ROBUSTNESS TO NON-NORMALITY OF REGRESSION TESTS

G.E.P. Box and G.S. Watson*

This research was supported in part by the United States Navy through the Office of Naval Research, under Contract Nonr-1202 (17), Project NR 042 222 and in part by the United States Army, through the Office of Ordnance Research under Contract DA 36-034-ORD-2297, Project No. 1715. Reproduction in whole or in part is permitted for any purpose of the United States Government.

* - University of Toronto and Research Triangle Institute, University of North Carolina
ROBUSTNESS TO NON-NORMALITY OF REGRESSION TESTS

G.E.P. Box and G.S. Watson

1. SUMMARY

A number of statistical procedures involve the comparison of a "regression" mean square with a "residual" mean square using the normal-theory F distribution for reference. The use of the procedure for the analysis of actual data implies that the distribution of the mean square ratio is insensitive to moderate non-normality. Many investigators, in particular Pearson (1931), Geary (1947), Cayen (1950), have considered the sensitivity of this distribution to parent non-normality for important special cases and a very general investigation was carried out by David and Johnson (1951a, 1951b).

The principal object of this paper is to demonstrate the overriding influence which the numerical values of the regression variables have in deciding sensitivity to non-normality and to demonstrate the essential nature of this dependency.

The distribution of we first obtain a simple approximation to the regression F statistic in the non-normal case. This shows that it is "the extent of non-normality" in the regression variables (the x's), which determines sensitivity to non-normality in the observations (the y's). Our results are illustrated for certain familiar special cases. In particular the well-known robustness of the analysis of variance test to compare means of equal-sized groups and the notorious lack of robustness of the test to compare two estimates of variance from independent samples are discussed in this context. We finally show that it is possible to choose the regression variables so that, to the order of approximation we employ, non-normality in the y's is without effect on the distribution of the test statistic. Our results demonstrate the effect which the choice of experimental design has in deciding robustness to non-normality.

2. THE MEAN SQUARE RATIO A

Suppose that the response \( y_u \), observed at the \( u \)th set of levels \( x_{1u}, x_{2u}, \ldots, x_{pu} \) of \( p \) regression variables, may be represented by the response function

\[
y_u = \beta_0 + \beta_1 x_{1u} + \cdots + \beta_p x_{pu} + \epsilon_u \quad (u = 1, 2, \ldots, N)
\]

(1)

where the error \( \epsilon_u = y_u - E(y_u) \) is a random variable. The regression variables may be quantitative or merely indicator variables denoting presence or absence of a certain quality. We speak of the \( N \) values \( x_{1u} \) as the...
elements of the $i^{th}$ regression vector $\hat{\beta}_i$ and suppose that the model is set up so that the regression vectors are linearly independent. We denote the usual least squares estimates of $\beta_1, \ldots, \beta_p$ by $\hat{\beta}_1, \ldots, \hat{\beta}_p$ and write $c_{ij} = u_{ij}^T x_{iu} x_{ju}$ ($i, j = 1, \ldots, p$).

To shed light on the plausibility of $\hat{\beta}_1^*, \hat{\beta}_2^*, \ldots, \hat{\beta}_p^*$ as possible values of the coefficients, the ratio of mean squares

$$R_p^* = \frac{S_{p*}}{S_E/(N-p-1)}$$

may be calculated. The appropriateness of this ratio can be seen from the fact that the regression sum of squares

$$S_{p*} = \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij} (\hat{\beta}_i - \hat{\beta}_i^*) (\hat{\beta}_j - \hat{\beta}_j^*)$$

is a measure of overall discrepancy between the least squares estimates $\hat{\beta}_i$ and the contemplated values $\hat{\beta}_i^*$, while the residual sum of squares

$$S_E = \sum_{u=1}^{N} (y_u - \hat{y}_u)^2$$

with $\hat{y}_u = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_{iu}$ measures the internal consistency of the data independently of the choice of $\hat{\beta}_i^*$. Furthermore, provided that the assumed form of response function is adequate, and that the errors have equal variances and are uncorrelated one with another, the expected values of the component sums of squares are:

$$E(S_{p*}) = \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij} (\beta_i - \beta_i^*) (\beta_j - \beta_j^*) + \sigma^2$$

$$E(S_E) = (N-p-1)\sigma^2$$

If in addition the errors could be supposed to follow normal distributions, that is, if the joint distribution of the $\epsilon_u$ was a spherical normal distribution, then when $\hat{\beta}_i^* = \beta_i$ ($i = 1, 2, \ldots, p$) $S_{p*}$ would have an $F$ distribution with $\nu_1 = p$ and $\nu_2 = N-p-1$ degrees of freedom.

Since in the analysis of real data we do not know the precise distribution of the $\epsilon_u$, one would like to know under what circumstances the $F$ distribution still supplied an adequate approximation to the distribution of $S_{p*}$. In particular, as has been shown by Fisher (1936) and Schöffe (1953), when the approximation was adequate, the statistic $R_i^*$ could be used not only to test hypotheses concerning particular values $\beta_i^*$ of the coefficients, but also to supply interval estimates for any linear combinations whatever of the $\beta_i$'s.

For the "special" case $\beta_i^* = 0$ ($i = 1, \ldots, p$) we denote the mean
square ratio by \( \lambda \). Thus

\[
\lambda = \frac{S_0^2}{S_0^2/(n-p-1)}
\]

where

\[
S_0 = \sum_{i=1}^{p} \sum_{j=1}^{n} c_{ij} b_{ij} = \sum_{i=1}^{p} b_{ii} - X^2
\]

In what follows we consider only this case since the more general situation always can be reduced to it. For example, if we calculate \( \lambda \) from the constructed observations \( \tilde{y}_u = y_u - \beta_1^* x_{1u} - \beta_2^* x_{2u} - \cdots - \beta_p^* x_{pu} \), then \( \lambda \) is identical with \( \lambda_{pm} \). In what follows the reference of \( \lambda \) to the normal-theory \( \chi^2 \) distribution is referred to as the general regression test.

