AN ABSTRACT OF THE THESIS OF

JAMES DONALD PUGH for the Master of Science
(Name) (Degree)

in Statistics presented on May 12, 1967
(Major) (Date)

Title TESTING EQUALITY OF TWO REGRESSIONS

Abstract approved [redacted]
Donald Guthrie, Jr.

This thesis addresses the problem of deciding whether or not two disjoint random samples that are fitted by the same regression equation emanate from the same parental population. Two situations are considered. The first situation is where the two disjoint samples are individually fitted with the regression equation. The second situation is where one sample and the union of the two disjoint samples are fitted.

For the first situation three cases are considered. Each case involves different assumptions about the residual variances for the populations from which the samples are drawn.

For the second situation two cases are considered. The first case requires knowledge of all the sums of squares and factors in the normal equations associated with the regression. The second case requires the knowledge of two sums of squares and results in a conservative approximation to an F statistic used for testing a hypothesis about the equality of the two coefficient vectors.

Testing Equality of Two Regressions

by

James Donald Pugh

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

June 1967

APPROVED:

[redacted]

Associate Professor of Statistics

In Charge of Major

[redacted]

Acting Chairman of Department of Statistics

[redacted]

Dean of Graduate School

Date thesis is presented May 12, 1967

Typed by Carol Baker for James Donald Pugh

ACKNOWLEDGMENT

The author is indebted to many; to his wife and children for their patience and understanding, to Dr. E. R. Anderson for his faith and encouragement, and to the U.S. Navy Electronics Laboratory for the opportunity to pursue a graduate program.

Thanks are warmly extended to the author's major professor, Dr. Donald Guthrie, Jr., and to the staff of the Statistics and Oceanography departments for their part in making the study at Oregon State University a worthwhile experience.

ERRATA

For the thesis "Testing Equality of Two Regressions" by James D. Pugh.

Page 8, line 10:

$$\lambda = \frac{1}{2} (B-C)' [\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}] (B-C) \quad \text{should be}$$

$$\lambda = \frac{1}{2} (B-C)' [\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}]^{-1} (B-C).$$

Page 10, line 12:

$$S = [(\hat{B}-\hat{C})-(B-C)]'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}][(\hat{B}-\hat{C})-(B-C)] \quad \text{should be}$$

$$S = [(\hat{B}-\hat{C})-(B-C)]'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}]^{-1}[\hat{B}-\hat{C})-(B-C)]$$

Page 12, line 5:

$[(n_1-p)\hat{\sigma}_1^2 + (n_2-p)\hat{\sigma}_2^2/\sigma^2]$ should be $[(n_1-p)\hat{\sigma}_1^2 + (n_2-p)\hat{\sigma}_2^2]/\sigma^2$.

Page 14, line 5:

$$\frac{\partial \log L}{\partial b_i} = \frac{1}{2\sigma_1^2} \sum_{m=1}^{n_1} 2(w_m - \sum_{k=1}^{p} u_{mk} b_k)(-u_{mi}) \quad \text{should be}$$

$$\frac{\partial \log L}{\partial b_i} = \frac{-1}{2\sigma_1^2} \sum_{m=1}^{n_1} 2(w_m - \sum_{k=1}^{p} u_{mk} b_k)(-u_{mi})$$

Page 17, line 6:

$$\hat{D} = \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} & 0 \\ 0 & \frac{2(n_1-p)\hat{\sigma}_1^2}{n_1^2} \end{bmatrix} \quad \text{should be} \quad D = \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} & 0 \\ 0 & \frac{2(n_1-p)^2(\hat{\sigma}_1^2)}{n_1^3} \end{bmatrix}$$

Page 17, line 9:

$$\hat{D} = \begin{bmatrix} \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_2-p)\hat{\sigma}_2^2}{n_2^2} \end{bmatrix} \text{ should be } D = \begin{bmatrix} \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_2-p)^2(\hat{\sigma}_2^2)^2}{n_2^3} \end{bmatrix}$$

Page 18, line 1:

$$\begin{bmatrix} \hat{B}-\hat{C} \\ \hat{\sigma}_1^2-\hat{\sigma}_1^2 \end{bmatrix}' \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} + \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_1-p)\hat{\sigma}_1^2}{n_1^2} + \frac{2(n_2-p)\hat{\sigma}_2^2}{n_2^2} \end{bmatrix} \begin{bmatrix} \hat{B}-\hat{C} \\ \hat{\sigma}_1^2-\hat{\sigma}_2^2 \end{bmatrix}$$

should be

$$\begin{bmatrix} \hat{B}-\hat{C} \\ \hat{\sigma}_1^2-\hat{\sigma}_2^2 \end{bmatrix}' \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} + \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_1-p)^2(\hat{\sigma}_1^2)^2}{n_1^3} + \frac{2(n_2-p)^2(\hat{\sigma}_2^2)^2}{n_2^3} \end{bmatrix}^{-1} \begin{bmatrix} \hat{B}-\hat{C} \\ \hat{\sigma}_1^2-\hat{\sigma}_2^2 \end{bmatrix}.$$

Page 18, line 4:

$$\lambda = \frac{1}{2} \begin{bmatrix} B-C \\ \sigma_1^2-\sigma_2^2 \end{bmatrix}' \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} + \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_1-p)\hat{\sigma}_1^2}{n_1^2} + \frac{2(n_2-p)\hat{\sigma}_2^2}{n_2^2} \end{bmatrix} \begin{bmatrix} B-C \\ \sigma_1^2-\sigma_2^2 \end{bmatrix}.$$

should be

$$\lambda = \frac{1}{2} \begin{bmatrix} B-C \\ \sigma_1^2-\sigma_2^2 \end{bmatrix}' \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} + \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_1-p)^2(\hat{\sigma}_1^2)^2}{n_1^3} + \frac{2(n_2-p)^2(\hat{\sigma}_2^2)^2}{n_2^3} \end{bmatrix}^{-1} \begin{bmatrix} B-C \\ \sigma_1^2-\sigma_2^2 \end{bmatrix}$$

Page 19, line 3:

$$\chi^2_{1-\alpha}(p+1) = \begin{bmatrix} (\hat{B}-\hat{C}) - (B-C) \\ \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2) - (\sigma_1^2-\sigma_2^2) \end{bmatrix}' \hat{D} \begin{bmatrix} (\hat{B}-\hat{C}) - (B-C) \\ \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2) - (\sigma_1^2-\sigma_2^2) \end{bmatrix}$$

should be

$$\chi^2_{1-\alpha}(p+1) = \begin{bmatrix} (\hat{B}-\hat{C}) - (B-C) \\ \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2) - (\sigma_1^2-\sigma_2^2) \end{bmatrix}' D \begin{bmatrix} (\hat{B}-\hat{C}) - (B-C) \\ \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2) - (\sigma_1^2-\sigma_2^2) \end{bmatrix}$$

TABLE OF CONTENTS

TESTING EQUALITY OF TWO REGRESSIONS

I. INTRODUCTION

A study of sea-surface temperatures off the Pacific Coast of the United States gave rise to some problems associated with the comparison of two subsamples based on certain sufficient statistics for the combined sample and one of the subsamples. The coefficients of a linear regression equation were determined by the method of least squares from a set of sea-surface temperature observations extending westward from the Pacific Coast to a given meridian, within one degree of latitude. From oceanographic considerations it was to be expected that observations taken over the continental shelf would come from a different parental population than the observations taken in the deep ocean. A regression equation was fitted to all of the data, and at a later date to the deep ocean data only. The output of the computer program which was used included the coefficients, the total, regression, and residual sums of squares, and the deviations of the estimates based on the regression equation using the estimated coefficients and the initial observations from which the coefficients were estimated.

