The Annals of Statistics
1987, Vol. 15, No. 1, 202–219

k-STATISTICS AND DISPERSION EFFECTS IN REGRESSION

BY PETER MCCULLAGH[1] AND DARYL PREGIBON

University of Chicago and AT & T Bell Laboratories

By the term k-statistic or polykay, we mean an unbiased estimate of a cumulant or product of cumulants [Fisher (1929) and Tukey (1950, 1956)]. In this paper, two sets of unbiased estimates are given for the case where the mean response, $E(Y)$, depends linearly on known covariates x. The k's are symmetric functions of the least-squares residuals and have previously been discussed by Anscombe (1961; 1981, Appendix 2). The l's are optimal in the sense of having minimum variance under the ideal assumption of normality [Pukelsheim (1980)]. The emphasis here on computability leads to the algebraic inversion of direct product matrices of order $n^3 \times n^3$ and $n^4 \times n^4$, a computation that is rarely feasible numerically, even on the fastest computers. This algebra leads to simple straightforward formulae for all statistics up to degree four. Conditions are given under which the k's are nearly or asymptotically optimal in the sense of being asymptotically equivalent to the corresponding l's. A small-scale simulation study provides a comparison between these statistics for finite n. An application to detecting heterogeneity of variance, avoiding the assumption of normality, is given. A new test statistic for detecting systematic dispersion effects is introduced and compared to existing ones. Two examples illustrate the methodology.

1. Introduction. In the usual linear model, the observations $Y^1, \ldots, Y^n$ are assumed to satisfy

$$E(Y^i) = x^i_r \beta^r,$$

where $\mathbf{X} = \{x^i_r\}$ is an $n \times p$ matrix of known constants, β is a p-dimensional vector of unknown parameters and the summation convention is applied to any index repeated once as a subscript and once as a superscript. In addition, it is commonly assumed that, for $i = 1, \ldots, n$, the errors $Y^i - x^i_r \beta^r$ are independently distributed with cumulants $\kappa_2 \delta^{ij}$, $\kappa_3 \delta^{ijk}$, $\kappa_4 \delta^{ijkl}$, and so on. In this paper, unbiased estimates are presented for the cumulants $\kappa_2, \kappa_3, \kappa_4, \ldots$, as well as the product, κ_2^2. The optimal estimates, $l_2, l_3, l_4, l_{22}, \ldots$, are the homogeneous polynomial functions of the residuals that have minimum variance in the ideal case where the observations are normally distributed and independent. The simpler estimates, k_2, k_3, k_4, k_{22}, that are symmetric homogeneous polynomial functions of the residuals, have been used by Anscombe (1961), who noted, with surprise, that the symmetric functions are not optimal in general.

Pukelsheim (1980) gives a formula for l_3, there denoted by $\sigma^3 \hat{\gamma}_1$, but his expression involves the generalized inverse of a $n^3 \times n^3$ matrix. Since n is typically fairly large, this formula is of little use for computation. Indeed, no

Received February 1986; revised June 1986.
[1]This research was supported in part by NSF Grant DMS86-01732.
AMS 1980 *subject classifications.* Primary 63E30, 62J05; secondary 62F35, 62N10.
Key words and phrases. Polykay, cumulant, residual test, tensor, generalized inverse, heterogeneity of variance.

numerical examples involving the computation of l_3 are given in Pukelsheim's paper. In this paper, we use the tensor notation of McCullagh (1984) to compute the necessary generalized inverses algebraically. The formulae are simplified to such an extent that routine computation is feasible. In fact, the l-statistics are not much more difficult to compute than the k-statistics.

Conditions are given under which the simpler k-statistics are optimal in the sense of having minimum variance under the ideal conditions. In fact, $k_2 \equiv l_2$ for all $\mathbf{X}$ and all n. More generally, and perhaps more usefully, it is shown that for large n, and under suitably mild limiting conditions on $\mathbf{X}$, that

$$n^{1/2}(k_3 - l_3) = O_p(1),$$

$$n^{1/2}(k_4 - l_4) = O_p(n^{-1/2}),$$

and

$$n^{1/2}(k_{22} - l_{22}) = O_p(n^{-3/2}).$$

Moreover, if the constant vector lies in the column space of $\mathbf{X}$, as would commonly be the case in applications, the first approximation above becomes

$$n^{1/2}(k_3 - l_3) = O_p(n^{-1}).$$

Note here that $n^{1/2}(k_3 - \kappa_3)$ is typically $O_p(1)$ for large n, and similarly for the other k's and l's so that the differences on the left of the above equations involve random variables that are $O_p(1)$. Thus the differences, apart from the first, are at least half an order of magnitude smaller than the random variables themselves. In other words, for large n the simpler symmetric functions of the residuals are nearly optimal.

This asymptotic result is important for two reasons. First, one would not normally consider estimating the higher-order cumulants unless the sample size was at least 50 and preferably more than 100. Second, the above analysis indicates that it is only when the constant vector is not included in the model that there is likely to be any appreciable difference between the k-statistics and the optimal l-statistics. In other words, there is reason to believe that the difference between the k-statistics and the l-statistics is likely to be negligible for all samples large enough to make estimation of the cumulants interesting and useful. This suspicion can be confirmed by simulation.

In the case of weighted regression as discussed in Section 5, the differences $n^{1/2}(k_3 - l_3)$ and $n^{1/2}(k_4 - l_4)$ are both $O_p(1)$ for large n. It follows that the k's are then not asymptotically optimal and the l-statistics are preferred. The difference between these statistics for finite n and for nonnormal samples is illustrated in Section 7 by means of a small-scale simulation experiment.

The latter part of the paper is devoted to using k-statistics in the analysis of dispersion effects in regression. This is an important application since classical tests for heterogeneity rely heavily on normality and k-statistics provide a means of accommodating excess skewness and kurtosis. This is done both directly and through a correction factor applied to classical tests. Two examples are introduced to illustrate the methodology. In both examples the numerical

differences between the k's and the optimal l's are negligible so that either statistic could be used to adjust for nonnormality.

2. Second-order k-statistics. We consider only homogeneous polynomial functions of the least-squares residuals, which are given in matrix notation by

$$\mathbf{R} = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{Y}. \tag{1}$$

For our purposes, it is more convenient to use index notation in the form

$$R^i = \rho^i_j Y^j,$$

where $\{\rho^i_j\}$, the residual projection matrix, has rank $\nu \le n$. If $\mathbf{X}$ has full rank then $\nu = n - p$. More generally, if $\text{rank}(\mathbf{X}) \le p$, then a suitable generalized inverse can be used in (1) and $\nu \ge n - p$.