3. A FAMILIAR COMPARISON

That the answer to the question "What effect has non-normality on the general regression test?" is profoundly influenced by the nature of the \( x \) vectors may be demonstrated by a familiar comparison. By different choices of the \( x \) vectors, almost the same regression model can be made to reproduce on the one hand a test to compare means which is little affected by non-normality and on the other a comparison of variances test which is notoriously sensitive to non-normality.

3.1 Comparison of means

If in the general regression model of equation (1) we write

\[ N = (p+1)n, \quad x_{1u} = p/(p+1) \] for \( u = n(i-1) + t \) \( (i = 1, \ldots, p; \ t = 1, 2, \ldots, n) \) and \( x_{1u} = -1/(p+1) \) otherwise, then as is well known for \( \lambda \) we obtain the standard analysis of variance mean square ratio appropriate to the comparison of \( p+1 \) means each of \( n \) observations:

\[
\lambda_m = \frac{n \sum_{i=1}^{p+1} (\bar{y}_i - \bar{y})^2 / p}{n \sum_{i=1}^{p+1} \sum_{t=1}^{n} (y_{it} - \bar{y}_i)^2 / n - p - 1}
\]

where \( \bar{y}_i \) is the mean for the \( i \)th group of \( n \) observations and \( \bar{y} \) the overall mean.

3.2 Comparison of variances

Suppose there are \( N-1 \) observations and in the general regression model let us temporarily relax the provision that \( \beta^*_0 = 0 \) and suppose that the constant term \( \beta^*_0 \) is known. Now put \( x_{1u} = 1 \) when \( u = i \) \( (i = 1, 2, \ldots, p) \) and \( x_{1u} = 0 \) otherwise. Then the mean square ratio
\[ R_V = \frac{\sum_{u=1}^{p} (y_u - \beta_0)^2 / p}{\sum_{s=1}^{n-p-1} (y_s - \beta_0)^2 / (n-p-1)} \]  

This has the form of the standard test for the comparison of variances (the population means being known to equal \( \beta_0 \)) for two independent samples of size \( n_1 = p \) and \( n_2 = n-p-1 \).

The criteria \( R_n \) and \( R_V \) can each be obtained therefore as particular cases of the general regression criterion. Furthermore, on the usual assumptions, the null distribution of each criterion is the normal-theory \( F \) with \( \nu_1 = p \) and \( \nu_2 = n-p-1 \) degrees of freedom. It is well known however that whereas \( R_n \) has a distribution which is remarkably insensitive to departures from normality in the parent population this is not the case for \( R_V \).

Specifically, using a method which we discuss later in more detail, Box and Andersen (1955) have shown that in the non-normal situation \( R_n \) and \( R_V \) have distributions which may be approximated by \( F \) distributions with modified degrees of freedom. They show that \( R_n \) is approximately distributed as \( F \) with \( \nu_1 = \delta m \) and \( \nu_2 = \delta (n-p-1) \) degrees of freedom and \( R_V \) is approximately distributed as \( F \) with \( \nu_1 = \delta m \) and \( \nu_2 = \delta (n-p-1) \) degrees of freedom, where for moderate non-normality and moderate numbers of observations the \( \delta \)'s are approximately

\[ \delta^{-1} = 1 - \frac{1}{m} \Gamma y \]

\[ \delta^{-1} = 1 + \frac{1}{2} \Gamma'y \]  

The measures of kurtosis \( \Gamma y \) and \( \Gamma'y \) in these expressions are basically similar one to the other and take zero values when the distribution is normal. Explicitly*

\[ \Gamma y = \frac{k_y}{k_y^2} \]

\[ \Gamma'y = E\{c_{y'}\} \]

\[ \Gamma'y = E\{c_{y'}\} = \frac{n+1}{n-1} \frac{m_y}{m_2} - 3 \]

where \( k_y \) and \( k_y' \) are the usual \( k \) statistics for the whole sample of \( n \) observations and \( m_y \) is the \( q \)th moment about the mean

\[ m_y = \frac{\sum_{u=1}^{n} (y_u - \beta_0)^q + \sum_{s=1}^{n-p-1} (y_s - \beta_0)^q}{n-1} \]

for the sample of \( n-1 \) observations.

The insensitivity to non-normality shown by \( R_n \) arises because the corrective factor is of order \( n^{-1} \) whereas that for \( R_V \) is of order \( 1 \).

* Asymptotic expansions for these constants in terms of the standardised cumulants of the parent distributions are given by Box and Andersen (1955).
From our present point of view the example serves to show that different choices of the $x$'s in the general regression model can change sensitivity to non-normality by a factor as large as $N/2$.

We shall now investigate the specific nature of this dependence of sensitivity on the nature of the $x$ vectors. We shall be able to show that the corrective factor for the general regression test is to order $N^{-1}$.

$$e^{-1} = 1 + \frac{\Gamma X}{2N}$$  \hspace{1cm} (13)

where $\Gamma X$ is a measure analogous to $\Gamma$ of "non-normality" in the $x$'s. In the two special examples considered above $\Gamma X$ approaches respectively its smallest possible value of $-2$ and its largest possible value of $N$.