After these calculations the research was inactive for a period of time. Upon reactivation, the hypothesis that the sea-surface temperature observations taken in the continental shelf area and the sea-surface temperature observation taken in the deep ocean come from the same parental population became of interest. If

$$y_i = \sum_{j=1}^{p} x_{ij}b_j + \varepsilon_i \quad \text{for} \quad i = 1(1)n_1$$

is the set of regression equations for a fit using the deep ocean data and $y_i = \sum_{j=1}^{p} x_{ij}c_j + \varepsilon_i$ for $i = n_1+1(1)n$ is the set of regression equations for a fit using the shelf area data, n being the total number of observations and n_1 being the number of observations taken in the deep ocean, then the usual procedure would be to test the hypothesis that $b_j = c_j$ for $j = 1(1)p$. Clearly, if any $b_j \neq c_j$, then the data taken in the deep ocean and the data taken over the continental shelf do not come from the same population.

However, to test this hypothesis additional computations are needed, since the continental shelf area data had not been fitted with the regression equation during the original analysis. During the time the research was inactive, the regression analysis computer program became obsolete and the raw data were mislaid, so additional computations were not readily obtainable.

For practical reasons, it was deemed necessary to utilize the computations that already existed, rather than compute the estimates of the coefficients for the shelf area data. It was felt that if the respective coefficients for the two fits (total sample, a_j, and deep ocean sample, b_j) were not equal it was due to the influence of the continental shelf area data which were included in the total sample fit and not included in the deep ocean fit. It was recognized that the estimates of the corresponding coefficients, $\hat{a}_j$ and $\hat{b}_j$ for $j = 1(1)p$, might be different due to random variations alone, hence a statistic was needed to test the hypothesis that the respective coefficients derived from the two fits of the data were equal. Furthermore, the statistic should utilize only the computations that already existed; that is, no new computations would be made.

The statistical procedures for such a situation are the topic of this thesis. The succeeding chapters will develop statistics which may be used to test hypotheses concerning the parental population from which the observations were drawn, and to construct confidence regions for the parameters associated with the regression equation, under varying assumptions about the data and availability of computations. The final statistic to be discussed is a conservative approximation to an F statistic which utilizes only the limited computations as mentioned previously.

II. NOTATION AND BACKGROUND INFORMATION

Let $Y = (y_1\ y_2 \cdots y_n)'$ be an $n \times 1$ vector of regressor variables. Let Y be partitioned so that $Y' = [W', Z']$ where W is $n_1 \times 1$ and Z is $n_2 \times 1$ and $n_1 + n_2 = n$. Henceforth, Y will be referred to as the total sample or Y-sample, W as the W-sample, Z as the Z-sample, and W and Z collectively as the disjoint or independent subsamples. Using this partitioning of Y it is clear that Y-sample is the combination of the disjoint W- and Z-samples. Let

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

be an $n \times p$ matrix of regressand variables, and let X be partitioned so that $X' = [U', V']$ where U is $n_1 \times p$ and V is $n_2 \times p$. Let $\varepsilon = (e_1\, e_2 \cdots e_n)'$ be an $n \times 1$ vector of random variables; let ε be partitioned so that $\varepsilon' = [\eta', \delta']$ where η is $n_1 \times 1$ and δ is $n_2 \times 1$. Let $A = (a_1\ a_2 \cdots a_p)'$, $B = (b_1\ b_2 \cdots b_p)'$ and $C = (c_1\ c_2 \cdots c_p)'$ be $p \times 1$ vectors.

The regression equations for the total sample may be written in matrix form as $Y = XA + \varepsilon$. If Y and X are partitioned as indicated, then the regression equations for the data in these partitions may be written as $W = UB + \eta$ and $Z = VC + \delta$.

It is assumed throughout that $n_1 > p$ and $n_2 > p$, and that the matrices $X'X$, $U'U$, and $V'V$ have full rank p.

The discussion in this thesis relies heavily on the theory of linear models of full rank. Graybill (2, p. 110-114) summarizes much of this theory in his Theorem 6.1 which is partially restated here for convenience.

<u>Theorem 2.1:</u> If $Y = XA + \varepsilon$ is a linear model of full rank and if ε is distributed $N(0, \sigma^2 I)$, the estimators

$$\hat{A} = (X'X)^{-1}Y'Y$$

and

$$\begin{aligned}\hat{\sigma}^2 &= Y'(I-X(X'X)^{-1}X')Y/(n-p)\\ &= (Y'Y-\hat{A}'X'Y)/(n-p)\\ &= (Y-X\hat{A})'(Y-X\hat{A})/(n-p)\end{aligned}$$

have the following properties:

(1) $\hat{A}$ is distributed $N(A, \sigma^2(X'X)^{-1})$,

(2) $(n-p)\hat{\sigma}^2/\sigma^2$ is distributed as χ^2_{n-p},

and

(3) $\hat{A}$ and $\hat{\sigma}^2$ are independent.

The use of a circumflex ($\wedge$) over a parameter indicates an estimator for that parameter and, thus, a random variable. For example, $\hat{A}$ is an estimator of A. Since least-squares estimators and maximum-likelihood estimators are equivalent when the residuals (ε) of a regression equation are distributed multivariate normal (5, p. 348), the circumflex will be interpreted to mean either one of these estimators in that situation but only the least-squares estimator when the residuals are not multivariate normal.

The matrix equations $\hat{A} = (X'X)^{-1}X'Y$ when written in the form $X'X\hat{A} = X'Y$ are known as the normal equations (2, p. 111).

III. TESTS AND CONFIDENCE REGIONS FOR THE DIFFERENCE BETWEEN TWO COEFFICIENT VECTORS FROM INDEPENDENT SAMPLES

The simplest case is the situation where both residual variances are known exactly. This is of theoretical interest only since such knowledge would never be available in practice, but the derivation of tests is informative in that it shows the underlying structure of the problem.