The covariance matrix of the residuals is

$$\begin{aligned}\text{cov}(R^i, R^j) &= \rho^i_k \rho^j_l \text{cov}(Y^k, Y^l) \\ &= \kappa_2 \rho^i_k \rho^j_l \delta^{kl} = \kappa_2 \rho^{i,j}.\end{aligned}$$

Note that $\rho^{i,j}$ is numerically identical to ρ^i_j, and the two arrays have the same matrix form but, as tensors under the permutation group, it is extremely helpful to maintain the distinction when using index notation. Note here that the permutation group is considered to be applied simultaneously to the rows of $\mathbf{X}$ and $\mathbf{Y}$. All derived statistics are required to be invariant under this operation.

The obvious and simplest estimate of κ_2 is based on the sum of squares of the residuals ignoring cross products. Thus,

$$k_2 = \nu^{-1}\delta_{ij}R^iR^j = \nu^{-1}S_2 \tag{2}$$

has expectation $\nu^{-1}\delta_{ij}\rho^{i,j}\kappa_2 = \kappa_2$, and variance

$$\nu^{-2}\delta_{ij}\delta_{kl}\left\{\kappa_2^2\left(\rho^{i,k}\rho^{j,l} + \rho^{i,l}\rho^{j,k}\right) + \kappa_4\rho^{i,j,k,l}\right\},$$

where

$$\rho^{i,j,k,l} = \rho^i_r\rho^j_s\rho^k_t\rho^l_u\,\delta^{rstu} = \sum_r \rho^i_r\rho^j_r\rho^k_r\rho^l_r.$$

Simplification gives

$$\text{var}(k_2) = 2\kappa_2^2/\nu + \kappa_4\sum\left(\rho^{i,i}\right)^2/\nu^2,$$

which reduces to $2\kappa_2^2/\nu$ under the ideal assumption of normality.

To derive the unbiased quadratic function of the residuals having minimum variance under the ideal assumptions, we apply the method of least squares to the vector having n^2 components R^iR^j. Most components are duplicated, but this duplication is of no consequence. Now, R^iR^j has expectation $\kappa_2\rho^{i,j}$ in the general case and covariance matrix

$$\begin{aligned}\text{cov}(R^iR^j, R^kR^l) &= \kappa_2^2\left(\rho^{i,k}\rho^{j,l} + \rho^{i,l}\rho^{j,k}\right) \\ &= \kappa_2^2 w^{ij,kl},\end{aligned}$$

under the ideal conditions. The Moore–Penrose generalized inverse of $w^{ij,kl}$ is easily seen to be

$$w_{ij,kl} = (\rho_{i,k}\rho_{j,l} + \rho_{i,l}\rho_{j,k})/4,$$

where $\rho_{i,j}$, the Moore–Penrose inverse of $\rho^{i,j}$, is numerically identical to $\rho^{i,j}$ and ρ^i_j. Thus the weighted least-squares estimate of κ_2 is

$$l_2 = \left(\rho^{i,j}w_{ij,kl}\rho^{k,l}\right)^{-1}\rho^{i,j}w_{ij,kl}R^kR^l, \tag{3}$$

and this is easily seen to be the same as k_2. More generally, it may be shown that if the design is quadratically balanced, so that $\rho^{i,i} = \nu/n$, then k_2 has minimum variance among positive quadratic forms for all parent distributions for which $\kappa_4 < \infty$ [Atiqullah (1962)].

Note that the choice of generalized inverse is not critical and it is in fact slightly simpler if we take $w_{ij,kl}$ to be

$$(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})/4 \quad \text{or even} \quad \delta_{ik}\delta_{jl}/2.$$

This may seem an extremely roundabout way of demonstrating a well-known general result, namely that unbiased estimates based on the sufficient statistic have minimum variance. However, the method that we have used extends easily to the higher-order l-statistics as we now show.

3. Third-order k-statistics. The simplest estimate of κ_3, based on the residuals and generalizing Fisher (1929) is

$$k_3 = \nu_3^{-1}\delta_{ijk}R^iR^jR^k = \nu_3^{-1}\sum_i (R^i)^3 = \nu_3^{-1}S_3, \tag{4}$$

where $\nu_3 = \Sigma_{ij}(\rho_{i,j})^3$ is assumed to be nonzero. Typically, but not invariably, ν_3 is a little smaller than ν and the condition that ν_3 be nonzero need not be a cause for concern in most circumstances. For a counterexample, however, see Anscombe (1961, Section 4.2).

In the simple case where $\mathbf{X} = \mathbf{1}$, the Y's comprise a simple random sample and k_3 reduces to Fisher's (1929) statistic. However, if $\mathbf{X}$ is degenerate so that the Y's comprise a simple random sample of zero mean, k_3 is the sum of cubes of the raw data and is therefore different from Fisher's statistic. In the degenerate case, both statistics are unbiased for κ_3 and both are symmetric functions. Evidently, in the latter case, symmetry does not guarantee uniqueness. Anscombe (1981, page 266) has shown in this case that it is preferable to use the sum of cubes about the sample mean than the sum of cubes of the raw data. Both of these are inferior to the estimate described below.

To compute the optimal estimate, we first consider the expectation of the triple product $R^iR^jR^k$, which is

$$\kappa_3\rho^i_r\rho^j_s\rho^k_t\,\delta^{rst} = \kappa_3\rho^{i,j,k}.$$

Under the ideal conditions, the covariance of $R^iR^jR^k$ and $R^lR^mR^n$ is $\kappa_2^3 w^{ijk,lmn}$, where

$$w^{ijk,lmn} = \rho^{i,l}\rho^{j,m}\rho^{k,n}[6] + \rho^{i,j}\rho^{k,l}\rho^{m,n}[9]$$

and the figures in square brackets denote summation over distinct partitions of the indices and induce the necessary symmetry. See McCullagh (1984) for more extensive uses of this notation. For an alternative notation involving Kronecker products, Hadamard products and other devices, see Pukelsheim (1980).

The simplest generalized inverse of w is

$$\delta_{il}\delta_{jm}\delta_{kn}/6 - \delta_{ij}\delta_{kl}\delta_{mn}/\{2(\nu + 4)\},$$

and the Moore–Penrose inverse is

$$w_{ijk,lmn} = \rho_{i,l}\rho_{j,m}\rho_{k,n}[6]/36 - \rho_{i,j}\rho_{k,l}\rho_{m,n}[9]/\{18(\nu + 4)\}.$$

These claims are easily verified directly by matrix multiplication. It follows that

$$\rho^{i,j,k}\rho^{l,m,n}w_{ijk,lmn} = \nu_3/6 - \sum_{ij}\rho^{i,i}\rho^{j,j}\rho_{i,j}/\{2(\nu + 4)\}$$

and

$$\rho^{i,j,k}w_{ijk,lmn}R^lR^mR^n = S_3/6 - S_2\sum_i \rho^i_i R^i/\{2(\nu + 4)\}.$$

Thus, the homogeneous cubic function of the residuals that is unbiased for κ_3 and has minimum variance under the ideal conditions is

$$l_3 = \frac{S_3 - 3S_2\Sigma\rho^i_i R^i/(\nu + 4)}{\nu_3 - 3\Sigma\rho^{i,i}\rho^{j,j}\rho_{i,j}/(\nu + 4)}. \tag{5}$$

It follows that if, and only if, the vector with elements $\rho^{i,i}$ lies in the column space of $\mathbf{X}$, then $l_3 \equiv k_3$.