4. PERMUTATION MOMENTS OF $R$

In matrix notation our model becomes

$$y = \mathbf{1}b_o + \mathbf{X}x + \varepsilon$$  \hspace{1cm} (14)

where $\mathbf{1}$ is a column of $N$ unities, $\mathbf{X}$ the $N \times p$ matrix of the levels of the $p$ regression variables, $y$ the column of the $N$ observed responses, and $\varepsilon$ the column of coefficients $\varepsilon_1, \ldots, \varepsilon_p$. Then, since we suppose that $\mathbf{1}'\mathbf{1} = N$, the least squares estimators $b_o$ and $b$ of $b_o$ and $\beta$ are given by

$$b_o = \bar{y}, \quad b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$  \hspace{1cm} (15)

Writing $z$ for the column of $N$ deviations $z_u = y_u - y$ ($u = 1, \ldots, N$) and $i = \{i_{uv}\}$ for the symmetric idempotent matrix $i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$, we have $b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'z$, whence the regression and residual sums of squares defined in (3) and (4) are given by

$$S_0 = \mathbf{z}'i\mathbf{z}, \quad S_E = \mathbf{z}'(\mathbf{I} - i)i\mathbf{z}$$  \hspace{1cm} (16)

We are concerned with the distribution of $R = \{(N-p)S_0\}/\{pS_E\}$ when $\varepsilon = 0$, that is when

$$y = \mathbf{1}b_o + \varepsilon$$

Now if the error vector $\varepsilon$ has a probability distribution of the form $p(\varepsilon) = f(\varepsilon)$ and in particular if $\varepsilon$ is spherically normally distributed then $R$ has a $F$ distribution with $\nu_1 = p$ and $\nu_2 = \nu_{E}$-1 degrees of freedom. Equivalently

$$F = \frac{S_0}{S_0 + S_E} = \frac{S_0}{\pi'\pi}$$  \hspace{1cm} (17)

has a $F$ distribution with $\nu_1 = p$ and $\nu_2 = \nu_{E}$-1 degrees of freedom.
To study the distribution of $R$ on less specific assumptions we suppose only that the errors have a symmetric distribution, that is to say we suppose that $p(\xi)$ is a symmetric function of the elements of $\xi$. As a special case this includes the chief possibility of interest to us here that $p(\xi) = \prod_{u=1}^{N} p(\xi_u)$ where $p(\xi_u)$ is any distribution whatever. Now suppose a probability density to be associated with a particular $\xi$, then this same probability density is associated with every rearrangement of the elements of $\xi$. Also, since $\xi_u = y_u - \bar{y} = \xi_u - \bar{\xi}$ it follows that whatever be the probability density associated with a particular $\xi$ this same probability density is associated with every rearrangement of the elements of $\xi$. The distribution obtained by associating a probability $1/N$ with each of the possible $N!$ values of $f(\tilde{z})$ obtained by rearranging the elements $\xi$ is called the permutation distribution of $f(\xi)$.

The mean and variance of this permutation distribution we denote by $E_p f(\tilde{z})$ and $V_p f(\tilde{z})$ respectively.

Now suppose we define different samples to mean vectors $\tilde{z}$ which cannot be made identical by rearrangement of the elements, then if $E_{\tilde{z}}$ denotes the expected value taken over all different samples $\tilde{z}$ the overall moments $Ef(\tilde{z})$ and $Vf(\tilde{z})$ of $f(\tilde{z})$ are given by

$$\begin{align*}
Ef(\tilde{z}) &= \frac{E}{\tilde{z}} \left\{ E_p f(\tilde{z}) \right\} \\
Vf(\tilde{z}) &= \frac{E}{\tilde{z}} \left\{ V_p f(\tilde{z}) \right\}
\end{align*}$$

To approximate the distribution of $W$ for any error distribution of form $p(\xi) = \prod_{u=1}^{N} p(\xi_u)$ it is convenient for our purpose to find $E(W)$ and $V(W)$ by first finding the appropriate permutation moments and then taking their expected values over all samples.

In what follows it is helpful to express our results in terms of the power sums

$$\sum_{u=1}^{N} z_u^r = S_r$$

which, of course, remain constant under permutations of the $\xi_u$. We have in particular

$$S_1 = 0; \quad S_2 = z^t \bar{z} = S_0 + S_2; \quad W = S_0/S_2$$
4.1 \( E_p(w) \)

We have

\[ E_p(z_{u}^2) = S_2/N \]

\[ E_p(z_u z_v) = -S_2 / N(N-1) \]

whence

\[ E_p(z z') = \frac{S_2}{N(N-1)} \left[ N I - 11' \right] \]  \hspace{1cm} (19)

Now \( z'Mz = \text{tr}(Mz z') \) where \( \text{tr}(A) \) denotes the trace of the matrix \( A \).

Using the linearity of the trace and expectation operators we have

\[ \sum E_p(w) = E_p(S_0) = E_p(z'Mz) = E_p(\text{tr}(Mz z')) \]

\[ = \text{tr} \left\{ \mathbf{H} E_p(z z') \right\} \]

But \( \mathbf{H} = 0 \) and \( \text{tr}(H) = \text{tr} \left\{ \left[ z'(X'X)^{-1} z \right] \right\} = \text{tr} \left\{ z'(X'X)^{-1} z \right\} = p \). Hence

on substituting (19) in (20) we obtain finally

\[ \sum E_p(w) = p S_2 / (N-1), \]

that is

\[ E_p(w) = p / (N-1), \] \hspace{1cm} (21)

as is obtained on classical regression assumptions. In particular we see that \( E_p(w) = E_p(w) \) where \( E_p(w) \) is the expected value of \( w \) on the usual normal theory.

