The Annals of Mathematical Statistics
1972, Vol. 43, No. 4, 1076-1088

MULTIPLE COMPARISON OF REGRESSION FUNCTIONS[1]

BY EMIL SPJØTVOLL

University of Oslo

1. Introduction. The situation discussed under the heading "Regression Analysis", treated in many textbooks, is that we have an $n \times 1$ random variable y which is $N(X'\beta, \sigma^2 I)$, where X' is a known $n \times p$ matrix, β is an unknown $p \times 1$ vector parameter and σ^2 is an unknown variance. The statistical problems are to estimate parameters and to test hypotheses concerning the parameters.

In practice, however, the situation is not so simple in most cases. Often the true form of the expectation of y is not known, but one has some variables which one suspects contribute to Ey in some way. Let these variables be the columns of the matrix X'. One then tries to describe Ey by $X'\beta$ for some β. But one cannot be sure, even with a large number p of variables that the statement "Ey is equal to $X'\beta$ for some β" is true. If y and the p variables of X' have a joint multinormal distribution, the above statement will be true conditionally given the p variables of X'. This argument cannot be used in many situations. In some, it can easily be seen that some of the variables in X' do not have a normal distribution, or maybe both a transform of a variable and the variable itself occur in X'. It is, however, possible (as remarked by a referee) that conditionally $Ey = X'\beta$ even if y and the variables in X' do not have a joint multinormal distribution.

Since usually too many variables are included in X', procedures have been developed for excluding variables which do not contribute to Ey; see, for example, Beale, Kendall and Mann (1967), and Draper and Smith (1966). An important class of such procedures are so-called stepwise regression methods. A description of some of these can be found in Draper and Smith (1966). It seems to be commonly accepted that stepwise regression methods lack justification by statistical theory. In particular the fact that different stepwise regression methods often give different results is confusing, see Hamaker (1962), and Draper and Smith (1966).

What one often ends up with is several regression functions which seem to be good candidates for the regression function to be used. Some of these regression functions might have been obtained by use of a stepwise procedure, some might have been obtained because the statistician has looked at some particular combination of variables. Lately, efficient computer procedures have been developed for computing all possible regression functions or certain subsets containing the "best" regression function; see [8] and [10]. The statistician, therefore, is faced with the problem of choosing among a (usually) large number

Received May 1, 1969; revised October 18, 1971.

[1] This paper was written while the author was visiting at the University of California, Berkeley, and revised at the University of Wisconsin and the University of Oslo.

of regression functions.

The aim of the present paper is to develop a technique for choosing among all possible regression functions in a given set. The technique is one of multiple comparison, and is analogous to, and can be applied in a way similar to, the S-method and T-method (see Scheffé (1959)) in the analysis of variance. Similar problems have been treated in [2], [7], [9], [12].

2. Assumptions and problems. We will assume that the $n \times 1$ random vector y is $N(\eta, \sigma^2 I)$, where η is unknown. Suppose we try to describe η by $X'\beta$ where X' is a known $n \times p$ matrix of rank p, and β is an unknown $p \times 1$ vector. Estimation of β by least squares gives the estimate $\hat{\beta} = (XX')^{-1}Xy$, with $E\hat{\beta} = (XX')^{-1}X\eta$. Using this estimate of β we estimate η by

$$\hat{\eta} = X'\hat{\beta} = X'(XX')^{-1}Xy \,. \tag{2.1}$$

We have $E\hat{\eta} = X'(XX')^{-1}X\eta$. The estimate $\hat{\eta}$ is the projection of y on $C(X')$ (we use the notation $C(A)$ to denote the space spanned by the columns of a matrix A), and $E\hat{\eta}$ is the projection of η on the same space. The latter is equal to η if and only if $\eta \in C(X')$. From the above it follows that if we try to write the expectation of y in the form $X'\beta$, and estimate β by least squares, then we are estimating the projection of Ey on $C(X')$.