Under normality, we have that

$$\operatorname{var}(l_3) = 6\kappa_2^3/\Big\{\nu_3 - 3\sum\rho^{i,i}\rho^{j,j}\rho_{i,j}/(\nu + 4)\Big\},$$

whereas

$$\operatorname{var}(k_3) = \kappa_2^3\big(6\nu_3 + 9\sum\rho^{i,i}\rho^{j,j}\rho_{i,j}\big)/\nu_3^2.$$

In the present context, the positions of the indices in the above formulae are immaterial, but in the context of weighted regression discussed in Section 5, the position of the indices becomes relevant.

4. Fourth-order k-statistics. Following Tukey (1950, 1956), we denote the estimates of κ_4 and κ_2^2 by k_4 and k_{22}, respectively. The optimal estimates are denoted by l_4 and l_{22}. The simplest estimates, based on symmetric homogeneous polynomial functions of degree four in the residuals are

$$k_4 = \big\{\nu(\nu + 2)S_4 - 3\nu_{22}S_2^2\big\}/\Delta, \tag{6}$$

$$k_{22} = \big\{\nu_4 S_2^2 - \nu_{22}S_4\big\}/\Delta, \tag{7}$$

where

$$S_4 = \sum(R^i)^4, \qquad \nu_4 = \sum(\rho_{i,j})^4, \qquad \nu_{22} = \sum(\rho_{i,i})^2,$$
$$\Delta = \nu(\nu - 1)\nu_4 + 3\big(\nu\nu_4 - \nu_{22}^2\big).$$

It is necessary here to assume that Δ is nonzero and this condition is typically satisfied if $\nu \geq 2$. Note that $\nu = 1$ implies $\Delta = 0$, as we might expect. In other words, there is no unbiased estimate of κ_4 or of κ_2^2 based on the residuals alone, unless the residual degrees of freedom are at least 2. Contrast Pukelsheim (1980, Lemma 2.2), where κ_2 is assumed known.

To compute the optimal estimates, we first consider the expectation of the product $R^iR^jR^kR^l$, which is

$$\kappa_4\rho^{i,j,k,l} + \kappa_2^2\rho^{i,j}\rho^{k,l}[3].$$

Under the ideal conditions, the covariance of $R^iR^jR^kR^l$ and $R^{i'}R^{j'}R^{k'}R^{l'}$ is $\kappa_2^4 w^{ijkl,i'j'k'l'}$, where

$$w^{ijkl,i'j'k'l'} = \rho^{i,i'}\rho^{j,j'}\rho^{k,k'}\rho^{l,l'}[24] + \rho^{i,j}\rho^{i',j'}\rho^{k,k'}\rho^{l,l'}[72].$$

By matrix multiplication, it is easily verified that the Moore–Penrose generalized inverse is

$$\begin{aligned} w_{ijkl,i'j'k'l'} = {} & \rho_{i,i'}\rho_{j,j'}\rho_{k,k'}\rho_{l,l'}[24]/576 \\ & - \rho_{i,j}\rho_{i',j'}\rho_{k,k'}\rho_{l,l'}[72]/\{288(\nu+6)\} \\ & + \rho_{i,j}\rho_{k,l}\rho_{i',j'}\rho_{k',l'}[9]/\{72(\nu+3)(\nu+6)\}. \end{aligned}$$

From the scalar products

$$\rho^{i,j,k,l}w_{ijkl,i'j'k'l'}R^{i'}R^{j'}R^{k'}R^{l'}$$

and

$$\rho^{i,j}\rho^{k,l}[3]w_{ijkl,i'j'k'l'}R^{i'}R^{j'}R^{k'}R^{l'},$$

which arise in computing the weighted least-squares estimate, it is easily seen that l_4 and l_{22} must be linear functions of

$$S_4 - 6S_2\sum\rho_i^i(R^i)^2/(\nu+6)$$

and S_2^2. The three quartics involved have the following expectations:

$$\begin{aligned} E(S_4) &= \nu_4\kappa_4 + 3\nu_{22}\kappa_2^2, \\ E\left(S_2^2\right) &= \nu_{22}\kappa_4 + \nu(\nu+2)\kappa_2^2, \\ E\left(S_2\sum\rho_i^i(R^i)^2\right) &= \mu\kappa_4 + \nu_{22}(\nu+2)\kappa_2^2, \end{aligned}$$

where $\mu = \sum\rho_i^i\rho_j^j(\rho^{i,j})^2$. If we write

$$\Delta_l = \Delta - \frac{6(\nu+2)}{\nu+6}\left\{\nu\mu - \nu_{22}^2\right\},$$

it follows that the optimal estimates are given by

$$\Delta_l l_4 = \Delta k_4 - 6\nu(\nu+2)S_2\varepsilon/(\nu+6), \tag{8}$$

$$\Delta_l l_{22} = \Delta k_{22} + 6\nu_{22}S_2\delta/(\nu+6), \tag{9}$$

where

$$\varepsilon = \sum \rho_i^i (R^i)^2 - \nu_{22} k_2,$$

$$\delta = \sum \rho_i^i (R^i)^2 - \mu\nu k_2/\nu_{22}.$$

In the quadratically balanced case, for which the residuals have equal variances, we have

$$\rho_{i,i} = \nu/n = \nu_{22}/\nu.$$

It follows then that $\varepsilon = 0$, $\nu\mu = \nu_{22}^2$ and hence that $\delta = 0$, $\Delta_l = \Delta$, $l_4 = k_4$ and $l_{22} = k_{22}$.

Note also that if $\nu = 1$, then $\Delta_l = \Delta = 0$, so that l_4 and l_{22} exist only if $\nu \geq 2$.

Under normality, we have that

$$\operatorname{var}(l_4) = \kappa_2^4 \frac{24\nu(\nu+2)(\nu+6)}{\nu_4\nu(\nu+2)(\nu+6) + 3\nu_{22}^2(\nu-2) - 6\mu\nu(\nu+2)},$$

whereas

$$\operatorname{var}(k_4) = \kappa_2^4 \frac{24\nu(\nu+2)\{\Delta + 3(\nu+2)(\nu\mu - \nu_{22}^2)\}}{\Delta^2}.$$

5. Weighted regression. We suppose now that the ith observation Y^i is the average of m_i independent and identically distributed, but unrecorded random variables

$$Y^i = \left(Z_1^i + \cdots + Z_{m_i}^i\right)/m_i.$$

It is assumed that

$$E\left(Z_j^i\right) = x_r^i \beta^r,$$

implying that Y^i satisfies the same linear model as the Z's. By assumption, the Y's are independently distributed with cumulants $m_i^{-1}\kappa_2, m_i^{-2}\kappa_3, m_i^{-3}\kappa_4, \ldots$, where κ_r is the rth cumulant of Z. To estimate these cumulants based on the observations (Y^i, m_i), minor changes are required in the expressions given in the previous sections.