4.2 \( v_p(w) \)

In what follows summations and permutations expectations are taken over all combinations for which the subscripts are unequal. Thus \( E_p(z_{1}^a z_{2}^a z_{4}^b) \) means the average value of \( z_{u}^a z_{v}^b z_{w}^c \) taken over all permutations for which \( t, u, v \) and \( w \) are unequal. Similarly \( E_{12} \) means \( \sum_{t \neq u \neq v \neq w} \text{tr}(M'uv) \).

Also, since the \( x's \) and hence the \( M's \) can be regarded as remaining fixed whilst the combinations of the \( z's \) pass through all possible permutations, we have, for example,

\[ E_p(z_{1}^a z_{2}^b z_{4}^c) z_{3}^d H_{123} = E_p(z_{1}^a z_{2}^b z_{3}^c z_{4}^d) \text{ tr}(M'uv) \]

On squaring the expression

\[ S_0 = E_{12} z_{1}^2 + E_{12} z_{2}^2 + E_{12} z_{3}^2 \]

and taking expectations, we then have

\[ E_p(S_0^2) = E_p(z_{1}^2) E_{11} + E_p(z_{2}^2) (2 E_{12} + E_{11} M_{22}) \]

\[ + E_p(z_{3}^2) (4 E_{12} M_{23} + 2 E_{11} M_{23}) \]

\[ + E_p(z_{4}^2) (4 E_{12} M_{34}) + E_{12} M_{34} \] \hspace{1cm} (22)
Using David and Kendall's tables (1949)

\[
\begin{align*}
N_1 E_p(s_1^2) &= S_4 \\
N_2 E_p(s_1^2 s_2^2) &= S_2^2 - S_4 \\
N_3 E_p(s_1^2 s_2^2 s_3) &= 2S_4 - S_2^2 \\
N_4 E_p(s_1^2 s_2^2 s_3) &= -S_6 \\
N_6 E_p(s_1^2 s_2^2 s_3 s_4) &= 3S_2^2 - 6S_4
\end{align*}
\]

(23)

where \( N^{(r)}(r) = \sum_{i=0}^{r-1} \frac{N!}{i!(N-i)!} \).

Also, using the relations \( M_1 = 0, \ M_2 = M, \ trM = p, \) we find rather remarkably that each of the sums involving the elements of \( M \) can be expressed in terms of \( m = \sum_{u=1}^{N} \frac{M_u^2}{u} \). In fact

\[
\begin{align*}
EM_{12} &= p^2 - m \\
EM_{12} M_{23} &= p^2 - m \\
EM_{12} X_{23} &= 2m - p \\
EM_{12} M_{13} &= 2m - p^2 \\
EM_{12} M_{12} &= -m \\
EM_{12} X_{12} &= -2m + 2p + p^2
\end{align*}
\]

(24)

On substituting (23) and (24) in (22) and writing \( S_4 \) and \( S_2 \) in terms of Fisher's \( k \) statistics

\[
\begin{align*}
(N-1)k_2 &= S_2 \\
(N-1)^3 k_3 &= N(N-1)S_4 - 3(N-1)S_2^2
\end{align*}
\]

we have

\[
V_p(\delta) = \frac{E_p(s_2^2)/S_2^2 - \left\{E_p(\delta)\right\}^2}{E_p(\delta)}
\]

\[
= \frac{2p(N-p-1)}{(N+1)(N-1)} + \frac{k_2}{(N+1)(N-1)} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\}.
\]

We now put \( C_y = k_2/N \) and for reasons which will be apparent later we write

\[
C_x = \frac{N(N-1)(N+1)}{p(N-p-1)(N-3)} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\}
\]

(25)

so that we have finally

\[
V_p(\delta) = V(\delta) \left\{ 1 + \frac{(N-3)C_x C_y}{2N(N-1)} \right\}
\]

(26)

where \( V(\delta) = \frac{2p(N-p-1)}{(N+1)(N-1)^2} \) is the variance of \( \delta \) on the usual normal theory.
5. \( C_x \) As a Measure of "Non-Normality" of the \( x \)'s

We now show that the function \( C_x \) of the \( x \)'s is analogous with the function \( C_y \) of the \( y \)'s and is a measure of "non-normality" of the \( x \)'s. We first notice that \( M \) is invariant under any non-singular transformation \( \mathbf{x} = \mathbf{x}' \). For \( \mathbf{M} = \mathbf{X}^{-1} \mathbf{x} = \mathbf{X}^{-1} (\mathbf{x}' - \bar{x}' \mathbf{1}) = \mathbf{X}^{-1} \mathbf{W}^{-1} \mathbf{U}' \mathbf{W}^{-1} \mathbf{W}^{-1} \mathbf{1}' \mathbf{U} \mathbf{X}^{-1} \mathbf{x} = \mathbf{X}^{-1} (\mathbf{W}^{-1} \mathbf{U}' \mathbf{W}^{-1} \mathbf{1}) \mathbf{x} \). If we now regard \( n = \sum_{u=1}^{N} u \mathbf{w}_u \) as a function of \( m \mathbf{W}^{-1} \mathbf{x} \) of the elements \( x_{1u} \) of \( n \), we see that its value is unchanged when every element \( x_{1u} \) is replaced by the corresponding \( w_{1u} \).