Let the regression equations for the two samples be $W = UB + \eta$ and $Z = VC + \delta$, where η is $N(0, \sigma_1^2 I)$ and δ is $N(0, \sigma_2^2 I)$. By Theorem 2.1 $\hat{B} = (U'U)^{-1}U'W$, $\hat{C} = (V'V)^{-1}V'Z$, $\hat{B}$ is $N(B, \sigma_1^2(U'U)^{-1})$, and $\hat{C}$ is $N(C, \sigma_2^2(V'V)^{-1})$. Since $\hat{B}$ and $\hat{C}$ are computed from independent samples, they are independent. Consequently, the moment generating function of $\hat{B}-\hat{C}$ will be the product of the moment generating functions of $\hat{B}$ and $-\hat{C}$. Letting $T = (t_1\ t_2 \cdots t_p)'$ we have (3, p. 343-347) these moment generating functions:

$$m_{\hat{B}}(T) = \exp\{T'B + \frac{1}{2} T'\sigma_1^2(U'U)^{-1}T\}$$

$$m_{-\hat{C}}(T) = m_{\hat{C}}(-T) = \exp\{-T'C + \frac{1}{2} T'\sigma_2^2(V'V)^{-1}T\} .$$

Hence,

$$m_{\hat{B}-\hat{C}}(T) = m_{\hat{B}}(T) \cdot m_{-\hat{C}}(T)$$
$$= \exp\{T'B + \frac{1}{2}T'\sigma_1^2(U'U)^{-1}T\} \cdot \exp\{-T'C + \frac{1}{2}T'\sigma_2^2(V'V)^{-1}T\}$$
$$= \exp\{T'(B-C) + \frac{1}{2}T'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}]T\},$$

the moment generating function of a p-variate normally distributed random variable with mean $B-C$ and dispersion matrix $\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}$.

<u>Theorem 3.1:</u> $(\hat{B}-\hat{C})'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})$ is distributed as $\chi'^2_{(p,\lambda)}$. That is, the quadratic form has a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = \frac{1}{2}(B-C)'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}](B-C)$.

Proof: Let $D = \sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}$. Because D is symmetric positive definite, there exists (2, p. 3) a non-singular matrix P such that $PDP' = I$ or $D = P^{-1}P'^{-1}$ or $D^{-1} = P'P$. Define a random variable $R = P(\hat{B}-\hat{C})$. The moment generating function for R is:

$$m_R(T) = m_{P(\hat{B}-\hat{C})}(T) = m_{(\hat{B}-\hat{C})}(P'T)$$
$$= \exp\{T'P(B-C) + \frac{1}{2}T'PDP'T\}$$
$$= \exp\{T'P(B-C) + \frac{1}{2}T'T\},$$

hence R is $N(P(B-C), I)$. The quadratic form $R'R$ has a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = \frac{1}{2}(B-C)'P'P(B-C) = \frac{1}{2}(B-C)'D^{-1}(B-C)$ (2, p. 83). But, $R'R = (\hat{B}-\hat{C})'P'P(\hat{B}-\hat{C}) = (\hat{B}-\hat{C})'D^{-1}(\hat{B}-\hat{C})$ which proves the theorem.

If one is interested only in testing a hypothesis about the homogeneity of the two independent samples, then a test of the hypothesis that $B = C$, when σ_1^2 and σ_2^2 are known, would be such a test and the non-central chi-square statistic in Theorem 3.1 would be the appropriate test statistic. However, if it is a confidence region that is of interest, a central chi-square variable involving the parameters B and C must be used.

Theorem 3.2: $[(\hat{B}-\hat{C})-(B-C)]'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}]^{-1}[(\hat{B}-\hat{C})-(B-C)]$ has a chi-square distribution with p degrees of freedom.

Proof: The moment generating function of a constant vector K, say, is

$$m_K(T) = m_1(K'T) = \exp\{T'K\} .$$

Consequently, the moment generating function for $-(B-C)$ is $m_{-(B-C)}(T) = \exp\{-T'(B-C)\}$. Because a constant and a random variable are independent, it follows that

$$m_{(\hat{B}-\hat{C})-(B-C)}(T) = m_{(\hat{B}-\hat{C})}(T) \cdot m_{-(B-C)}(T)$$

$$= \exp\{\frac{1}{2}T'[\sigma_1^2(U'U)^{-1}+\sigma_2^2(V'V)^{-1}]T\} .$$

Hence, $(\hat{B}-\hat{C})-(B-C)$ is $N(0, \sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1})$. By the same reasoning employed in Theorem 3.1, the quadratic form

$$[(\hat{B}-\hat{C})-(B-C)]'[\sigma_1^2(U'U)^{-1} + \sigma_2^2(V'V)^{-1}]^{-1}[(\hat{B}-\hat{C})-(B-C)]$$

has a non-central chi-square distribution with p degrees of freedom and non- centrality parameter

$$\lambda = \frac{1}{2}[(B-C)-(B-C)]'[\sigma_1^2(U'U)^{-1}+\sigma_2^2(V'V)^{-1}]^{-1}[(B-C)-(B-C)] = 0 .$$

Therefore, the quadratic form has a central chi-square distribution.

If

$$S = [(\hat{B}-\hat{C})-(B-C)]'[\sigma_1^2(U'U)^{-1}+\sigma_2^2(V'V)^{-1}][(\hat{B}-\hat{C})-(B-C)]$$

and $\chi^2_{1-\alpha}(p)$ is the $1-\alpha$ point of the chi-square distribution, then $P[S < \chi^2_{1-\alpha}(p)] = 1-\alpha$ determines a confidence region in p-space. The boundary of the $100(1-\alpha)\%$ confidence region is given by the ellipsoid

$$[(\hat{B}-\hat{C})-(B-C)]'[\sigma_1^2(U'U)^{-1}+\sigma_2^2(V'V)^{-1}]^{-1}[(\hat{B}-\hat{C})-(B-C)] = \chi^2_{1-\alpha}(p)$$

in p-space centered at the point $(\hat{b}_1-\hat{c}_1, \hat{b}_2-\hat{c}_2, \cdots, \hat{b}_p-\hat{c}_p)$ (5, p. 265).

Let us now consider the case where the variances are unknown but equal. If the values of σ_1^2 and σ_2^2 are unknown, but the two parameters are known to be equal in value $(\sigma_1^2=\sigma_2^2=\sigma^2)$, then one may construct a non-central F statistic to test the homogeneity hypothesis, that is, test the hypothesis $B = C$. It is also possible to construct a confidence region for $(B-C)$ in p-space using a central F statistic. This construction is based on the above non-central χ^2 statistic and the estimate of σ^2 obtained by pooling the residual sums of squares from the two regressions. This development is summarized by

Theorem 3.3: The ratio

$$\{(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})(n-2p)\}/\{[(n_1-p)\hat{\sigma}_1^2+(n_2-p)\hat{\sigma}_2^2]p\}$$

has a non-central F distribution with p and n-2p degrees of freedom and non-centrality parameter

$$\lambda = (B-C)'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(B-C)/(2\sigma^2).$$