Let $\mathbf{W} = \operatorname{diag}\{m_1, \ldots, m_n\}$ and $\mathbf{V} = \mathbf{W}^{-1}$ so that $\operatorname{cov}(\mathbf{Y}) = \kappa_2 \mathbf{V}$. The residual vector is

$$\mathbf{R} = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)\mathbf{Y},$$

in matrix notation, or

$$R^i = \rho_j^i Y^j,$$

in index notation, so that ρ_j^i is no longer symmetrical. The covariance matrix of the residuals is

$$\kappa_2\left(\mathbf{V} - \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\right)$$

or $\kappa_2 \rho^{i,j} = \kappa_2 \rho_j^i / m_j$ using indices. The Moore–Penrose inverse is

$$\kappa_2^{-1}\left(\mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)$$

or $\kappa_2^{-1}\rho_{i,j}$ using indices. Thus $\rho^{i,j} = \rho^i_j/m_j$ while $\rho_{i,j} = m_i\rho^i_j$, no summation intended.

If we define

$$S_r = \sum_i m_i(R^i)^r$$

for the weighted rth power sum of the residuals and

$$\nu_r = \sum_{ij} m_i m_j(\rho^{i,j})^r,$$

$$\nu_{22} = \sum_i m_i(\rho^{i,i})^2,$$

it follows that $\nu_2 = \nu$ and that Equations (2)–(9) give the required unbiased estimates of the cumulants. Note that the positions of the indices in (5), (8) and (9) are now important.

6. Large sample approximation. We consider now the formal limit in the equally weighted case where the number of observations becomes large and the number of parameters in the model remains fixed. In other words, n is assumed to be large and hence ν is large. Under the usual assumptions regarding the limiting behavior of $\mathbf{X}$, namely that the eigenvalues of $n^{-1}\mathbf{X}^T\mathbf{X}$ have positive limits as $n \to \infty$, we may write

$$\rho_{i,i} = 1 - a_i/n$$

and

$$\rho_{i,j} = a_{ij}/n, \qquad i \neq j,$$

where a_i and a_{ij} are $O(1)$ for large n.

In the case of the third cumulant, it follows from (5) that $\nu_3 = \nu + O(1)$ and

$$l_3 = \frac{k_3 - 3k_2\overline{R} + O_p(n^{-1})}{1 - 3\Sigma\rho_{i,j}/\nu^2 + O(n^{-2})}, \tag{10}$$

where $\overline{R} = n^{-1}\Sigma R^i$. Hence,

$$n^{1/2}(k_3 - l_3) \sim 3n^{1/2}k_2\overline{R} = O_p(1).$$

However, if the constant vector lies in the column space of $\mathbf{X}$, then $\overline{R} = 0$ and a similar analysis then gives

$$n^{1/2}(k_3 - l_3) = O_p(n^{-1}).$$

For statistics of degree four, we have

$$\Delta = O(n^3), \qquad \Delta_l - \Delta = O(1) \quad \text{and} \quad \varepsilon = O_p(1).$$

Hence, from (8) and (9) we find

$$l_4 = k_4 + O_p(n^{-1})$$

and

$$l_{22} = k_{22} + O_p(n^{-2}).$$

These give the required asymptotic results as stated in Section 1.

A similar analysis may be used in the (unequally) weighted case provided that $n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}$ has positive limiting eigenvalues and that $\Sigma(m_i - \bar{m})^2/n$ tends to a nonzero limit as n becomes large. However, when the weights are unequal, the average residual $\bar{R}$ in (10) is no longer zero unless the constant vector lies in the column space of $\mathbf{W}\mathbf{X}$ and there is no reason to expect this in general. This leads to the conclusion that, when the weights are unequal, $n^{1/2}(k_3 - l_3) = O_p(1)$ for large n. In the case of the fourth-order statistics, we find $\Delta_l - \Delta = O(n^2)$ and $\varepsilon = O(n^{1/2})$, leading to the conclusion that

$$n^{1/2}(k_4 - l_4) = O_p(1),$$

$$n^{1/2}(k_{22} - l_{22}) = O_p(n^{-1/2}).$$

7. Simulation results. We report results from a small-scale simulation experiment to calibrate the theory of the preceding sections. The results demonstrate the improvement of l-statistics over k-statistics not only in the Gaussian case (for which the l's are optimal) but also for sampling from a gamma distribution.

We use the *car insurance claims* example of McCullagh and Nelder (1983, page 158) as a basis for the simulation. This example consists of average insurance claims in a $4 \times 4 \times 8$ design. Each average is based on widely different numbers of individual claims ranging from 1 to 434. There are 5 empty cells in the table leaving $n = 123$ observations. We chose this example since it is of sufficient size to contemplate estimating high-order cumulants and since it provided us an opportunity to check the algebra in Section 5, the case of weighted regression.

Let m_i denote the number of observations that each average claim was based upon. For each of 1000 normal samples

$$\{y_i \sim \mathsf{N}(0, m_i^{-1}) : i = 1, \ldots, 123\}$$

and for each of 1000 gamma samples

$$\{y_i \sim \mathsf{G}(m_i)/m_i : i = 1, \ldots, 123\},$$

we computed third- and fourth-order k- and l-statistics after fitting the weighted regression model defined by the example. For computational efficiency, quantities involving only the projection matrix $\{\rho_j^i\}$ were computed only once, outside the simulation, e.g., the ν's, Δ's, and μ. Table 1 summarizes the results of the experiment.

For both normal and gamma samples, we find good agreement between empirical, and where available, theoretical quantities. Estimates appear to be unbiased and the minimum variance property of the l's, under normality, is clearly demonstrated. A gratifying finding is that the l's are moderately (20–30%) more efficient than the k's even for gamma errors.

TABLE 1
Results from limited simulation experiment based on the car insurance claims example

	Normal samples				Gamma samples		
	Mean		Variance		Mean		Variance
	Theoretical value	Empirical value	Theoretical value	Empirical value	Theoretical value	Empirical value	Empirical value
k_2	1.	0.998	0.0183	0.0172	1.	0.998	0.0248
k_3	0.	0.042	0.911	0.962	2.	2.00	9.87
k_4	0.	−0.173	11.0	10.5	6.	6.03	975.
k_{22}	1.	0.994	—	0.0687	1.	0.996	0.0997
l_2	1.	0.998	0.0183	0.0172	1.	0.998	0.0248
l_3	0.	0.0229	0.481	0.465	2.	2.00	8.07
l_4	0.	−0.0208	3.40	3.12	6.	5.98	733.
l_{22}	1.	0.994	—	0.0686	1.	0.996	0.101

Since this is a weighted regression problem, we expect to find differences between the k's and l's, through not in their means, and the simulation quantifies the extent of such differences. Although paired t-tests show no significant differences at the standard 5% level, certain anomalies do appear when the paired differences in k-statistics are plotted against the optimal l's. (See Figure 1.) For gamma samples, the obvious feature displayed in the plots is that k_3 and k_4 are consistently larger than l_3 and l_4 for large values of the latter. Since both are unbiased, the implication is that, for small values of l_3 and l_4, the k's are smaller on average than the l's. This phenomenon is readily apparent when the plots are redrawn over a more limited range of the x-axis (not shown). For normal samples, near-random scatter is displayed for third-order k-statistics though there is a suggestion that for large $|l_3|$, the corresponding $|k_3|$ are somewhat larger. For fourth-order k-statistics, the pattern of dependence between k_4 and l_4 is similar to that in the gamma samples where for small $|l_4|$, there is strong negative association between $k_4 - l_4$ and l_4, and for large l_4, the corresponding k_4 are consistently larger.