The usual estimate of σ^2 is

$$s^2 = (y - \hat{\eta})'(y - \hat{\eta})/(n - p) \,. \tag{2.2}$$

Here $(n - p)s^2/\sigma^2$ has a chi-square distribution with $n - p$ degrees of freedom, and is statistically independent of $\hat{\eta}$. It is a central chi-square distribution if $\eta \in C(X')$, if not, it is noncentral with noncentrality parameter $\sigma^{-2}\eta'(I - X'(XX')^{-1}X)\eta$. It follows that

$$Es^2 = \sigma^2 + \eta'(I - X'(XX')^{-1}X)\eta/(n - p) \,. \tag{2.3}$$

Suppose we have to choose between two regression functions $X_1'\beta_1$ and $X_2'\beta_2$ where X_i' is a known $n \times p_i$ matrix of rank p_i, and $C(X_i') \subset C(X')$, $i = 1, 2$. The estimates of η are $\hat{\eta}_i = X_i'(X_iX_i')^{-1}X_iy$, $i = 1, 2$, with expected values $\eta_i = X_i'(X_iX_i')^{-1}X_i\eta$, $i = 1, 2$. As a measure of the goodness of fit of the regression function $X_i'\beta$ we can use the squared length of the difference between η_i and η, smaller values indicating better fit. This is found to be

$$(\eta - \eta_i)'(\eta - \eta_i) = \eta'\eta - \eta'X_i'(X_iX_i')^{-1}X_i\eta \,.$$

Alternatively we could use the cosine of the angle between η and η_i, which is

$$\cos(\eta, \eta_i) = (\eta'\eta)^{-\frac{1}{2}}(\eta'X_i'(X_iX_i')^{-1}X_i\eta)^{\frac{1}{2}} \,,$$

larger values indicating better fit. It is seen that both these measures are equivalent to

$$\eta_i'\eta_i = \eta'X_i'(X_iX_i')^{-1}X_i\eta \,, \tag{2.4}$$

larger values indicating better fit. (2.4) is the squared length of the projection

of η on $C(X_i')$. Our definition of goodness will now be based upon (2.4). If we have two (or more) regression functions, the function with the largest $\eta_i'\eta_i$ will be said to be best.

In Section 5 simultaneous comparison of functions of the form (2.4) will be considered. Of course the maximum value of (2.4) is obtained for $X_i' = X'$, but some matrix X_i' with fewer than p columns may be equally good in the sense that $\eta_i'\eta_i$ also is equal to the maximum value of (2.4), and in that case we prefer to use X_i'. In some situations one is not interested in all possible matrices X_i', but for example only matrices with a given number $k < p$ of columns, e.g., see Hamaker (1962). Then (2.4) can be used to compare these. The above criterion of goodness of a regression function is, of course, one of many possible criteria.

3. Multiple comparisons of quadratic functions. Let $\hat{\psi} = (\hat{\psi}_1, \cdots, \hat{\psi}_q)'$ be a random vector with distribution $N(\hat{\psi}, \sigma^2 B)$ where $\psi = (\psi_1, \cdots, \psi_q)'$ and σ^2 are unknown parameters, and B is a known positive definite matrix. Let s^2 be an estimate of σ^2 independent of $\hat{\psi}$ such that $\nu s^2/\sigma^2$ has a chi-square distribution with ν degrees of freedom. In Scheffé ((1959) page 69) simultaneous confidence intervals for all linear functions of ψ are given; i. e. functions of the form $h'\psi$ where h is some given coefficient vector. The purpose of this section is to find simultaneous confidence intervals for all quadratic functions of ψ i. e. functions of the form $\psi'C\psi$ where C is some given symmetric $q \times q$ matrix. Equation (2.4) indicates why we are interested in quadratic functions of the unknown parameters, and the application to problems in regression analysis will be studied in more detail in Section 5.

Before stating the main theorem of this section, we will give a lemma, which is quite trivial, but which exhibits clearly the technique used in the proof of the theorem.

LEMMA 1. *Let the distribution of X depend upon parameters (θ, ξ), and let $\{S(x)\}$ be a family of $1 - \alpha$ confidence sets for θ. Let F be a family of functions of θ. Then a family of simultaneous $1 - \alpha$ confidence sets for the values of the functions in F is given by $\{f(S(x))\}$, $f \in F$, i. e. $P_{\theta,\xi}\{f(\theta) \in f(S(X)) \text{ for all } f \in F\} \geqq 1 - \alpha$.*

PROOF. We prove the lemma by proving that for all $f \in F$, $\{x: \theta \in S(x)\} \subset \{x: f(\theta) \in f(S(x))\}$. Suppose $x_0 \in \{x: \theta \in S(x)\}$, then $\theta \in S(x_0)$, and hence $f(\theta) \in f(S(x_0))$, for all $f \in F$.