Proof: From Theorem 2.1 we know that $(W-U\hat{B})'(W-U\hat{B})/\sigma_1^2 = (n_1-p)\hat{\sigma}_1^2/\sigma_1^2$ has a chi-square distribution with n_1-p degrees of freedom, $(Z-V\hat{C})'(Z-V\hat{C})/\sigma_2^2=(n_2-p)\hat{\sigma}_2^2/\sigma_2^2$

has a chi-square distribution with n_2-p degrees of freedom, $\hat{B}$ and $\hat{\sigma}_1^2$ are independent, and that $\hat{C}$ and $\hat{\sigma}_2^2$ are independent. By virtue of the fact that W-sample and Z-sample are disjoint, we also know that these four terms $(\hat{B}, \hat{C}, \hat{\sigma}_1^2$, and $\hat{\sigma}_2^2)$ are mutually independent. If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $[(n_1-p)\hat{\sigma}_1^2 + (n_2-p)\hat{\sigma}_2^2/\sigma^2]$ has a chi-square distribution with n-2p degrees of freedom (3, p. 139-140). By Theorem 3.1 $(\hat{B}-\hat{C})'[(U'U)^{-1} + (V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/\sigma^2$ has a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = (B-C)'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(B-C)/2\sigma^2$. Therefore, the ratio

$$\{(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/(p\sigma^2)\}/\ \{[(n_1-p)\hat{\sigma}_1^2+(n_2-p)\hat{\sigma}_2^2]/(n-2p)\sigma^2\}$$

$$= \{(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})(n-2p)\}/\{[(n_1-p)\hat{\sigma}_1^2+(n_2-p)\hat{\sigma}_2^2]\,p\}$$

has a non-central F distribution with p and n-2p degrees of freedom and non-centrality parameter given above.

The independence of the numerator and denominator in the non-central F statistic is not affected by making the numerator a central chi-square statistic by subtracting the mean (B-C) from $(\hat{B}-\hat{C})$. Therefore,

$$F = \{[(\hat{B}-\hat{C})-(B-C)]'[(U'U)^{-1}+(V'V)^{-1}]^{-1}[(\hat{B}-\hat{C})-(B-C)](n-2p)\}/\{[(n_1-p)\hat{\sigma}_1^2+(n_2-p)\hat{\sigma}_2^2]\,p\}$$

has a central F distribution with p and n-2p degrees of freedom.

If $F_{1-\alpha}(p, n-2p)$ is the $1-\alpha$ point of the F distribution with the indicated degrees of freedom, then $P[F < F_{1-\alpha}(p,n-2p)] = 1-\alpha$ determines a $100(1-\alpha)\%$ confidence ellipsoid for $(B-C)$ in p-space, bounded by the surface $F = F_{1-\alpha}(p, n-2p)$, and centered at $(\hat{b}_1-\hat{c}_1, \hat{b}_2-\hat{c}_2, \cdots, \hat{b}_p-\hat{c}_p)$.

As is usually the case, the problem becomes much more difficult when the variances are both unknown and unequal, the Behrens-Fisher problem. It is now necessary to rely on tests and confidence regions based on approximate distributions arising from the large sample properties of maximum likelihood estimators. Let L be the likelihood function for a probability density function that has m parameters $Q = (q_1\ q_2 \cdots q_m)'$. Then (4, p. 51-55), the asymptotic distribution of the maximum likelihood estimate $\hat{Q}$ is multivariate normal with the parameters as means, and dispersion matrix D, where $D^{-1} = ((r_{ij}))$ with

$$r_{ij} = -E\left\{ \frac{\partial^2 \log L}{\partial q_i\, \partial q_j} \right\} .$$

Ultimately, the distribution of $[(\hat{B}-\hat{C})', (\hat{\sigma}_1^2-\hat{\sigma}_2^2)]'$ is needed. In order to derive it, the distribution of $[\hat{B}', \hat{\sigma}_1^2]'$ and $[\hat{C}', \hat{\sigma}_2^2]'$ must be determined. Let

$$L = f(W; B, \sigma_1^2) = (2\pi)^{-\frac{n_1}{2}} \cdot (\sigma_1^2)^{-\frac{n_1}{2}} \cdot \exp\{-\frac{1}{2\sigma_1^2}(W-UB)'(W-UB)\}$$

$$= (2\pi)^{-\frac{n_1}{2}} \cdot (\sigma_1^2)^{-\frac{n_1}{2}} \cdot \exp\{-\frac{1}{2\sigma_1^2} \cdot \sum_{m=1}^{n_1} (w_m - \sum_{k=1}^{p} u_{mk} b_k)^2\},$$

then

$$\log L = -\frac{n_1}{2} \cdot \log(2\pi) - \frac{n_1}{2} \cdot \log(\sigma_1^2) - \frac{1}{2\sigma_1^2} \sum_{m=1}^{n_1} (w_m - \sum_{k=1}^{p} u_{mk} b_k)^2,$$

$$\frac{\partial \log L}{\partial b_i} = \frac{1}{2\sigma_1^2} \sum_{m=1}^{n_1} 2(w_m - \sum_{k=1}^{p} u_{mk} b_k)(-u_{mi})$$

$$= \frac{1}{\sigma_1^2} \sum_{m=1}^{n_1} u_{mi}(w_m - \sum_{k=1}^{p} u_{mk} b_k),$$

and

$$\frac{\partial^2 \log L}{\partial b_i \partial b_j} = \frac{1}{\sigma_1^2} \sum_{m=1}^{n_1} u_{mi}(-u_{mj}) = -\frac{1}{\sigma_1^2} \sum_{m=1}^{n_1} u'_{im} u_{mj}$$

$$= -\frac{1}{\sigma_1^2} \cdot (i, j \text{ th element of } U'U).$$

Also

$$\frac{\partial \log L}{\partial(\sigma_1^2)} = -\frac{n_1}{2\sigma_1^2} + \frac{1}{2\sigma_1^4}\sum_{m=1}^{n_1}(w_m - \sum_{k=1}^{p} u_{mk}b_k)^2 ,$$

$$\frac{\partial^2 \log L}{\partial^2(\sigma_1^2)} = \frac{n_1}{2\sigma_1^4} - \frac{1}{\sigma_1^6}\sum_{m=1}^{n_1}(w_m - \sum_{k=1}^{p} u_{mk}b_k)^2 ,$$

and

$$\frac{\partial^2 \log L}{\partial b_i \partial\sigma_1^2} = -\frac{1}{\sigma_1^4}\sum_{m=1}^{n_1} u_{mi}(w_m - \sum_{k=1}^{p} u_{mk}b_k) .$$

Therefore the elements r_{ij} of the inverse of the dispersion matrix for $[\hat{B}', \hat{\sigma}_1^{2\,\prime}]$ are

$$r_{ij} = -E\left\{\frac{\partial^2 \log L}{\partial b_i \partial b_j}\right\} = -E\left\{-\frac{1}{\sigma_1^2}(\text{i, jth element of } U'U)\right\}$$

$$= \frac{1}{\sigma_1^2}(\text{i, jth element of } U'U)$$

for $i, j = 1(1)p$,

$$r_{i(p+1)} = r_{(p+1)i} = -E\left\{-\frac{1}{\sigma_1^4}\sum_{m=1}^{n_1} u_{mi}(w_m - \sum_{k=1}^{p} u_{mk}b_k)\right\}$$

$$= \frac{1}{\sigma_1^4}\sum_{m=1}^{n_1} u_{mi}E(w_m - \sum_{k=1}^{p} u_{mk}b_k) = 0$$

for $i = 1(1)p$, and

$$r_{(p+1)(p+1)} = - E \left\{ \frac{n_1}{2\sigma_1^4} - \frac{1}{\sigma_1^6} \sum_{m=1}^{n_1} (w_m - \sum_{k=1}^{p} u_{mk} b_k)^2 \right\}$$

$$= - \frac{n_1}{2\sigma_1^4} + \frac{1}{\sigma_1^6} \sum_{m=1}^{n_1} E(w_m - \sum_{k=1}^{p} u_{mk} b_k)^2$$

$$= - \frac{n_1}{2\sigma_1^4} + \frac{1}{\sigma_1^6} \sum_{m=1}^{n_1} \sigma_1^2 = \frac{n_1}{2\sigma_1^4} .$$