8. Applications: tests for systematic dispersion effects. Classical tests for heterogeneity of variance are known to be sensitive to nonnormality. Bartlett's test, for example, is sensitive to excess kurtosis [Box (1953)]. In this section, we develop a test that is designed to detect systematic trends in variance of an easily understood type. The results developed in the preceding sections are used to take account of excess skewness and kurtosis as measured by κ_3 and κ_4, so that reliance on the normal distribution can be avoided. Bickel (1978) discusses an alternative approach to robust tests for heterogeneity.

8.1. *Methodology.* Suppose that having fitted the linear model $E(\mathbf{Y}) = \mathbf{X}\beta$ by least squares, we wish to test for systematic trends in variance. We proceed

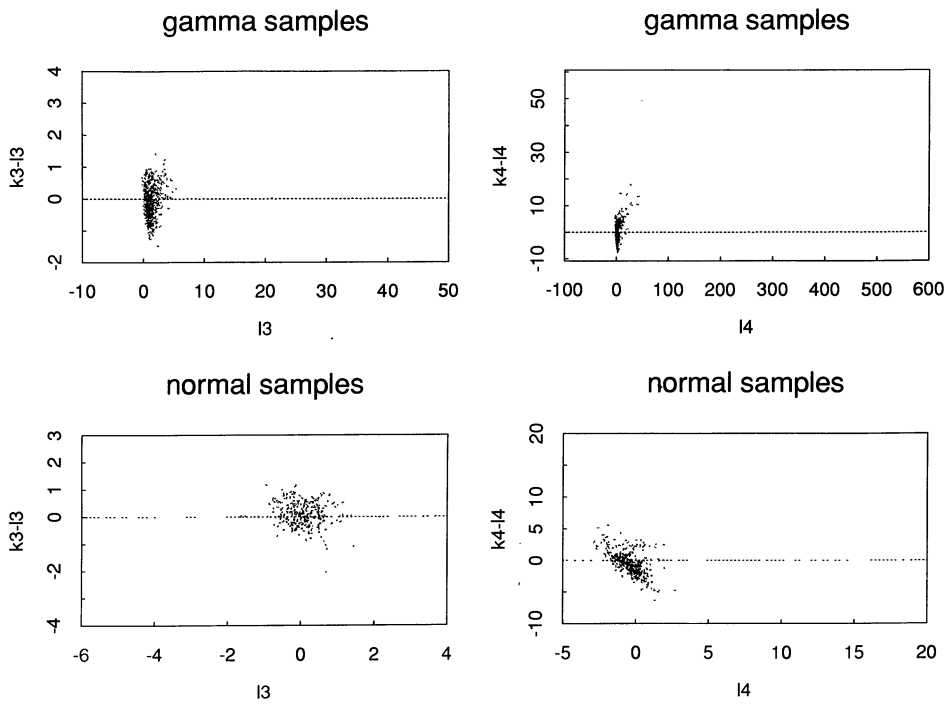

FIG. 1. *Scatter plots of the difference in k-statistics versus the optimal l's for the simulation study in Section 7. The top panels display $k_m - l_m$ versus l_m, $m = 3, 4$, for the gamma samples. The bottom panels display the same for the normal samples. The horizontal line superimposed on each plot has zero intercept.*

by computing specified linear combinations of squared residuals and comparing these with the observed value of κ_2.

Let $d^i = (Y^i - x_r^i \hat{\beta}^r)^2$ denote the ith squared residual. We base our test on the q linear combinations $\mathbf{Z}^T\mathbf{d}$, where $\mathbf{Z}$ is a given $n \times q$ matrix. Since the distribution of $\mathbf{Z}^T\mathbf{d}$ depends on the unknown cumulants $\kappa_2, \kappa_3, \ldots$, and not on β, our proposal is to base our inferences on the conditional distribution of $\mathbf{Z}^T\mathbf{d}$ given the observed value of k_2. The conditional distribution is approximately free of the unknown parameters. In principle, we should also condition on k_3 and k_4 but this additional conditioning seems not to affect the conclusions to the order of approximation used here.

Let $\mathbf{A}$ be the $n \times n$ symmetric matrix with elements

$$a_{ij} = 2(\rho^{i,j})^2\kappa_2^2 + \rho^{i,i,j,j}\kappa_4.$$

Then the covariance matrix of k_2 and $\mathbf{Z}^T\mathbf{d}$ is

$$\begin{bmatrix} \mathbf{1}^T\mathbf{A}\mathbf{1}/\nu^2 & \mathbf{1}^T\mathbf{A}\mathbf{Z}/\nu \\ \mathbf{Z}^T\mathbf{A}\mathbf{1}/\nu & \mathbf{Z}^T\mathbf{A}\mathbf{Z} \end{bmatrix},$$

while the unconditional mean of $\mathbf{Z}^T\mathbf{d}$ is

$$E(\mathbf{Z}^T\mathbf{d}) = \kappa_2\mathbf{Z}^T\boldsymbol{\rho},$$

where we have written $\boldsymbol{\rho}$ for the vector with elements $\rho^{i,i}$. To the extent that the central limit theorem holds for large n, it follows that

$$E(\mathbf{Z}^T\mathbf{d}|k_2) \simeq k_2\mathbf{Z}^T\boldsymbol{\rho},$$

$$\operatorname{cov}(\mathbf{Z}^T\mathbf{d}|k_2) \simeq \mathbf{Z}^T\mathbf{A}(\mathbf{I} - \mathbf{P})\mathbf{Z},$$

where $\mathbf{P} = \mathbf{1}(\mathbf{1}^T\mathbf{A}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{A}$ is a projection matrix and $\mathbf{I} - \mathbf{P}$ projects onto the orthogonal complement of $\mathbf{1}$. If the columns of $\mathbf{Z}$ sum to zero as henceforth assumed, we have that $(\mathbf{I} - \mathbf{P})\mathbf{Z} \simeq \mathbf{Z}$, and hence $\operatorname{cov}(\mathbf{Z}^T\mathbf{d}|k_2) \simeq \mathbf{Z}^T\mathbf{A}\mathbf{Z}$.