Now let us choose \( T \) so that \( \mathbf{W}^{-1} \mathbf{M} \mathbf{W}^{-1} \) is diagonal with the \( 1 \)th element equal to \( \sum_{u=1}^{N} w_{1u}^2 \). Then

\[
m = \sum_{u=1}^{N} w_{1u}^2 \sum_{i=1}^{P} \left( \frac{w_{1i}^2}{\sum_{u=1}^{N} w_{1u}^2} \right)^2
\]

That is

\[
m = \frac{P}{1} \frac{1}{4} \left( \frac{t_{21}^0}{\sum_{i=1}^{P} t_{21}^i} \right)^2 + \frac{P}{1} \frac{1}{2} \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \frac{x_{1j}^i}{x_{11}^i}
\]

where \( t_{21}^0 = \frac{\sum_{u=1}^{N} w_{1u}^2}{\sum_{u=1}^{N} w_{1u}^4}, \quad t_{21}^i = \frac{\sum_{u=1}^{N} w_{1u}^4}{\sum_{u=1}^{N} w_{1u}^2}, \quad x_{1j}^i = \frac{\sum_{u=1}^{N} w_{1u}^4}{\sum_{u=1}^{N} w_{1u}^2} \quad (27) \)

Now, because of the invariant property of \( m \), equation (27) is still true if the corresponding power sums for the \( x \)'s (which we write as \( S_{21}^0, S_{21}^i, S_{21j}^i \) replace the power sums \( e_{21}^0, e_{21}^i, e_{21j}^i \) for the \( W \)'s.

Defining \( k \) statistics for the \( x \)'s in the usual way by \((N-1)k_{21}^0 = S_{21}^0, \quad (N-1)k_{21}^i = S_{21}^i, \quad (N-1)k_{21j}^i = S_{21j}^i \), we find after a little reduction that

\[
m = \frac{(N-1)(N-2)}{(N-1)^2} \left[ \frac{P}{1} \sum_{i=1}^{P} \left( k_{21}^i / (k_{21}^0)^2 \right)^2 \right] + \frac{P}{1} \sum_{i=1}^{P} \sum_{j=1}^{P} \left( k_{21j}^i / (k_{21j}^i)^2 \right)
\]

Substituting this expression in (25) we have finally

\[
C_x = \frac{(N-2)}{(N-1)^2} \frac{P}{1} \sum_{i=1}^{P} \left( k_{21}^i / (k_{21}^0)^2 \right)^2 + \frac{P}{1} \sum_{i=1}^{P} \sum_{j=1}^{P} \left( k_{21j}^i / (k_{21j}^i)^2 \right)
\]

We see that just as \( C_y = k_{21}^0 / k_{21}^2 \) is a measure of non-normality (specifically a measure of kurtosis) for the \( y \)'s so \( C_x \) can be regarded as its multi-variable analogue for the \( x \)'s. In particular:

(a) If \( p = 1 \) then \( C_x \) is the same function of the elements of the single vector \( x \) as \( C_y \) is of the elements of \( y \).

(b) The expected value of \( C_x \) is zero for samples from a normal population. This is so because the ratio of \( k \) statistics which \( C_x \)
contains are homogeneous functions of degree zero in the \( x \)'s. All such functions of normal variates are distributed independently of scaling statistics \( k_{ij}^2 \) \( (i = 1, 2, \ldots, p) \) and consequently the expected values of these ratios are the ratios of the expectations.

But for a normal distribution \( E(k_{ij}^2) = E(\{k_{ij}^2\}) = 0 \) \( (i \neq j = 1, 2, \ldots, p) \).

(c) \( C_{XX} \) is invariant under linear transformations of the \( x \)'s. Any sensible measure of multi-variate non-normality would clearly need to possess this property.

**Upper and lower bounds for \( C_{XX} \)**

We first show that

\[
\begin{align*}
P \leq m \leq \frac{N-1}{N} p
\end{align*}
\]

Because of the invariant property of \( m \) we may suppose, without loss of generality, that \( X'X = I \). Then

\[
\begin{align*}
N \sum_{u=1}^{N} M_{uu} &= \sum_{i=1}^{p} \sum_{u=1}^{N} X_{iu}^2 = p
\end{align*}
\]

But

\[
\begin{align*}
N \sum_{u=1}^{N} M_{uu}^2 &= \left( \sum_{u=1}^{N} M_{uu} \right)^2 / N \geq 0
\end{align*}
\]

Hence

\[
\begin{align*}
m &= \sum_{u=1}^{N} M_{uu}^2 \geq p^2 / N
\end{align*}
\]

Now suppose \( N-p \) further columns are added to \( X \), the first of which is \( X_{-1} \), to form an orthogonal matrix \( \bar{X} \). Then since the sums of squares of the elements of each row of \( \bar{X} \) is unity

\[
\begin{align*}
\bar{X}_{uu} &= \sum_{i=1}^{N} X_{iu}^2 \leq 1 - 1/N
\end{align*}
\]

whence

\[
\begin{align*}
\bar{X}_{uu} \leq M_{uu}(1 - 1/N)
\end{align*}
\]

But

\[
\begin{align*}
\bar{X}_{uu} = m
\end{align*}
\]

whence

\[
\begin{align*}
m = \sum_{u=1}^{N} M_{uu}^2 \leq \left( (N-1)/N \right) \sum_{u=1}^{N} M_{uu} = p(N-1)/N
\end{align*}
\]

Substituting these bounds in (25) we have finally

\[
\begin{align*}
-2 \leq \frac{N^2}{N-1} C_{XX} \leq N-1
\end{align*}
\]

We shall show later that the lower bound is actually obtainable. The upper bound is approached but cannot be attained in finite samples as is clear from the manner of its derivation.
6. APPROXIMATE DISTRIBUTION OF R