This lemma gives us an easy way to find confidence sets for functions of parameters for which we already have a confidence set. We need only find the maps of $S(x)$ by the functions f for each x.

In the present problem $(\psi - \hat{\psi})' B^{-1}(\psi - \hat{\psi})/qs^2$ has an F-distribution with q and ν degrees of freedom. Hence

$$P[(\psi - \hat{\psi})' B^{-1}(\psi - \hat{\psi}) \leqq qs^2 F_{\alpha,q,\nu}] = 1 - \alpha\,, \tag{3.1}$$

where $F_{\alpha,q,\nu}$ is the upper α-point of the F-distribution with q and ν degrees of

freedom. Hence for observed values of $\hat{\phi}$ and s^2 we have a confidence set for ϕ. We shall apply Lemma 1 with $\theta = \phi$, and F equal to the family of all quadratic functions $\phi' C\phi$, where C varies over all symmetric matrices with real elements. Hence for each C we have to find the set of possible values of $\phi' C\phi$ when ϕ is contained in the ellipsoid $(\phi - \hat{\phi})' B^{-1}(\phi - \hat{\phi}) \leqq qs^2 F_{\alpha,q,\nu}$.

For a given matrix C there exists a nonsingular $q \times q$ matrix P so that $P'B^{-1}P = I$ and $P'CP = D$, where D is a diagonal matrix with real diagonal elements $d_1, \cdots, d_q$ equal to the roots of $|C - dB^{-1}| = 0$ (see, e. g., [1] pages 341–342). Define the vector $\hat{\gamma} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_q)'$ by $\hat{\gamma} = P'\hat{\phi}$. Then the problem above is equivalent to that of finding the set of possible values of $\gamma' D\gamma$ where γ is in the sphere $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) \leqq qs^2 F_{\alpha,q,\nu}$.

Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimum and maximum root, respectively, of

$$\sum_{i=1}^{q} d_i^2 \hat{\gamma}_i^2/(d_i - \lambda)^2 = qs^2 F_{\alpha,q,\nu} . \tag{3.2}$$

Let $a = -\min(\min_i d_i, \lambda_{\min})$ and $b = \max(\max_i d_i, \lambda_{\max})$. Then we have the following result.

THEOREM 1. *With the above assumptions and notation the probability is at least* $1 - \alpha$ *that simultaneously for all quadratic functions* $\phi' C\phi$ *we have*

$$A_1 \leqq \phi' C\phi \leqq A_2 ,$$

where

$$A_1 = a(\sum_{i=1}^{q} d_i \hat{\gamma}_i^2/(a + d_i) - qs^2 F_{\alpha,q,\nu})$$

and

$$A_2 = b(\sum_{i=1}^{q} d_i \hat{\gamma}_i^2/(b - d_i) + qs^2 F_{\alpha,q,\nu})$$

with exception of the following cases:

$$A_1 = 0 \quad \textit{if all} \quad d_i \geqq 0 \quad \textit{and} \quad \sum_{d_i>0} \hat{\gamma}_i^2 \leqq qs^2 F_{\alpha,q,\nu} ,$$

and

$$A_2 = 0 \quad \textit{if all} \quad d_i \leqq 0 \quad \textit{and} \quad \sum_{d_i<0} \hat{\gamma}_i^2 \leqq qs^2 F_{\alpha,q,\nu} .$$

REMARK 1. Note that $\hat{\gamma}$, D, $\lambda_{\min}$ and $\lambda_{\max}$ depend upon the particular matrix C. Hence for each C we have to calculate anew the values of these variables.

REMARK 2. Note that we can write

$$A_1 = \sum_{i=1}^{q} d_i \hat{\gamma}_i^2/(1 + d_i/a) - aqs^2 F_{\alpha,q,\nu}$$

and

$$A_2 = \sum_{i=1}^{q} d_i \hat{\gamma}_i^2/(1 - d_i/b) + bqs^2 F_{\alpha,q,\nu} .$$

Hence if d_i/a and d_i/b are small in absolute values, the confidence intervals are approximately of the form

$$[\sum_{i=1}^{q} d_i \hat{\gamma}_i^2 - aqs^2 F_{\alpha,q,\nu}, \sum_{i=1}^{q} d_i \hat{\gamma}_i^2 + bqs^2 F_{\alpha,q,\nu}] ,$$

which is intuitively reasonable.