The conditional covariance matrix of $\mathbf{Z}^T\mathbf{d}$ can be estimated from the data using the formulae derived in the previous sections. To be precise, an unbiased estimate of the approximate conditional covariance matrix is $\hat{\operatorname{cov}}(\mathbf{Z}^T\mathbf{d}|k_2) = \mathbf{Z}^T\hat{\mathbf{A}}\mathbf{Z}$, where $\hat{\mathbf{A}}$ denotes the $n \times n$ symmetric matrix with elements

$$\hat{a}_{ij} = 2(\rho^{i,j})^2k_{22} + \rho^{i,i,j,j}k_4.$$

If $q = 1$, one sided significance levels may be computed by referring the standardized statistic,

$$T_1 = \mathbf{Z}^T(\mathbf{d} - k_2\boldsymbol{\rho})/(\mathbf{Z}^T\hat{\mathbf{A}}\mathbf{Z})^{1/2},$$

to standard normal percentiles. More generally, if $q > 1$, we may compute the quadratic form

$$T_1^2 = (\mathbf{d} - k_2\boldsymbol{\rho})^T\mathbf{Z}(\mathbf{Z}^T\hat{\mathbf{A}}\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{d} - k_2\boldsymbol{\rho}), \tag{11}$$

whose null distribution is approximately χ_q^2 provided only that ν is sufficiently large. Large values are taken as evidence of a systematic trend in variance.

An alternative model-based approach is to assume normality and to rely on standard asymptotic theory based on log-likelihood derivatives. Cook and Weisberg (1983) entertain the model $Y^i \sim N(\mu^i, \sigma^2\phi^i)$ independently for each i where

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \qquad \log\boldsymbol{\phi} = \mathbf{Z}\boldsymbol{\gamma}.$$

Since σ^2 is unknown, it may be assumed without loss of generality that the columns of $\mathbf{Z}$ sum to zero. It follows that the log-likelihood derivative with respect to $\boldsymbol{\gamma}$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\gamma} = \mathbf{0}$ is proportional to $\mathbf{Z}^T\mathbf{d}$, in part, justifying our use of this statistic in the previous calculations. To test the hypothesis $H_0: \boldsymbol{\gamma} = \mathbf{0}$, standard likelihood-based arguments lead to the scalar score statistic

$$T_2^2 = \mathbf{d}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{d}/(2\sigma^4), \tag{12}$$

where in practice, σ^4 is typically estimated by the maximum likelihood estimate $\hat{\sigma}^4$, rather than by k_2^2. The unbiased estimate k_{22} is not a component of the sufficient statistic under H_0.

Under normality, the asymptotic null distribution of T_2^2 is χ_q^2 independently of all unknown parameters and hence the statistic may be used to test H_0.

Without the normality assumption, however, the asymptotic distribution of T_2^2 is $\{1 + \kappa_4/(2\kappa_2^2)\}\chi_q^2$, so that large values of the statistic could be interpreted as due to excess kurtosis rather than to systematic trends in variance. For that reason, it seems preferable to use the statistic T_1 or T_1^2, which have built-in corrections for kurtosis. Note that to a first order of approximation, we have

$$T_2^2 \simeq (1 + k_4/(2k_{22}))T_1^2,$$

suggesting a simple multiplicative adjustment. For comparative purposes, denote by $\tilde{T}_1^2$, the approximation to T_1^2 obtained by applying the adjustment factor $c = (1 + k_4/(2k_{22}))^{-1}$ to T_2^2.

Another alternative procedure leading essentially to the same conclusions is based on the Wald test. This requires maximizing the normal-theory log likelihood to estimate the parameters (β, γ), and their asymptotic covariance matrix under H_0,

$$\begin{pmatrix} \kappa_2(\mathbf{X}^T\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & 2(\mathbf{Z}^T\mathbf{Z})^{-1} \end{pmatrix}.$$

This is correct under normality but not otherwise. More generally, the asymptotic covariance matrix of $(\hat{\beta}, \hat{\gamma})$ is given by

$$\begin{pmatrix} \kappa_2(\mathbf{X}^T\mathbf{X})^{-1} & \kappa_3/\kappa_2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1} \\ \kappa_3/\kappa_2(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} & (2 + \kappa_4/\kappa_2^2)(\mathbf{Z}^T\mathbf{Z})^{-1} \end{pmatrix},$$

which reduces to the diagonal form under normality. From these calculations, it appears that the Wald test based on $\hat{\gamma}$, namely $\hat{\gamma}^T\mathbf{Z}^T\mathbf{Z}\hat{\gamma}/2$, also needs to be adjusted by the factor c to account for nonnormality.

It is worth pointing out at this stage that standard asymptotic theory leading to the χ_q^2 approximation for (12) requires, in addition to normality, that $p = \dim(\beta)$ be fixed as $n \to \infty$. By contrast, the approximation to (11) is valid if, say, $p = n/2$ and the only requirement is that $\nu = n - p \to \infty$. This difference can be important particularly in the context of fractional factorial designs where the number of unknown regression parameters is often an appreciable fraction of n.

8.2. *Examples.*

EXAMPLE 1. In their paper, Cook and Weisberg discuss an example concerned with estimating tree "volume" as a function of tree height and diameter [Ryan et al. (1976, page 278)]. They entertain the model

$$\text{Volume}^{1/3} = \beta_0 + \beta_1\text{Height} + \beta_2\text{Diameter} + e.$$

Cook and Weisberg (Table 2, page 7) provide values of the score statistic for assessing the dependence of the error variance on diameter and height, both separately and jointly. As an example, the hypothesis of no (log-linear) dependence of variance on height results in $T_2^2 = 3.24$ (using $\hat{\sigma}^4$). This value is not

significant at the 5% level though graphical procedures proposed by these authors display "an obvious wedge shape."

These data are not extensive ($n = 31$), but we proceed to demonstrate the formulae of the previous sections. The standardized k-statistics for these data are

$$k_2 = 0.00686, \qquad \frac{k_3}{k_2^{3/2}} = -0.0814, \qquad \frac{k_4}{k_2^2} = -0.708, \qquad \frac{k_{22}}{k_2^2} = 0.955.$$

The optimal l-statistics are quite similar and not recorded here. Based on the above values, an adjustment factor of $c = 1.59$ should be applied to the score test, T_2^2, to account for nonnormality. In this particular example, this adjustment is enough to change an apparently insignificant dependence of variance on height ($T_2^2 = 3.24$), to a significant one ($\tilde{T}_1^2 = 5.15$). Using (11) directly we obtain $T_1^2 = 5.02$. This value better coincides with the informal graphical procedure than T_2^2 and seems to substantiate the claim that the dependence of variance on height is real. Of course, other models for these data, in particular multiplicative ones, would lead to different and possibly more easily interpretable conclusions.