Case 1: Permutation Distribution of R

Following Pitman and Welch we now approximate the permutation distribution of W with a β distribution with degrees of freedom adjusted so as to have the correct mean and variance. If \( \nu_1 = p \) and \( \nu_2 = N-p-1 \) are the degrees of freedom of the β distribution appropriate on normal theory, then it is readily shown that the approximating distribution has degrees of freedom \( 1/\nu_1 \) and \( 1/\nu_2 \) where

\[
\frac{1}{\nu_1} = 1 + \frac{(N+1) a_1}{N-1-2a_1}, \quad \text{with} \quad a_1 = \frac{N-3}{2N(N-1)} \beta_\chi \beta_y
\]  

(31)

Equivalently the permutation distribution of R is approximated by an F distribution with degrees of freedom \( \nu_1 = 1/\nu_1 \) and \( \nu_2 = 1/\nu_2(N-p-1) \). In those cases where \( \beta_\chi \) and \( \beta_y \) are not simultaneously close to upper bounds we have to order \( N^{-1} \)

\[
\frac{1}{\nu_1} = 1 + \frac{\beta_\chi \beta_y}{2N}
\]  

(32)

Case 2: Distribution of R under general non-normality of y, \( \chi \) fixed

If we now take expectations of the permutation moments over all samples, then \( \beta_y \) in \( W(W) \) is replaced by \( E(\beta_y) \) which we denote by \( \Gamma_y \). Again using the β distribution approximation we have that if \( p(e) \) is any symmetric function of the elements of \( e \) and in particular if the \( e \)'s may be regarded as independent random drawings from any probability distribution whatever, R is distributed approximately as F with

\[
\nu_1 = 1/\nu_1 \quad \text{and} \quad \nu_2 = 1/\nu_2(N-p-1) \quad \text{degrees of freedom, where}
\]

\[
\frac{1}{\nu_2} = 1 + \frac{(N+1) a_2}{N-1-2a_2} \quad \text{and} \quad a_2 = \frac{N-3}{2N(N-1)} \beta_\chi \Gamma_y
\]  

(33)

or to order \( N^{-1} \)

\[
\frac{1}{\nu_2} = 1 + \frac{\beta_y \Gamma_y}{2N}
\]  

(34)

Case 3: Distribution of R under general non-normality of y and \( \chi \)

Finally we may suppose that the regression variables themselves are random variables distributed independently of the y's in some p-variate distribution and that it is the deviations of these variates from their sample means that is recorded in the matrix X.

Taking expectations of the moments over all realizations of X,
$C_X$ in (31) must be replaced by $E[C_X]$ which we denote by $\Gamma_X$.

Once again using the $\beta$ approximation, $R$ will be approximately distributed as $F$ with $\nu_1 = \delta_3 \rho$ and $\nu_2 = \delta_3 (N-\rho-1)$ degrees of freedom where

$$\delta_3^{-1} = 1 + \frac{(N+1)\delta_3}{N-2\rho} \quad \text{with} \quad \delta_3 = \frac{N-3}{2(N-1)} \Gamma_X \Gamma_Y$$  \hspace{1cm} (35)

or to order $N^{-1}$

$$\delta_3^{-1} = 1 + \frac{\Gamma_X \Gamma_Y}{2N} \hspace{1cm} (36)$$

7. ACCURACY OF THE APPROXIMATION

It should be noticed that the above approximations do not depend for their accuracy on the fitting of an arbitrary distribution to the first two moments of $W$. The distribution of $W$ is known to be given exactly by the approximation when in case 2 the observations $y$ are normally distributed and when in case 3 either the observations $y$ or the regression variables $x$, or both, are normally distributed. We might expect to be able to represent moderate departures from normality by suitable changes in the mean and variance of a system of curves which are of the right basic shape. Evidence that such a hope is justified are:

(a) For case 1 Pitman has shown for analysis of variance tests that except in very unlikely samples the third and fourth moments of $W$ agree fairly closely with those of the approximating $\beta$ distribution. We show later that the analysis of variance test for equal groups corresponds to the general regression test when $C_X$ attains its lower bound.

(b) For case 1 Box and Andersen showed by means of sampling experiments that the permutation distribution appropriate to the comparison of two independent variances was well represented by the approximation for distributions as non-normal as the rectangular and the double exponential. As we shall see, the test they considered can be very nearly reproduced in the general regression framework when $C_X$ approaches its upper bound.

(c) For case 2 Box and Andersen showed the close agreement between the results of Gayen (1950) and results obtained by this approximation in analysis of variance tests. This confirms in particular the
appropriateness of the general regression approximation at the lower bound of $C_\alpha$.

(d) For case 2 Box (1953), using an argument not employing permutation theory, obtained an $F$ approximation to the distribution of $R$ for the comparison of two independent sample variances. In this the degrees of freedom were modified by functions which to order $N^{-1}$ are identical with those given by the approximating $F$ distribution derived via permutation theory. This supports the essential validity of the approximation as $C_\alpha$ approaches its upper bound.

8. SPECIAL CASE OF THE GENERAL ONE-WAY CLASSIFICATION

We can readily check our formulae against those of Welch for the one-way classification analysis of variance. In the general regression model we arrange that the $u$ th element $x_{iu}$ of the $i$ th $x$ vector is $1 - \frac{n_i}{N}$ when $x_u$ falls in the $i$ th group and $\frac{n_i}{N}$ otherwise. As is well known the general regression test then reduces to the usual "one way classification" analysis of variance test for the comparison of $p+1$ means. Then

$$S_0 = x'Mx = X'MX = \sum_{i=1}^{p+1} n_i y_i^2 - y^{'2}$$

where the overall sample mean is $\bar{y}$ and the sample mean for $n_i$ observations in the $i$ th group is $\bar{y}_i$.