If D is the dispersion matrix for $[\hat{B}', \hat{\sigma}_1^2 ']$, then combining the above results gives

$$D^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} (U'U) & 0 \\ 0 & \frac{n_1}{2\sigma_1^4} \end{bmatrix}$$

and consequently (6, p. 48-51),

$$D = \begin{bmatrix} \sigma_1^2 (U'U)^{-1} & 0 \\ 0 & \frac{2\sigma_1^4}{n_1} \end{bmatrix} .$$

According to large sample theory, if the maximum-likelihood estimators for the parameters in the dispersion matrix are substituted for the parameters, a good approximation of the density function for $[\hat{B}', \hat{\sigma}_1^2]'$ is obtained (5, p. 264). Therefore, $\begin{bmatrix} \hat{B} \\ \hat{\sigma}_1^2 \end{bmatrix}$ is approximately multivariate normal with mean $\begin{bmatrix} B \\ \sigma_1^2 \end{bmatrix}$ and dispersion matrix/1

$$\hat{D} = \begin{bmatrix} \frac{n_1-p}{n_1} \hat{\sigma}_1^2 (U'U)^{-1} & 0 \\ 0 & \frac{2(n_1-p)\hat{\sigma}_1^2}{n_1^2} \end{bmatrix} .$$

Similarly, for large samples $\begin{bmatrix} \hat{C} \\ \hat{\sigma}_2^2 \end{bmatrix}$ is approximately multivariate normal with mean $\begin{bmatrix} C \\ \sigma_2^2 \end{bmatrix}$ and dispersion matrix

$$\hat{D} = \begin{bmatrix} \frac{n_2-p}{n_2} \hat{\sigma}_2^2 (V'V)^{-1} & 0 \\ 0 & \frac{2(n_2-p)\hat{\sigma}_2^2}{n_2^2} \end{bmatrix} .$$

Using the same argument as in Theorem 3.1, it clearly follows that for large samples the quadratic form

---

/1 $\hat{\sigma}_1^2$ is the unbiased estimator of σ_1^2. The maximum likelihood estimator of σ_1^2 is $(n_1-p)\hat{\sigma}_1^2/n_1$.

$$\begin{bmatrix} \hat{B}-\hat{C} \\ \hat{\sigma}_1^2-\hat{\sigma}_2^2 \end{bmatrix}' \begin{bmatrix} \frac{n_1-p}{n_1}\hat{\sigma}_1^2(U'U)^{-1}+\frac{n_2-p}{n_2}\hat{\sigma}_2^2(V'V)^{-1} & 0 \\ 0 & \frac{2(n_1-p)\hat{\sigma}_1^2}{n_1^2}+\frac{2(n_2-p)\hat{\sigma}_2^2}{n_2^2} \end{bmatrix} \begin{bmatrix} \hat{B}-\hat{C} \\ \hat{\sigma}_1^2-\hat{\sigma}_2^2 \end{bmatrix}$$

has approximately a non-central chi-square distribution with $p+1$ degrees of freedom and non-centrality parameter

$$\lambda=\tfrac{1}{2}\begin{bmatrix} B-C \\ \sigma_1^2-\sigma_2^2 \end{bmatrix}' \begin{bmatrix} \frac{n_1-p}{n_1}\hat{\sigma}_1^2(U'U)^{-1}+\frac{n_2-p}{n_2}\hat{\sigma}_2^2(V'V)^{-1} & 0 \\ 0 & \frac{2(n_1-p)\hat{\sigma}_1^2}{n_1^2}+\frac{2(n_2-p)\hat{\sigma}_2^2}{n_2^2} \end{bmatrix} \begin{bmatrix} B-C \\ \sigma_1^2-\sigma_2^2 \end{bmatrix}.$$

This non-central chi-square statistic can be used in an approximate test of the hypothesis that the two subsamples (W-sample and Z-sample) have the same parameters.

Letting D be the matrix in the above quadratic form and using the same argument as in Theorem 3. 2 we see that

$$\begin{bmatrix} (\hat{B}-\hat{C})-(B-C) \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2)-(\sigma_1^2-\sigma_2^2) \end{bmatrix}' D \begin{bmatrix} (\hat{B}-\hat{C})-(B-C) \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2)-(\sigma_1^2-\sigma_2^2) \end{bmatrix}$$

has approximately, for large samples, a central chi-square distribution with $p+1$ degrees of freedom. Therefore, an ellipsoidal $100\,(1-\alpha)\,\%$ confidence region with center at

$(b_1-c_1,\ b_2-c_2\ ,\ \cdots\ ,\ b_p-c_p,\ \sigma_1^2-\sigma_2^2)$ can be constructed with the boundary given by

$$\chi^2_{1-\alpha}(p+1) = \begin{bmatrix} (\hat{B}-\hat{C}) - (B-C) \\ \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2)-(\sigma_1^2-\sigma_2^2) \end{bmatrix}' \quad \hat{D} \begin{bmatrix} (\hat{B}-\hat{C}) - (B-C) \\ \\ (\hat{\sigma}_1^2-\hat{\sigma}_2^2)-(\sigma_1^2-\sigma_2^2) \end{bmatrix} \quad .$$

IV. PROCEDURES WHEN TOTAL SAMPLE AND ONE SUBSAMPLE ARE FITTED

If it happens that the regression equation is fitted to the total sample and one of the subsamples rather than to the two disjoint subsamples, then the derived coefficient vectors will not be independent nor will the residual sums of squares. For the sake of discussion and without loss of generality, assume that the W-sample is the fitted subsample. The question arises, can the information obtained from these two fits be used to detect any inhomogeneity among the two disjoint subsamples ?

In the case where the normal equations and total sums of squares are known, this problem has a direct solution. If the factors in the normal equations $(X'X,\ X'Y,\ U'U, \text{ and } U'W)$ and the total sums of squares $(Y'Y \text{ and } W'W)$ are known, $\hat{C}$ and $(Z-V\hat{C})'(Z-V\hat{C})$ may be determined directly from these terms without having to fit the Z-sample.