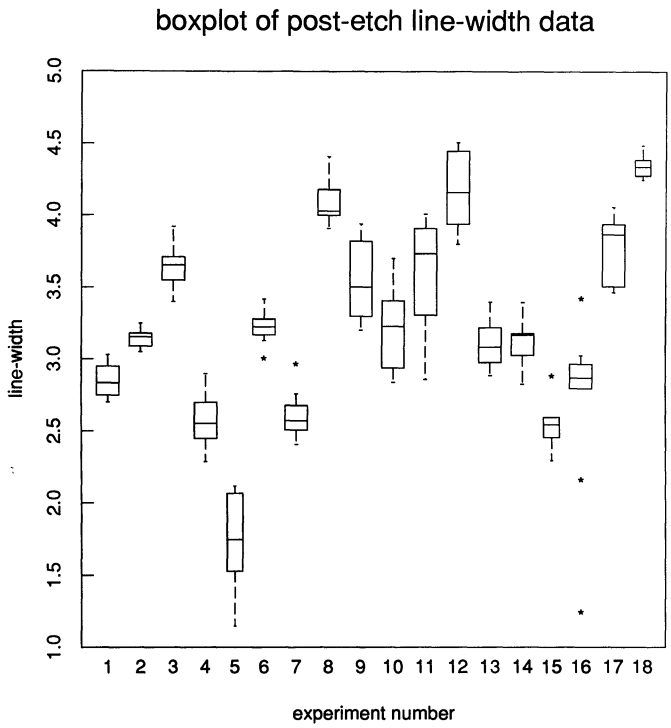

FIG. 2. *Boxplot of the post-etch line-width data. Units are micrometers. The width of each box is proportional to the square root of the number of replications. Experiments* 5, 15, *and* 18 *had only five replications; all others had ten.*

EXAMPLE 2. Phadke, Kackar, Speeney and Grieco (1983) present data on optimizing the production process of a certain integrated circuit. The experiment was designed to estimate the main effects of 8 factors on chip yield using 10 replications in each of 18 experimental settings. Slight imbalances occurred due to either broken or unavailable wafers during the time of the experiment, so that only 165 observations were available. Of the three measures of chip yield studied by these authors, we illustrate our method on the *post-etch line-width* data displayed graphically in Figure 2. The plot shows marked location effects, and possibly some dispersion effects. Experiment 16 stands out as being either highly variable or particularly outlier-prone.

The 18 experiments each provide an independent estimate of variability, s_i^2, which can be tested for homogeneity using Bartlett's test. This results in a value of 81.76 on 17 df, indicating either significant heterogeneity of variance or nonnormality. Figures 3 and 4 display probability plots to help sort out the ambiguity. Figure 3 is a normal probability plot of the within-experiment differences $Y^{ij} - \bar{Y}^i$. These differences are approximately iid if no dispersion effects are present. Under this assumption, the probability plot indicates that the data are apparently nonnormal, especially as regards the large negative outlier. Figure 4 is a normal probability plot of the $\log s_i^2$'s (base 10). These

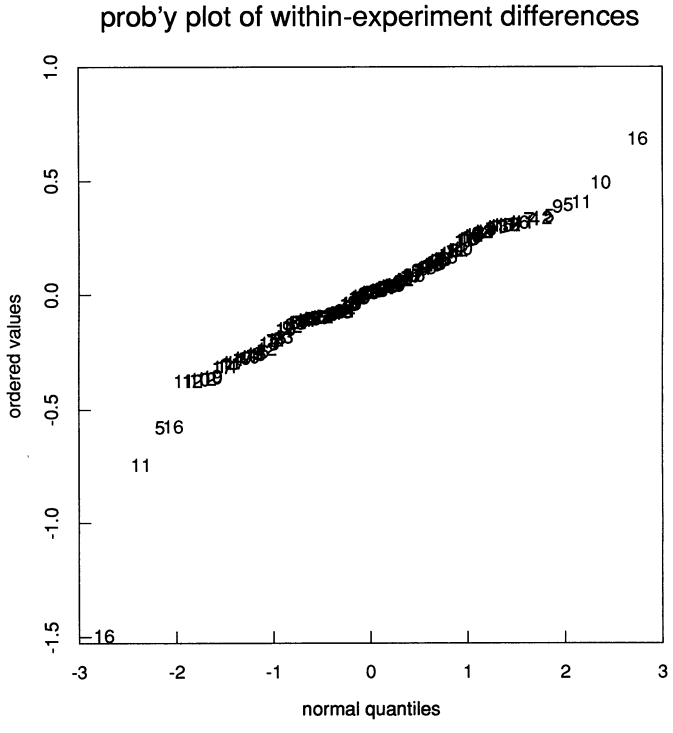

FIG. 3. *Normal probability plot of within-experiment differences* $Y^{ij} - \bar{Y}^i$. *The experiment number is used as the plotting character.*

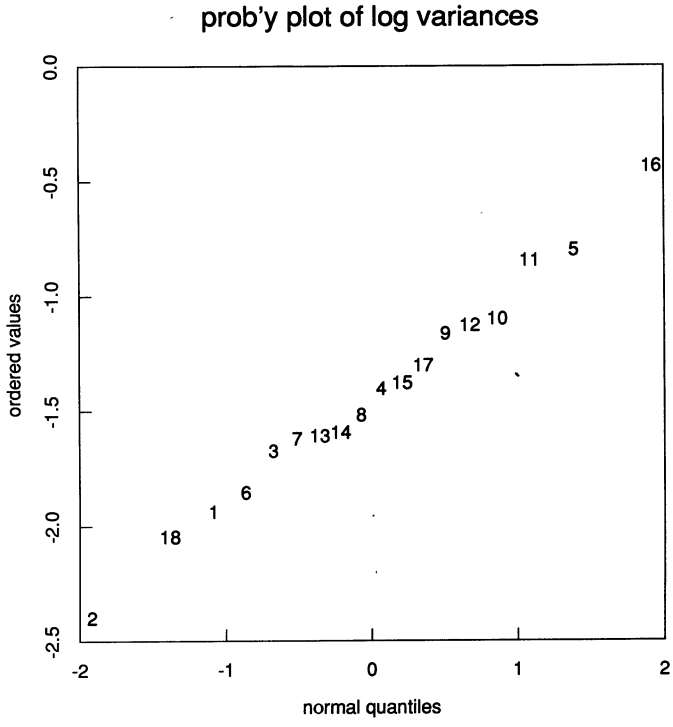

FIG. 4. *Normal probability plot of the logarithms (base 10) of the sample variances,* $\log s_i^2$. *The experiment number is used as the plotting character.*

values are iid if no dispersion effects are present. The plot is remarkably linear, suggesting that the sample variances are log-normal. Note, however that the slope is approximately 0.45 whereas normal-theory would suggest a slope of about 0.20. The graph therefore suggests that the kurtosis is about 8.0, which is consistent with the estimates presented below.