Now if $y_u$ is any one of the $n_i$ observations in the $i$ th group then the corresponding diagonal element $K_{uu}$ of $K$ is $n_i^{-1} - N^{-1}$. Hence

$$K_{uu} = \sum_{i=1}^{p+1} n_i (n_i^{-1} - N^{-1})$$

$$= \sum_{i=1}^{p+1} n_i - (2p+1)/N$$

Substituting this expression in (21) gives the value obtained by Welch for $V_1(x)$ and the corresponding $F$ approximation given by Box and Andersen.

6.1 Equal Groups

If the number of observations is the same in every group so that $n_i = n = N/(p+1)$ ($i = 1, 2, \ldots, N$), we have using equation (31)
\[ m = \frac{y^2}{N} \quad \text{and} \quad \sum_{n=1}^{N-2} c_n = -2 \]  \hspace{1cm} (39)

The lower bound of the inequality (30) is thus actually attained for the equal groups analysis of variance test.

Substituting the result in (31) we obtain correctly

\[ s^{-1} = 1 - \frac{1}{N} \left( \frac{N+N-1}{N-1+2y/N} \right) c_y \]

or to order \( y^{-1} \)

\[ s^{-1} = 1 - c_y/N \]  \hspace{1cm} (40)

The corresponding result with \( \gamma_y = E(c_y) \) replacing \( c_y \) provides the appropriate correction factor for case 2 exemplifying the approximate effect of parent non-normality.

### 8.2 Very unequal groups

It is not possible to reproduce exactly the test for the comparison of two variances from the present general regression set-up. As we have already mentioned in Section 3.2 we can however reproduce such a test with a slightly modified model in which the overall mean \( \beta_0 \) is supposed known.

The permutation approximation given by Box and Andersen for the set-up of Section 3.2 with \( n_1 + n_2 = N-1 \) observations is

\[ s^{-1} = 1 + \frac{1}{2} \left( \frac{N+y}{N-2+c_y} \right) c_y \]

or to order \( y^0 \)

\[ s^{-1} = 1 + \frac{1}{2} c_y \]  \hspace{1cm} (41)

where \( c_y \) is defined in equation (12). The corresponding result which gives the effect of non-normality on this comparison of variances is obtained as before by replacing \( c_y \) with \( \gamma_y = E(c_y) \). It is also possible by a different modification in which both the group means are eliminated to reproduce the usual test to compare two variances when their group means are estimated from the samples. With the present unmodified set-up we come closest to reproducing the test for the comparison of two variances by selecting the \( i \)th vector to have an element \( x_{1u} = \frac{N-1}{N} \) when \( u = i \) and \( x_{1u} = -\frac{1}{N} \) otherwise. This then corresponds to the analysis of variance test discussed above with one observation in each of \( p \) groups and the remaining \( N-p \) in the remaining groups. After a little manipulation we obtain for this case
\[ R = \left( \sum_{u=1}^{p} (y_u - \bar{y}')^2 + \frac{p(N-p)}{N} (\bar{y}' - \bar{y})^2 \right) / p \]
\[ \text{where } \bar{y}' \text{ is now used for the sample mean of the first } p \text{ observations and } \bar{y}'' \text{ for the sample mean of the last } N-p \text{ observations.} \]

The usual test criteria to compare the variances of two samples of \( p \) and \( N-p \) observations would be

\[ R' = \left( \frac{\sum_{u=1}^{p} (y_u - \bar{y}')^2}{\sum_{s=p+1}^{N} (y_s - \bar{y}'')^2} \right) / (p-1) \]
\[ \text{which differs from } R \text{ above only in that the latter contains a single extra comparison which contrasts the mean } \bar{y}' \text{ of the first } p \text{ observations with the overall mean } \bar{y}. \]

For this extreme case of very unequal groups which nearly reproduces the comparison of variances test we have

\[ m = \frac{p}{N} \left( \frac{N-1}{N-p} \right) \text{ and } C_x = N-1 - \frac{N-p}{N} \]
\[ \text{Provided that the ratio of } p \text{ to } N \text{ was small, } C_x \text{ would approach its upper bound of } N-1 \text{ quite closely.} \]

If we write \( \hat{N} \) for \( N-1 - \frac{N-p}{N} \) the modifying factor appropriate to this case is

\[ \delta^{-1} = 1 + \frac{1}{4} \frac{N-1}{\hat{N}} \left( \frac{\hat{N}}{\hat{N} - 1 - \frac{N-p}{N}} \right) C_y \]
\( \text{which is very similar to the corresponding expression in equation (41)} \)
\( \text{and once more to order } N^0, \ \delta^{-1} = 1 + \frac{1}{4} C_y. \)

3.3 Group sizes which approximately nullify the effect of non-normality

We have seen that particular choices of group sizes can be made so that \( C_x \) approaches a lower value of -2 and an upper value of 0, giving rise to proportionate corrections. Provided the number of observations is not very small, the slight corrective increase in the degrees of freedom at the lower extreme is usually of little concern but the considerable corrective decrease at the upper extreme is much more serious. Consideration of equations (31), (33) and (35) shows that in general if in cases 1 and 2 we choose \( X \) so that \( C_x = 0 \) or in case 3 sample
from a population in which \(\frac{\sum x_i}{n} = 0\), the corrective factor supplied by our approximations are all zero irrespective of the \(y_i\)'s or their distribution. One way of exploiting this fact in experimental design theory has been noted by Box (1952).