<u>Theorem 4.1:</u> $\hat{C} = (X'X-U'U)^{-1}(X'Y-U'W)$ and

$$(Z-V\hat{C})'(Z-V\hat{C}) = (Y'Y-W'W)-(X'Y-U'W)'(X'X-U'U)^{-1}(X'Y-U'W).$$

Proof: $$Y'Y = [W', Z'] \begin{bmatrix} W \\ Z \end{bmatrix} = W'W + Z'Z,$$

$$X'X = [U', V'] \begin{bmatrix} U \\ V \end{bmatrix} = U'U + V'V,$$

and

$$X'Y = [U', V'] \begin{bmatrix} W \\ Z \end{bmatrix} = U'W + V'Z .$$

Therefore

$$Z'Z = Y'Y - W'W, \quad V'V = X'X - U'U,$$

and

$$V'Z = X'Y - U'W.$$

Now,

$$\hat{C} = (V'V)^{-1}V'Z,$$

consequently

$$\hat{C} = (X'X - U'U)^{-1}(X'Y - U'W);$$

and from Theorem 2.1

$$\begin{aligned}(Z-V\hat{C})'(Z-V\hat{C}) &= Z'(I-V(V'V)^{-1}V')Z \\ &= Z'Z - Z'V(V'V)^{-1}V'Z \\ &= (Y'Y-W'W)-(X'Y-U'W)'(X'X-U'U)^{-1}(X'Y-U'W).\end{aligned}$$

We therefore see that with this amount of information available the tests and procedures in Chapter 3 may be used to test whether the W-sample and the Z-sample come from the same population.

In the case where the factors in the normal equations are unknown, the problem is somewhat more involved. Assume, again, that the total sample and one subsample were fitted with the same regression equation, only this time the factors of the normal equations $(X'X,\ X'Y,\ U'U,$ and $U'W)$ are not available. The only information available being the residual sum of squares, $(W-U\hat{B})'(W-U\hat{B})$, from the fitted subsample and a way to compute $(W-U\hat{A})'(W-U\hat{A})$. If the residuals $(Y-X\hat{A})$ are available, as they were in the oceanography example cited in the introduction, then $(W-U\hat{A})'(W-U\hat{A})$ may be computed by summing the squares of the components of $(Y-X\hat{A})$ over the W-sample space only. To be more precise, if

$$K = \begin{bmatrix} I(n_1 \times n_1) & 0\,(n_1 \times n_2) \\ 0(n_2 \times n_1) & 0\,(n_2 \times n_2) \end{bmatrix} ,$$

then

$$\begin{aligned} K(Y-X\hat{A}) &= KY-KX\hat{A} \\ &= K\begin{bmatrix} W \\ Z \end{bmatrix} - K\begin{bmatrix} U \\ V \end{bmatrix}\hat{A} \\ &= (W-U\hat{A}). \end{aligned}$$

Therefore

$$(W-U\hat{A})'(W-U\hat{A}) = (Y-X\hat{A})'K'K(Y-X\hat{A}) = (Y-X\hat{A})'K(Y-X\hat{A})$$

which is the sum of squares of the components of $(Y-X\hat{A})$ over the W-sample. With this information it is possible to obtain the quadratic form $(\hat{B}-\hat{A})' U' U(\hat{B}-\hat{A})$ which, if its distribution were known and tabulated, would be useful in making inferences about the homogeneity of the total population.

In order to study the quadratic form $(\hat{B}-\hat{A})' U' U(\hat{B}-\hat{A})$ some properties that are consequences of the normal equations will be needed. These properties are stated in the following theorems.

Theorem 4.2: Let $Y = XA+\varepsilon$ be any general linear hypothesis model where $\hat{A}$ is the least-square estimator of A, then $X'(Y-X\hat{A}) = 0 \ (p\times 1)$.

Proof: From the normal equations $X'X\hat{A} = X'Y$ we get $X'Y-X'X\hat{A} = 0 \ (p\times 1)$ and finally, $X'(Y-XA) = 0 \ (p\times 1)$.

Theorem 4.3: Let $Y = XA+\varepsilon$, $W = UB+\eta$, and $Z=VC+\delta$ be the regression equation for the total sample and the two disjoint subsamples, respectively, as defined in Chapter 2. Then $U'U(\hat{B}-\hat{A}) = V'V(\hat{A}-\hat{C})$.

Proof: From Theorem 4.1 we know that $X'X = U'U + V'V$ and that $X'Y = U'W + V'Z$. From the normal equations $X'XA = X'Y$ we get $(U'U+V'V)\hat{A} = U'W + V'Z$. But $U'U\hat{B} = U'W$ and $V'V\hat{C} = V'Z$; therefore, $U'U\hat{A} + V'V\hat{A} = U'U\hat{B} + V'V\hat{C}$, and then

$U'U\hat{B} - U'U\hat{A} = V'V\hat{A} - V'V\hat{C}$, and finally $U'U(\hat{B}-\hat{A}) = V'V(\hat{A}-\hat{C})$.

Corollary 4.3.1: Since $X'X = U'U + V'V$, we have $V'V = X'X - U'U$. Therefore,

$$U'U(\hat{B}-\hat{A}) = (X'X - U'U)(\hat{A}-\hat{C}),$$

$$U'U(\hat{B}-\hat{A}) = X'X(\hat{A}-\hat{C}) - U'U(\hat{A}-\hat{C}),$$

then

$$U'U(\hat{B}-\hat{A}) + U'U(\hat{A}-\hat{C}) = X'X(\hat{A}-\hat{C}),$$

from which it follows that

$$U'U(\hat{B}-\hat{C}) = X'X(\hat{A}-\hat{C}).$$

$(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})$ is obtained by finding the difference of the known sums of squares $(W-U\hat{A})'(W-U\hat{A})$ and $(W-U\hat{B})'(W-U\hat{B})$. This may be seen by partitioning the sum of squares $(W-U\hat{A})'(W-U\hat{A})$.

Theorem 4.4: $(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A}) = (W-U\hat{A})'(W-U\hat{A}) - (W-U\hat{B})'(W-U\hat{B})$.

Proof: $(W-U\hat{A})'(W-U\hat{A}) = (W-U\hat{B}+U\hat{B}-U\hat{A})'(W-U\hat{B}+U\hat{B}-U\hat{A}),$

$$(W-U\hat{A})'(W-U\hat{A}) = [(W-U\hat{B})+U(\hat{B}-\hat{A})]'[(W-U\hat{B})+U(\hat{B}-\hat{A})],$$

$$(W-U\hat{A})'(W-U\hat{A}) = (W-U\hat{B})'(W-U\hat{B}) + (W-U\hat{B})'U(\hat{B}-\hat{A})$$
$$+ (\hat{B}-\hat{A})'U'(W-U\hat{B}) + (\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A}).$$

By Theorem 4.2,

$$(W-U\hat{B})'U = U'(W-U\hat{B}) = 0 .$$

Therefore,

$$(W-U\hat{A})'(W-U\hat{A}) = (W-U\hat{B})'(W-U\hat{B}) + (\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})$$

and hence,

$$(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A}) = (W-U\hat{A})'(W-U\hat{A})-(W-U\hat{B})'(W-U\hat{B}).$$

Thus the quadratic form $(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})$ may be computed without explicitly knowing $U'U$.

To study the distribution of $(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})$, let us first look at the distribution of $(\hat{B}-\hat{A})$.