Consider fitting individual experiment means as location effects, so that the vector of squared residuals is $\mathbf{d} = (Y^{ij} - \overline{Y}^i)^2$. Let $\mathbf{Z}$ denote the matrix of contrasts corresponding to the experimental factors under study. The values of T_1^2 and T_2^2 based on $\mathbf{Z}^T\mathbf{d}$ are, respectively, 23.38 and 143.0 (using $\hat{\sigma}^4$), the former depending on the estimated standardized cumulants:

$$k_2 = 0.06629, \qquad \frac{k_3}{k_2^{3/2}} = -1.741 \qquad \frac{k_4}{k_2^2} = 10.04, \qquad \frac{k_{22}}{k_2^2} = 0.9265.$$

Due to the near balance in the experiment, the optimal l's are identical to k's to four significant digits. The k-statistics can also be used to adjust T_2^2, yielding $\tilde{T}_1^2 = 22.26$. It appears that tests based heavily on the normal likelihood function strongly support the existence of dispersion effects whereas tests based on our method indicate otherwise.

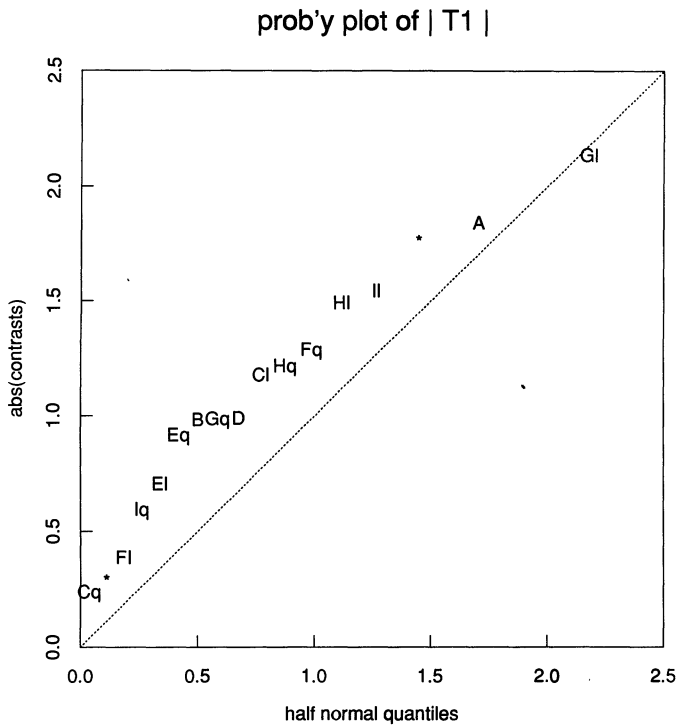

FIG. 5. *Half-normal probability plot of dispersion test statistics based on single degree-of-freedom contrasts. The contrast label is used as the plotting character. The two labeled "*" correspond to "error contrasts" that were not of interest to the experimenter. The second character of some labels indicates the linear or quadratic component of the three-level factors. The line superimposed on the plot has unit slope corresponding to the standard error of the standardized contrasts.*

Further insight into these data is possible by considering the 17 single degree-of-freedom contrasts $\mathbf{Z}_1, \ldots, \mathbf{Z}_{17}$. Each of the linear combinations $\mathbf{Z}_i^T\mathbf{d}$ leads to a standardized statistic $T_1(i)$. Figure 5 is a half-normal probability plot of $|T_1(i)|$. Although nonlinear, the plot is not indicative of signficant effects. Since for this example, $T_1^2(i) \approx 0.1648\ T_2^2(i)$ for $i = 1, \ldots, 17$, the half-normal probability plot of $|T_2(i)|$ is nearly identical to Figure 5. The important difference between the two, however, is that the half-normal plot of $T_1(i)$ should have approximately unit slope but that based on $T_2(i)$ has unknown slope depending on κ_4. For comparative purposes, the proportionality constant suggested by the adjustment factor c is 0.1557.

To assess the effect of the apparent outlier on our analysis, we removed it from the data and repeated the calculations. Qualitatively similar conclusions were obtained even though quite large differences in detail emerged. Bartlett's test decreased from 81.76 to 52.95, but still significant. The value of T_2^2 also decreased (to 55.97) but the value of T_1^2 increased (to 32.39). For the single degree-of-freedom contrasts $\mathbf{Z}_i$, we obtained $T_1^2(i) \approx 0.5644T_2^2(i)$ for $i =$

1, ..., 17. The increase in the constant of proportionality from 0.1648 to 0.5644 is due to the large change in the estimated standardized cumulants:

$$k_2 = 0.04980, \qquad \frac{k_3}{k_2^{3/2}} = -0.4509, \qquad \frac{k_4}{k_2^2} = 1.334, \qquad \frac{k_{22}}{k_2^2} = 0.9784.$$

The proportionality constant suggested by the adjustment factor c is 0.5946. The probability plot of dispersion effects again showed no significant factors, though the ordering of effects changed substantially.

We conclude that there are no significant dispersion effects in these data and that the apparent heterogeneity is due to nonnormality, and specifically, excess kurtosis as measured by κ_4. Furthermore, it appears that even though the estimated k-statistics are quite sensitive to outliers, the effect on the resulting test statistic, T_1, is substantially less. Based on a single example, however, we hesitate to claim that T_1 is more resistant to outliers than tests based on the normal likelihood function. If anything, the example reinforces the importance of supplementing the calculation of any test statistic with graphical displays.

REFERENCES

ANSCOMBE, F. J. (1961). Examination of residuals. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 1–36. Univ. California Press.

ANSCOMBE, F. J. (1981). *Computing in Statistical Science through APL*. Springer, New York.

ATIQULLAH, M. (1962). The estimation of residual variance in quadratically balanced problems and the robustness of the F-test. *Biometrika* **49** 83–91.

BICKEL, P. J. (1978). Using residuals robustly, I: Tests for heteroscedasticity, nonlinearity. *Ann. Statist.* **6** 266–291.

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

COOK, R. D. and WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70** 1–10.

FISHER, R. A. (1929). Moments and product moments of sampling distributions. *Proc. London Math. Soc.* (2) **30** 199–238. Reprinted (1972) as paper 74 in *Collected Papers of R. A. Fisher* (J. H. Bennett, ed.) **2** 351–354. Univ. of Adelaide Press.

MCCULLAGH, P. (1984). Tensor notation and cumulants of polynomials. *Biometrika* **71** 461–476.

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

PHADKE, M. S., KACKAR, R. N., SPEENEY, D. V. and GRIECO, M. J. (1983). Off-line quality control in integrated circuit fabrication using experimental design. *Bell System Tech. J.* **62** 1273–1309.

PUKELSHEIM, F. (1980). Multilinear estimation of skewness and kurtosis in linear models. *Metrika* **27** 103–113.

RYAN, B. F., JOINER, B. L. and RYAN, T. A. (1976). *Minitab Student Handbook*. Duxbury, North Scituate, Mass.

TUKEY, J. W. (1950). Some sampling simplified. *J. Amer. Statist. Assoc.* **45** 501–519.

TUKEY, J. W. (1956). Keeping moment-like sampling computations simple. *Ann. Math. Statist.* **27** 37–54.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637

ROOM 2C-264
AT & T BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974-2070