PROOF. By the remarks preceding the theorem our problem is to determine the minimum and maximum of

$$T(\gamma) = \sum_{i=1}^{q} d_i \gamma_i^2$$

subject to $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) \leqq c^2$ where $c^2 = qs^2 F_{\alpha,q,\nu}$. We shall now consider the problem of finding the minimum (the problem of finding the maximum is analogous).

First consider the case when all $d_i \geqq 0$ and $\sum_{d_i>0} \hat{\gamma}_i^2 \leqq c^2$. Since the d_i are nonnegative $T(\gamma)$ is always $\geqq 0$. But in this case the value 0 is attained by setting $\gamma_i = 0$ for the indexes i with $d_i > 0$ and $\gamma_i = \hat{\gamma}_i$ for indexes i with $d_i = 0$.

Next consider the case when all $d_i \geqq 0$ and $\sum_{d_i>0} \hat{\gamma}_i^2 > c^2$. Then also $\sum_{i=1}^{q} \hat{\gamma}_i^2 > c^2$. Hence the sphere $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) \leqq c^2$ does not contain the origin. The ellipsoid $T(\gamma) = \sum_{i=1}^{q} d_i \gamma_i^2$ is centered at the origin. The minimum of $T(\gamma)$ when γ is in the sphere is obtained by expanding the ellipsoid until it touches the sphere. The minimum is therefore obtained at some point satisfying $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) = c^2$.

Finally, consider the case when not all $d_i \geqq 0$. We can write

$$T(\gamma) = \sum_{d_i<0} d_i(\gamma_i - \hat{\gamma}_i + \gamma_i)^2 + \sum_{d_i>0} d_i \gamma_i^2 .$$

Any value of $T(\gamma)$ with $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) < c^2$ can be decreased by increasing $|\gamma_i - \hat{\gamma}_i|$ for some γ_i with $d_i < 0$ and letting $\gamma_i - \hat{\gamma}_i$ have the same sign as $\hat{\gamma}_i$. Hence the minimum takes place for a γ on the sphere $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) = c^2$.

Similarly, it can be shown that, apart from the exception in the theorem, the maximum of $T(\gamma)$ is also attained for $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) = c^2$.

The solution to the problem of finding the minimum and maximum of $\gamma'/D\gamma$ subject to $(\gamma - \hat{\gamma})'(\gamma - \hat{\gamma}) = c^2$ can be found from the results in a paper by Forsythe and Golub (1965). They consider the problem of finding all stationary points of $\Phi(x) = (x - b)'A(x - b)$ subject to $x'x = 1$. Our problem can be transformed to theirs by making the correspondence $x = (\gamma - \hat{\gamma})/c$, $b = -\hat{\gamma}/c$, $D = A$, $\Phi(x) = T(\gamma)/c^2$. The characteristic roots $\lambda_1, \cdots, \lambda_n$ of A correspond to $d_1, \cdots, d_q$. From Theorem (4.1) and Equation (3.10) of Forsythe and Golub we obtain the expressions A_1 and A_2 in our theorem. This completes the proof.

It may be of interest to see the form of the confidence interval given in Theorem 1 in the case when $\psi' C\psi = \varphi^2$, where $\varphi = a'\psi$ is a linear function of ψ. In this case, we will directly construct the variables γ and $\hat{\gamma}$ used in the proof of the theorem. Let $\hat{\gamma}_1 = (a'Ba)^{-\frac{1}{2}} a'\hat{\psi}$, and adjoin $\hat{\gamma}_2, \cdots, \hat{\gamma}_q$ so that they are all independent with variance σ^2. We have $\gamma_1 = (a'Ba)^{-\frac{1}{2}}\varphi$, $d_1 = a'Ba$ and $d_2 = \cdots = d_q = 0$. By using Theorem 1 we get the interval

$$d_1(\max(0, |\hat{\gamma}_1| - c))^2 \leqq \varphi^2 \leqq d_1(|\hat{\gamma}_1| + c)^2 ,$$

where the form of the left member of the inequality reflects the possibility of the exception in the theorem. Let $\hat{\varphi} = a'\hat{\psi}$. An estimate of Var $\hat{\varphi}$ is $d_1 s^2$; call this $\sigma_{\hat{\varphi}}^2$. With this notation the above confidence interval can be written

$$(3.3) \qquad (\max(0, |\hat{\varphi}| - \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}}))^2 \leqq \varphi^2 \leqq (|\hat{\varphi}| + \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}})^2 .$$