Theorem 4.5: $(\hat{B}-\hat{A})$ has a multivariate normal distribution with mean $[I-(X'X)^{-1}U'U]B-(X'X)^{-1}V'VC$ and dispersion matrix $\sigma_1^2[(U'U)^{-1}-2(X'X)^{-1}+(X'X)^{-1}U'U(X'X)^{-1}] + \sigma_2^2(X'X)^{-1}V'V(X'X)^{-1}$.

Proof: From Theorem 2.1 $\hat{B} = (U'U)^{-1}U'W$

and

$$\hat{A} = (X'X)^{-1}X'Y ;$$

therefore

$$(\hat{B}-\hat{A}) = (U'U)^{-1}U'W-(X'X)^{-1}X'Y,$$

$$(\hat{B}-\hat{A}) = (U'U)^{-1}U'W-(X'X)^{-1}(U'W+V'Z) ;$$

therefore,

$$(\hat{B}-\hat{A}) = [(U'U)^{-1}-(X'X)^{-1}]U'W-(X'X)^{-1}V'Z.$$

The moment generating function for W is

$m_W(T) = \exp\{T'UB + \frac{1}{2}T'\sigma_1^2 T\}$. Therefore the moment generating function for $[(U'U)^{-1}-(X'X)^{-1}]U'W$ is

$$\begin{aligned} m_{[(U'U)^{-1}-(X'X)^{-1}]U'W}(T) &= m_W(U[(U'U)^{-1}-(X'X)^{-1}]T) \\ &= \exp\{T'[(U'U)^{-1}-(X'X)^{-1}]U'UB \\ &\quad + \frac{1}{2}T'[(U'U)^{-1}-(X'X)^{-1}]U'\sigma_1^2 U[(U'U)^{-1}-(X'X)^{-1}]T\} \\ &= \exp\{T'[I-(X'X)^{-1}U'U]B \\ &\quad + \frac{1}{2}T'\sigma_1^2[(U'U)^{-1}-2(X'X)^{-1}+(X'X)^{-1}U'U(X'X)^{-1}]T\}. \end{aligned}$$

The moment generating function for Z is

$m_Z(T) = \exp\{T'VC + \frac{1}{2}T'\sigma_2^2 T\}$. Therefore, the moment generating function for $-(X'X)^{-1}V'Z$ is

$$\begin{aligned} m_{-(X'X)^{-1}V'Z}(T) &= m_Z(-V(X'X)^{-1}T) \\ &= \exp\{-T'(X'X)^{-1}V'VC + \frac{1}{2}T'(X'X)^{-1}V'\sigma_2^2 V(X'X)^{-1}T\} \\ &= \exp\{-T(X'X)^{-1}V'VC + \frac{1}{2}T'[\sigma_2^2(X'X)^{-1}V'V(X'X)^{-1}]T\}. \end{aligned}$$

Using the fact that W and Z are independent we have

$$m_{\hat{B}-\hat{A}}(T) = m_{[(U'U)^{-1}-(X'X)^{-1}]U'W-(X'X)^{-1}V'Z}(T)$$

$$= m_{[(U'U)^{-1}-(X'X)^{-1}]U'W}(T) \cdot m_{-(X'X)^{-1}V'Z}(T)$$

$$= \exp\{T'[B-(X'X)^{-1}U'UB-(X'X)^{-1}V'VC]$$

$$+\frac{1}{2}T'[\sigma_1^2((U'U)^{-1}-2(X'X)^{-1}+(X'X)^{-1}U'U(X'X)^{-1})$$

$$+\sigma_2^2(X'X)^{-1}V'V(X'X)^{-1}]T\}.$$

This last expression is the moment generating function for a multivariate normal random variable with mean $[I-(X'X)^{-1}U'U]B-(X'X)V'VC$ and dispersion matrix

$\sigma_1^2[(U'U)^{-1}-2(X'X)^{-1}+(X'X)^{-1}U'U(X'X)^{-1}]+\sigma_2^2(X'X)^{-1}V'V(X'X)^{-1}$.

When it can be assumed that $\sigma_1^2 = \sigma_2^2$, the dispersion matrix of $(\hat{B}-\hat{A})$ will be $\sigma_1^2[(U'U)^{-1}-(X'X)^{-1}]$.

Consider the quadratic form $(\hat{B}-\hat{A})'(\frac{U'U}{\sigma_1^2})(\hat{B}-\hat{A})$ and make the assumption $\sigma_1^2 = \sigma_2^2$, then $(\hat{B}-\hat{A})'(\frac{U'U}{\sigma_1^2})(\hat{B}-\hat{A})$ will have a non-central chi-square distribution if and only if the product of the matrix of the quadratic form with the dispersion matrix for $(\hat{B}-\hat{A})$ is idempotent (2, p. 84). However,

$$(U'U/\sigma_1^2)\cdot\{\sigma_1^2[(U'U)^{-1}-(X'X)^{-1}]\} = [I-U'U(X'X)^{-1}]$$

and this matrix is not idempotent. Therefore, $(\hat{B}-\hat{A})'(U'U/\sigma_1^2)(\hat{B}-\hat{A})$ does not have a non-central chi-square distribution, even under the restriction that $\sigma_1^2 = \sigma_2^2$.

Rather than pursue the distribution of $(\hat{B}-\hat{A})'(U'U/\sigma_1^2)(\hat{B}-\hat{A})$, consider the ratio $(\hat{B}-\hat{A})' U'U(\hat{B}-\hat{A})/(p\hat{\sigma}_1^2)$ as a conservative approximation to the F statistic

$$F_1 = (\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/(p\hat{\sigma}_1^2).$$

When $\sigma_1^2 = \sigma_2^2$, $(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/\sigma_1^2$ has a non-central chi-square distribution with p degrees of freedom by Theorem 3.1. From Theorem 3.3, recall that $(W-U\hat{B})'(W-U\hat{B})/\sigma_1^2 = (n_1-p)\hat{\sigma}_1^2/\sigma_1^2$ has a central chi-square distribution with n_1-p degrees of freedom and that $(n_1-p)\hat{\sigma}_1^2/\sigma_1^2$ is independent of $(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/\sigma_1^2$. Consequently, the ratio

$$F_1 = \{(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/(\sigma_1^2 p)\}/\{(n_1-p)\hat{\sigma}_1^2/[\sigma_1^2(n_1-p)]\}$$
$$= (\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/(p\hat{\sigma}_1^2)$$

has a non-central F distribution with p and n_1-p degrees of freedom.

If it can be shown that

$$(\hat{B}-\hat{A})' U' U(\hat{B}-\hat{A}) \leq (\hat{B}-\hat{C})' [(U'U)^{-1}+(V'V)^{-1}]^{-1} (\hat{B}-\hat{C})$$

then

$$F_2 = \{(\hat{B}-\hat{A})' U' U(\hat{B}-\hat{A})/(\sigma_1^2 p)\} / \{(n_1-p)\hat{\sigma}_1^2/[(n_1-p)\sigma_1^2]\}$$

$$= (\hat{B}-\hat{A})' U' U(\hat{B}-\hat{A})/(p\hat{\sigma}_1^2) \leq F_1 .$$

In order to show that $(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A}) \leq (\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})$ it is necessary to prove an additional theorem.