Returning, now, to the one-way classification analysis of variance we see from equation (33) that, if we choose the group sizes \(n_i\) so that

\[
\frac{\sum n_i-1}{n} \sum n_i - (p+1)^2 = \frac{2n(n-p-1)}{n+1}
\]

then \(C_X = 0\) and the correction term is zero. Rather surprisingly therefore the effect of non-normality as measured by our approximation is not smallest for equal group sizes.

As an example, suppose there are just two groups of size \(n_1 = rn\) and \(n_2 = (1-r)n\). Substituting in (46) we obtain

\[
r = \frac{1}{2} \left[ 1 \pm \frac{\sqrt{-2}}{\sqrt{n}} \right]
\]

for the optimum ratio of sub-group sizes. If \(n = 12\), for example, then approximately the optimal group sizes are 9 and 3. For large \(n\) the optimal sizes are approximately in the ratio 4:1.

At first sight the idea that unequal group sizes could produce less sensitivity to non-normality seemed sufficiently surprising as to be almost unacceptable. In fact as we show in the next section, it is not difficult to explain it. The result itself may be confirmed independently by study of the results of Gayen (1950). This author obtained the exact distribution of \(R\) for the one-way classification analysis of variance test for a parent population expressed by an Edgeworth series. His method of derivation is, of course, quite different from that used here. His corrective factor for kurtosis is proportional to a quantity which he denotes by

\[
\frac{\nu_{22}}{2} = \frac{2\nu_{12} - (k^2 - k'^2)(\nu_1 + \nu_2 + 2)}{3(\nu_1 + \nu_2 + 1)(\nu_1 + \nu_2 + 2)}
\]

In our notation \(\nu_1 = p\), \(\nu_2 = n-p-1\), \(k = (p+1)\), \(k' = (p+1)^2 - \frac{p+1}{n} \sum_{i=1}^{n} n_i\). Making the necessary substitution we see that when (46) is satisfied Gayen's quantity \((\nu_{22})\) is zero, providing verification of our result.

We do not of course suggest that one would deliberately seek unequal sample sizes to lessen the effect of non-normality in a test to compare
means. The reduction in precision with which comparison among the means could be made and the increase in sensitivity to variance inequalities which would result would certainly not be worth the small increase in robustness to non-normality.

9. ROBUSTNESS DETERMINED BY "NORMALITY" IN THE $x_i$'s

We shall conduct the following discussion in terms of the situation (cases 2 and 3) where the error vector $\epsilon$ is drawn from any symmetric distribution. We have seen that our modifying factor involves a measure of non-normality in the $y_i$'s multiplying an analogous measure of "non-normality" in the $x_i$'s. That such factors would be involved symmetrically is to be expected from geometric considerations. The criterion $R$ is a function of the angle between the observation vector and the plane of the $x$ vectors and, as was first shown by Fisher, will follow its normal theory distribution if the $y_i$'s or the $x_i$'s or both are normally distributed. The multiplicative characteristic shows how non-normality in the $y_i$'s is magnified or diminished depending on whether the $x_i$'s are "normal-looking" or not.

The effects can be understood intuitively by means of particular examples. Consider the analysis of variance test for two equal groups. In this case the single vector $x_1$ has for its elements $n/2$ values equal to $-\frac{1}{2}$ and $n/2$ values equal to $+\frac{1}{2}$. The distribution of individual $x_i$'s is like that shown in figure 1(a) (for $N = 20$) and represents the most "platykurtic" distribution possible, for which the value $C_X$ tends to its lower limit of $-2$. The set-up involving very unequal groups, discussed in Section 8.2, comes closest to representing the comparison of two independent variances. The $x$ vectors then each have one value equal to $(N-1)/N$ and the remaining $N-1$ values equal to $-1/N$. The distribution is that of the full circles shown in figure 1(c). This distribution has the same measure of kurtosis as when its mirror image, shown by open circles, is added. This represents the most leptokurtic distribution possible and the value of $C_X$ tends to its upper limit, $N$. The full circles in figure 1(b) show a distribution in which $1/4$ of the observations are set a distance $3/4$ from the origin and the remaining $3/4$ of the observations are set a distance $-1/4$. This distribution
Figure 1. Some distributions of the elements of $x$ for $N = 20$. 

(a) $C_x$ approaches $-2$ 

(b) $C_x$ approaches $0$ 

(c) $C_x$ approaches $N$
is close to that expected to minimize the effect of non-normality in the y's. Again the mirror image distribution of open circles is added. We see that in this example the distribution of the x's is doing its best to approximate a normal curve which accounts for the resulting insensitivity to non-normality in the y's.

10. CONCLUSION

Our results may be summarised in the simple statement that sensitivity to non-normality in the y's is determined by the extent of the "non-normality of the x's". The small effect in one direction experienced with the equal groups analysis of variance test and the much larger effect in the opposite direction found in the test for the comparison of independent variances provide extremes of sensitivity within which the sensitivity of the general test will be found. In the analysis of data arising from experimental designs such as factorials, a small and usually unimportant degree of sensitivity characteristic of the equal groups analysis of variance may be expected. With data in which the x's themselves are drawn from near normal distributions an even smaller degree of sensitivity is to be expected. Tests which employ x vectors in which one or two elements are very different in magnitude from the remainder may be expected to show much greater sensitivity to non-normality in the y's. In addition to the tests for comparing variances, certain tests concerned with outliers and with missing observations will show this greater sensitivity. One would expect that the usual normal theory multivariate overall regression tests will have analogous corrective factors.
REFERENCES

Box, G.E.P. (1952), Biometrika, 29, 49-57.
Box, G.E.P. (1953), Biometrika, 40, 318-335.
Pearson, E.S. (1931) Biometrika, 22, 114-133.
Scheffe, H. (1953), Biometrika, 40, 87-104.