The confidence interval for φ obtained by Scheffé's S-method of multiple comparison is

$$\hat{\varphi} - \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}} \leqq \varphi \leqq \hat{\varphi} + \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}} . \tag{3.4}$$

It is easily seen that the interval (3.3) consists of those φ^2 for which φ is in the interval (3.4). Note that we cannot deduce (3.4) from (3.3), since φ^2 is not a one-one function of φ.

If for a particular problem we are interested in both quadratic and linear functions at the same time, we can use the S-method for linear functions, and the method given in Theorem 1 for quadratic functions. The probability that all confidence intervals cover the true values would be $1 - \alpha$. This follows from Lemma 1. The following two inequalities can be helpful when one has to determine $\lambda_{\min}$ and $\lambda_{\max}$ numerically.

Let F be the set of indexes i for which $\hat{\gamma}_i \neq 0$. We have

$$\max_{i \in F} d_i < \lambda_{\max} \leqq \max_{i \in F} d_i + (\textstyle\sum_{i=1}^{q} d_i^2 \hat{\gamma}_i^2/c^2)^{\frac{1}{2}} , \tag{3.5}$$

and

$$\min_{i \in F} d_i - (\textstyle\sum_{i=1}^{q} d_i^2 \hat{\gamma}_i^2/c^2)^{\frac{1}{2}} \leqq \lambda_{\min} \leqq \min_{i \in F} d_i . \tag{3.6}$$

To prove this we can compare (3.5) and

$$\textstyle\sum_{i=1}^{q} d_i^2 \hat{\gamma}_i^2/(\max_{i \in F} d_i - \lambda)^2 = c^2 . \tag{3.7}$$

As a function of λ the left member of (3.7) is always at least as large as the left member of (3.2) when $\lambda > \max_{i \in F} d_i$. Hence the solution of (3.7) with respect to λ with $\lambda > \max_{i \in F} d_i$ is greater than the corresponding solution of (3.2) which is $\lambda_{\max}$. The solution of (3.7) with $\lambda > \max_{i \in F} d_i$ is $\max_{i \in F} d_i + (\sum_{i=1}^{q} d_i^2 \hat{\gamma}_i^2/c^2)^{\frac{1}{2}}$. This proves (3.5). (3.6) is proved in a similar way.

Scheffé's method of multiple comparison of all linear functions is such that the set of points ψ satisfying all confidence statements about the linear functions is equal to the set of points satisfying $(\psi - \hat{\psi})' B^{-1}(\psi - \hat{\psi}) \leqq c^2$. Hence the two sets have the same probability $1 - \alpha$. The next theorem gives the corresponding result for the above multiple comparison method of quadratic functions.

THEOREM 2. *The probability that all the confidence statements in Theorem* 1 *are true simultaneously is* $P\{((\psi - \hat{\psi})' B^{-1}(\psi - \hat{\psi}) \leqq qs^2 F_{\alpha,q,\nu}) \cup ((\psi + \hat{\psi})' B^{-1}(\psi + \hat{\psi}) \leqq qs^2 F_{\alpha,q,\nu}))\}$. *This is greater than* $1 - \alpha$ *if* $\psi \neq 0$, *and equal to* $1 - \alpha$ *if* $\psi = 0$.

PROOF. We prove the theorem by showing that the set of ψ and $\hat{\psi}$ satisfying all the inequalities of the confidence sets in Theorem 1 is equal to the set of all ψ and $\hat{\psi}$ satisfying

$$(\psi - \hat{\psi})' B^{-1}(\psi - \hat{\psi}) \leqq qs^2 F_{\alpha,q,\nu} \tag{3.8}$$

or

$$(\psi + \hat{\psi})' B^{-1}(\psi + \hat{\psi}) \leqq qs^2 F_{\alpha,q,\nu} . \tag{3.9}$$

If ψ and $\hat{\psi}$ satisfy (3.8) then it is shown in Theorem 1 that ψ and $\hat{\psi}$ satisfy all

the confidence statements. The same is true if ψ and $\hat{\psi}$ satisfy (3.9) since changing the sign of ψ or $\hat{\psi}$ will not change the confidence statements.

