambridge Univ. Press, London.

HUBER, P. J. (1972). Robust statistics: A review. *Ann. Math. Statist.* **43** 1041–1067.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

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

HULL, C. H. and NIE, N. H. (1981). *SPSS Update 7-9: New Procedures and Facilities for Releases 7-9.* McGraw-Hill, New York.

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43** 1449–1458.

JUREČKOVÁ, J. (1971). Nonparametric estimation of regression coefficients. *Ann. Math. Statist.* **42** 1328–1338.

JUREČKOVÁ, J. (1973). Central limit theorem for Wilcoxon rank statistics process. *Ann. Statist.* **1** 1046–1060.

KENNEDY, W. J., JR. and GENTLE, J. E. (1980). *Statistical Computing.* Dekker, New York.

KOENKER, R. and BASSETT, G., JR. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KOUL, H. L. (1969). Asymptotic behavior of Wilcoxon-type confidence regions in multiple linear regression. *Ann. Math. Statist.* **40** 1950–1979.

KOUL, H. L. (1970). A class of ADF tests for subhypotheses in multiple linear regression. *Ann. Math. Statist.* **41** 1273–1281.

KOUL, H. L., SIEVERS, G. L. and MCKEAN, J. (1987). An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scand. J. Statist.* **14** 131–141.

KRUSKAL, W. H. and WALLIS, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* **47** 583–612.

LAPLACE, P. S. (1793). Sur quelques points du systeme du monde. *Mem. Acad. Roy. Sci. Paris* (1789) 1–87.

LEHMANN, E. L. (1963a). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34** 957–966.

LEHMANN, E. L. (1963b). Asymptotically nonparametric inference: An alternative approach to linear models. *Ann. Math. Statist.* **34** 1494–1506.

LEHMANN, E. L. (1963c). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.* **34** 1507–1512.

LEHMANN, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Ann. Math. Statist.* **35** 726–734.

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

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MCKEAN, J. W. and HETTMANSPERGER, T. P. (1976). Tests of hypotheses based on ranks in the general linear model. *Comm. Statist. A—Theory Methods* **5** 693–709.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, Mass.

POLICELLO, G. E., II and HETTMANSPERGER, T. P. (1976). Adaptive robust procedures for the one-sample location problem. *J. Amer. Statist. Assoc.* **71** 624–633.

PORTNOY, S. L. (1977). Robust estimation in dependent situations. *Ann. Statist.* **5** 22–43.

PORTNOY, S. L. (1979). Further remarks on robust estimation in dependent situations. *Ann. Statist.* **7** 224–231.

PURI, M. L. and SEN, P. K. (1969). A class of rank order tests for a general linear hypothesis. *Ann. Math. Statist.* **40** 1325–1343.

PURI, M. L. and SEN, P. K. (1973). A note on asymptotically distribution free tests for subhypotheses in multiple linear regression. *Ann. Statist.* **1** 553–556.

PURI, M. L. and SEN, P. K. (1977). Asymptotically distribution-free aligned rank-order tests for composite hypotheses for general multivariate linear models. *Z. Wahrsch. verw. Gebiete* **39** 175–186.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd ed. Wiley, New York.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–837.

RUPPERT, D. and CARROLL, R. J. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10** 429–441.

RYAN, B. F., JOINER, B. L. and RYAN, T. A., JR. (1985). *Minitab Handbook,* 2nd. ed. Duxbury, Boston.

SAS INSTITUTE, INC. (1985). *SAS User's Guide: Statistics,* Version 5 ed. SAS Institute, Inc., Cary, N. C.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

SCHRADER, R. M. and HETTMANSPERGER, T. P. (1980). Robust analysis of variance based on a likelihood ratio criterion. *Biometrika* **67** 93–101.

SCHRADER, R. M. and MCKEAN, J. W. (1977). Robust analysis of variance. *Comm. Statist. A—Theory Methods* **6** 879–894.

SCHUSTER, E. (1974). On the rate of convergence of an estimate of a functional of a probability density. *Scand. Actuar. J.* **1** 103–107.

SCHWEDER, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scand. J. Statist.* **2** 113–126.

SCHWEDER, T. (1981). Correction note. *Scand. J. Statist.* **8** 55.

SEAL, H. L. (1967). The historical development of the Gauss linear model. *Biometrika* **54** 1–24.

SEN, P. K. (1982). On M tests in linear models. *Biometrika* **69** 245–248.

SHEATHER, S. J. (1985). A new method of estimating the asymptotic standard error of the Hodges-Lehmann estimator based on least squares. Research Report 18, 1985, Dept. Statistics, Univ. Melbourne.

SHEATHER, S. J. and HETTMANSPERGER, T. P. (1985). A data-based algorithm for choosing the window width when estimating the integral of $f^2(x)$. Research Report 3, 1985, Dept. Statistics, Univ. Melbourne.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *Amer. J. Psych.* **15** 72–101.

SPJØTVOLL, E. (1968). A note on robust estimation in analysis of variance. *Ann. Math. Statist.* **39** 1486–1492.

SRIVASTAVA, M. S. (1972). Asymptotically most powerful rank tests for regression parameters in MANOVA. *Ann. Inst. Statist. Math.* **24** 285–297.

STIGLER, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard Univ. Press, Cambridge, Mass.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation.* Johns Hopkins Univ. Press, Baltimore, Md.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, ed.) 445–485. Stanford Univ. Press, Stanford, Calif.

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.

WEGMAN, E. J. (1972). Nonparametric probability density estimation: I. A survey of available methods. *Technometrics* **14** 533–546.

WEISBERG, S. (1984). Robustness and diagnostics: Black box or Pandora's box? Presented at the American Statistical Association annual meeting, Philadelphia, August 14, 1984.

WEISBERG, S. (1985). *Applied Linear Regression,* 2nd ed. Wiley, New York.

WELSH, A. H. (1987a). The trimmed mean in the linear model (with discussion). *Ann. Statist.* **15** 20–45.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1** 80–83.