<u>Theorem 4.6:</u> $[(U'U)^{-1}+(V'V)^{-1}]^{-1} = U'U-U'U(X'X)^{-1}U'U.$

Proof: Consider

$$(U'U)^{-1}+(V'V)^{-1} = (V'V)^{-1}(V'V+U'U)(U'U)^{-1}$$

$$= (V'V)^{-1}(X'X)(U'U)^{-1} .$$

Then

$$[(U'U)^{-1}+(V'V)^{-1}]^{-1} = U'U(X'X)^{-1}V'V$$

$$= U'U(X'X)^{-1}(X'X-U'U)$$

$$= U'U-U'U(X'X)^{-1}U'U.$$

<u>Theorem 4.7:</u> $(\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A}) \leq (\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C}).$

Proof: By Theorem 4.6

$$(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C}) = (\hat{B}-\hat{C})'[U'U-U'U(X'X)^{-1}U'U](\hat{B}-\hat{C})$$

$$= (\hat{B}-\hat{C})'U'U(\hat{B}-\hat{C})-(\hat{B}-\hat{C})'U'U(X'X)^{-1}U'U(\hat{B}-\hat{C}).$$

From Corollary 4.3.1,

$$(\hat{A}-\hat{C}) = (X'X)^{-1}U'U(\hat{B}-\hat{C}) ;$$

therefore,

$$(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C}) = (\hat{B}-\hat{C})'U'U(\hat{B}-\hat{C})-(\hat{B}-\hat{C})'U'U(\hat{A}-\hat{C})$$

$$= (\hat{B}-\hat{C})U'U(\hat{B}-\hat{C}-\hat{A}+\hat{C})$$

$$= (\hat{B}-\hat{C})U'U(\hat{B}-\hat{A})$$

$$= [(\hat{B}-\hat{A})+(\hat{A}-\hat{C})]'U'U(\hat{B}-\hat{A})$$

$$= (\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})+(\hat{A}-\hat{C})'U'U(\hat{B}-\hat{A}).$$

From Theorem 4.3, $U'U(\hat{B}-\hat{A}) = V'V(\hat{A}-\hat{C})$; therefore,

$$(\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C}) = (\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})+(\hat{A}-\hat{C})'V'V(\hat{A}-\hat{C}).$$

Since $V'V$ is positive definite, the theorem follows, with equality holding only when $\hat{A} = \hat{C}$, in which case $\hat{B} = \hat{A}$ and $\hat{B} = \hat{C}$.

As a direct consequence of Theorem 4.7

$$F_2 = (\hat{B}-\hat{A})'U'U(\hat{B}-\hat{A})/(p\hat{\sigma}_1^2) = \{(W-U\hat{A})'(W-U\hat{A})-(W-U\hat{B})'(W-U\hat{B})\}/(p\hat{\sigma}_1^2)$$

$$\leq F_1 = (\hat{B}-\hat{C})'[(U'U)^{-1}+(V'V)^{-1}]^{-1}(\hat{B}-\hat{C})/(p\hat{\sigma}_1^2).$$

Consequently, if $F_2 > F_{1-\alpha}(p, n_1-p)$, the $(1-\alpha)$ point of the central F distribution with p and n_1-p degrees of freedom, a hypothesis that $\hat{B} = \hat{C}$ (under the assumption that $\sigma_1^2 = \sigma_2^2$) would be rejected. It is to be emphasized, however, that obtaining a non-critical F_2 does not necessarily mean that F_1 would also be non-critical. F_2, then, is the conservative approximation to the F-statistic that was being sought.

V. SUMMARY AND CONCLUSION

Available procedures for testing the homogeneity of two samples fitted by the same regression equation have been discussed. The tests vary according to how much is known, or can be assumed, about the two samples. Essentially these procedures involve testing a hypothesis that the coefficient vectors from the two fits are equal when their residual variances are known, unknown but equal, and unknown. Ellipsoidal confidence regions for the unknown parameters in the regression equations are also obtainable under these conditions.

When the two samples together and one sample alone are fitted separately by the regression equation, the parameters associated with the unfitted sample are determinable if the sums of squares and the factors in the normal equations for both fits are known. Hence, the statistical techniques for constructing confidence regions, and testing hypotheses, that are applicable where the two disjoint samples are individually fitted, are available. If total sums of squares and the factors in the normal equations are not known, or are not available, then with little more than the residual sums of squares for each fit a conservative approximation to an exact test is available to test the homogeneity hypothesis, if it can be assumed that the residual variances for the two disjoint samples are equal.

When there is a choice, the approach wherein the two disjoint samples are individually fitted by the regression equation is preferable. In the event that it is desirable to fit one sample and the union of the two disjoint samples, retain all of the sums of squares and cross-product matrices that enter into the normal equations, as well as the total sums of squares, so that the parameters of the unfitted sample may be determined. Then the same statistical techniques are available as when the two disjoint samples are fitted individually. The approximation to the F statistic that utilizes only two sums of squares should be used only in situations where all of the above information is not available.

BIBLIOGRAPHY

1. Anderson, E. R. Sea-surface temperature estimation ··· use of regression models for time-space interpolation of sea-surface temperature observations. San Diego, 1967. 34 p. (U.S. Navy Electronics Laboratory. Technical report no. 1429)

2. Graybill, Franklin A. An introduction to linear statistical models. Vol. 1. New York, McGraw-Hill, 1961. 463 p.

3. Hogg, Robert V. and Allen T. Craig. Introduction to mathematical statistics. 2d ed. New York, Macmillan, 1965. 383 p.

4. Kendall, Maurice G. and Alan Stuart. The advanced theory of statistics. Vol. 2. New York, Hafner, 1961. 676 p.

5. Mood, Alexander M. and Franklin A. Graybill. Introduction to the theory of statistics. 2d ed. New York, McGraw-Hill, 1963. 443 p.

6. Stephenson, G. Matrices, sets and groups; an introduction for students of science and engineering. New York, American Elsevier, 1965. 164 p.

NUWC TP 36

nuwc

STATISTICAL TESTS OF THE EQUALITY OF TWO REGRESSIONS

by J. D. Pugh

Ocean Sciences Department • San Diego, California

May 1968

NAVAL UNDERSEA WARFARE CENTER
An activity of the Naval Material Command

G. H. Lowe, Capt., USN
Commander

Wm. B. McLean, Ph.D.
Technical Director

Work was performed under SR 104 03 01, Task 0586 (NUWC L405) by the Oceanometrics Division. This report covers work from July 1966 to May 1967 and was approved for publication 13 May 1968.

This report is a revision of the thesis the author used to fulfill one of the requirements for the degree of master of science at Oregon State University.

Released by
E. R. Anderson, Head
Oceanometrics Division

Under authority of
G. B. Anderson, Acting Head
Ocean Sciences Department

The distribution control point for this report is NUWC, San Diego, California 92152.