Conversely, if all the inequalities of the confidence statements are satisfied, it follows that (3.3) is satisfied for all linear functions φ. (3.3) implies

$$\hat{\varphi} - \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}} \leqq \varphi \leqq \hat{\varphi} + \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}}$$

or

$$\hat{\varphi} - \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}} \leqq -\varphi \leqq \hat{\varphi} + \hat{\sigma}_{\hat{\varphi}}(qF_{\alpha,q,\nu})^{\frac{1}{2}} .$$

Since this is true for all linear functions φ, it follows by the equivalence of the F-test and the S-method (see Scheffé 1959)) that either (3.8) or (3.9) are satisfied.

4. Multiple testing. From the results of Section 3 we can derive simultaneous tests of hypotheses

$$H\colon \psi' C\psi = 0$$

for any set of symmetric matrices C. We reject the hypothesis $\psi' C\psi = 0$ if and only if the confidence interval for $\psi' C\psi$ given in Theorem 1 does not cover the point 0. The overall significance level of these simultaneous tests is $\leqq \alpha$, since if we reject one of the hypotheses falsely, then the corresponding confidence interval does not cover the true parameter value (which is 0). The probability of this is $\leqq \alpha$ by Theorem 2, hence the probability of false rejections is $\leqq \alpha$.

5. Multiple comparison of regression functions. We now return to the problem of Section 2. Assume that the most comprehensive regression function is $X'\beta$. Our estimate of β is $\hat{\beta} = (XX')^{-1}Xy$ which is $N(\beta, (XX')^{-1}\sigma^2)$ where by definition we have $\beta = (XX')^{-1}X\eta$. We have

$$P\{(\beta - \hat{\beta})' XX'(\beta - \hat{\beta}) \leqq ps^2 F_{\alpha,p,n-p}\} \geqq 1 - \alpha , \tag{5.1}$$

where s^2 is given by (2.2). Equality holds in (5.1) if $Ey = X'\beta$. If $Ey \neq X'\beta$, then $(n - p)s^2/\sigma^2$ has a noncentral chi-square distribution. To prove that (5.1) also holds in that case we can use the technique of a proof in Scheffé (1959) pages 136–137.

Consider the problem of comparing two regression functions $X_1'\beta_1$ and $X_2'\beta_2$ where $C(X_i') \subset C(X')$, rank $X_i = p_i$, $i = 1, 2$. The goodness of fit of $X_i'\beta_i$ is measured by $\eta' X_i'(X_iX_i')^{-1}X_i\eta$, which is the squared length of the projection $X_i'(X_iX_i')^{-1}X_i\eta$ of η on $C(X_i')$. This projection can also be written in the form $X_i'(X_iX_i')^{-1}X_iX'(XX')^{-1}X\eta$ since $C(X_i') \subset C(X')$, and $X'(XX')^{-1}X\eta$ is the projection of η on $C(X')$. Using the last expression for the projection of η on $C(X_i')$ we find that the goodness is measured by $\beta' XX_i'(X_iX_i')^{-1}X_iX'\beta$. Hence the difference of goodness between the two regression functions is

$$\beta'(XX_2'(X_2X_2')^{-1}X_2X' - XX_1'(X_1X_1')^{-1}X_1X')\beta . \tag{5.2}$$

When we compare all possible regression functions with the column space contained in $C(X')$, we have by (5.2) a problem of multiple comparison of

functions of the form $\beta' C\beta$ where a confidence region for β is given in (5.1). A solution to this problem is given in Theorem 1 in Section 3, where β corresponds to ψ and (5.1) corresponds to (3.1). In Section 3 it was required that $\nu s^2/\sigma^2$ has a central chi-square distribution. In this section this chi-square distribution might be noncentral, in which case we have strict inequality in (5.1). Theorem 1 still holds, since what was important in the proof was that the probability (3.1) is $1 - \alpha$ or greater.

The equation for determining the roots d_i is

$$|XX_2'(X_2X_2')^{-1}X_2X' - XX_1'(X_1X_1')^{-1}X_1X' - dXX'| = 0 . \tag{5.3}$$

We shall now derive a lemma stating that many of the roots are 0 or 1. Suppose without loss of generality that $p_1 \leqq p_2$. Let p_0 be the dimension of the intersection of $C(X_1')$ and $C(X_2')$.

LEMMA 2. *In* (5.3) $p - p_1 - p_2 + 2p_0$ *roots are equal to* 0, $p_2 - p_1$ *roots are equal to* 1 *and the remaining* $2(p_1 - p_0)$ *roots occur in pairs of the form* $\pm d_i$, $i = 1, \cdots, p_1 - p_0$.

PROOF. Let $\alpha_{p_1+p_2-2p_0+1}, \cdots, \alpha_{p_1+p_2-p_0}$ be an orthonormal basis for the intersection. Choose vectors $\alpha_1, \cdots, \alpha_{p_1+p_2-2p_0}$ so that $\alpha_1, \cdots, \alpha_{p_1-p_0}, \alpha_{p_1+p_2-2p_0+1}, \cdots, \alpha_{p_1+p_2-p_0}$ constitute an orthonormal basis for $C(X_1')$ and $\alpha_{p_1-p_0+1}, \cdots, \alpha_{p_1+p_2-p_0}$ constitute an orthonormal basis for $C(X_2')$. Adjoin orthonormal vectors $a_{p_1+p_2-p_0+1}, \cdots, \alpha_p$ orthogonal to the $p_1 + p_2 - p_0$ first vectors so that the set of all the vectors span $C(X')$. The vectors can be chosen so that $\alpha_i'\alpha_j = 0$, $i = 1, \cdots, p_1 - p_0$, $j = 2p_1 - 2p_0 + 1, \cdots, p_1 + p_2 - 2p_0$.

Define matrices Q_1, Q_2 and Q_3 by $Q_1 = (\alpha_1, \cdots, \alpha_{p_1-p_0})$, $Q_2 = (\alpha_{p_1-p_0+1}, \cdots, \alpha_{2p_1-2p_0})$ and $Q_3 = (\alpha_{2p_1-2p_0+1}, \cdots, \alpha_p)$. We have $Q_1'Q_i = I$, $i = 1, 2, 3$, $Q_i'Q_3 = 0$, $i = 1, 2$. (If $p_1 = p_0$, the matrices Q_1 and Q_2 do not exist. We shall return to that case in Section 6(a)). Define the vector $\theta = (\theta_1, \cdots, \theta_p)'$ by $\theta_i = \alpha_i'\eta$, $i = 1, \cdots, p$. We can also write

$$\theta = [Q_1Q_2Q_3]'\eta . \tag{5.4}$$

In terms of θ, (5.2) can be written

$$\sum_{i=p_1-p_0+1}^{p_1+p_2-2p_0} \theta_i^2 - \sum_{i=1}^{p_1-p_0} \theta_i^2 = \theta' G\theta , \tag{5.5}$$

where the equality (5.5) defines the matrix G.

Define $\hat{\theta}$ by

$$\hat{\theta} = [Q_1Q_2Q_3]'y . \tag{5.6}$$

The covariance matrix of $\hat{\theta}$ is $A\sigma^2$ where

$$A = \begin{bmatrix} I & Q_1'Q_2 & 0 \\ Q_2'Q_1 & I & 0 \\ 0 & 0 & I \end{bmatrix} . \tag{5.7}$$

Since $\hat{\theta}$ is a nonsingular linear transform of $\hat{\beta}$, (5.3) can be written $|G - dA^{-1}| = 0$.

Using (5.5) and (5.7) we find that (5.3) reduces to

$$(5.8) \qquad (1-d)^{p_2-p_1}d^{p-p_1-p_2+2p_0}\left|\begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} - d\begin{bmatrix} I & Q_1'Q_2 \\ Q_2'Q_1 & I \end{bmatrix}^{-1}\right| = 0 .$$

Hence we immediately obtain $p - p_1 - p_2 + 2p_0$ roots d_i equal to 0, and $p_2 - p_1$ roots equal to 1.

After multiplying (5.8) by the determinant of the matrix

$$\begin{bmatrix} I & Q_1'Q_2 \\ Q_2'Q_1 & I \end{bmatrix}$$

we find that the equation for the remaining roots of (5.8) becomes

$$(5.9) \qquad \left|\begin{matrix} -(1+d)I & -Q_1'Q_2 \\ Q_2'Q_1 & (1-d)I \end{matrix}\right| = 0 .$$

The determinant in (5.9) can be evaluated as (see e. g. Anderson (1957) page 344),

$$|(1-d)I|\,|-(1+d)I + (1-d)^{-1}Q_1'Q_2Q_2'Q_1| .$$

Hence (5.9) can be written

$$(5.10) \qquad |Q_1'Q_2Q_2'Q_1 - (1-d)^2 I| = 0 .$$

It follows that if $e_1, \cdots, e_{p_1-p_0}$ are the characteristic roots of the positive definite matrix $Q_1'Q_2Q_2'Q_1$, then the remaining $2(p_1 - p_0)$ roots of (5.3) are given by $d = \pm(1-e_i)^{\frac{1}{2}}$, $i = 1, \cdots, p_1 - p_0$. Hence instead of p unknown roots in (5.3), there are only $p_1 - p_0$ unknown. This completes the proof of the lemma.

As for the vector $\hat{\gamma}$ used in Theorem 1 it is not necessary to determine the $\hat{\gamma}_i$ corresponding to $d_i = 0$ since for these $\hat{\gamma}_i$ we have $d_i\hat{\gamma}_i = 0$, and they will not contribute to the values of A_1 and A_2 in Theorem 1. The $\hat{\gamma}_i$ corresponding to $d_i = 1$ can be chosen equal to $\hat{\theta}_i$, $i = 2p_1 - 2p_0 + 1, \cdots, p_1 + p_2 - 2p_0$.

It can be instructive to relate some of the quantities introduced above to the residual sums of squares. The residual sum of squares using a certain regression equation $X_i'\beta_i$ is

$$SS(X_i') = (y - X_i'\hat{\beta}_i)'(y - X_i'\hat{\beta}_i) .$$

In particular we have $SS(X') = (n-p)s^2$. We can write $SS(X_i') = y'y - (X_i'\hat{\beta}_i)'(X_i'\hat{\beta}_i)$. Hence $SS(X_i')$ is the squared length of y minus the squared length of the projection of y on $C(X_i')$. Consider now two functions $X_i'\beta_i$ and $X_2'\beta_2$. In terms of $\hat{\theta}$ we get

$$(5.11) \qquad SS(X_1') - SS(X_2') = \sum_{i=p_1-p_0+1}^{p_1+p_2-2p_0}\hat{\theta}_i^2 - \sum_{i=1}^{p_1-p_0}\hat{\theta}_i^2 .$$

This can be regarded as an estimate of (5.2) and (5.5). It is not unbiased, unless $p_2 = p_2$, since we have

$$E(SS(X_1') - SS(X_2')) = \sum_{i=p_1-p_0+1}^{p_1+p_2-2p_0}\theta_i^2 - \sum_{i=1}^{p_1-p_0}\theta_i^2 + (p_2 - p_1)\sigma^2 .$$

We conclude this section with some remarks. In some cases we have some

variables which we want to include in each regression equation. Suppose that we have $p - p^*$ of these variables and p^* other variables. A regression function $X'\beta$ can then be written in the form $X^{0\prime}\beta^0 + X^{*\prime}\beta^*$, where $X^{0\prime}\beta^0$ refers to the variables always included, $X^{*\prime}\beta^*$ refers to the other p^* variables, and $C(X^{0\prime}) \perp C(X^{*\prime})$. In that case we can start with

$$P\{(\beta^* - \hat{\beta}^*)'X^*X^{*\prime}(\beta^* - \hat{\beta}^*) \leqq s^2 p^* F_{\alpha;p^*,n-p}\} \geqq 1 - \alpha ,$$

instead of (5.1), and apply our multiple comparison procedure with β^* and $\hat{\beta}^*$ instead of β and $\hat{\beta}$, and the projection of η on $C(X^{*\prime})$ instead of the projection on $C(X')$.

Application of the above multiple comparison method will not usually give a unique best regression equation, but a set of many equations none of which gives a "significantly better" fit than any other on the basis of the statistical criterion adopted here. It seems to the present author that this is the most we can hope for when he compares it with the S-method and T-method in the analysis of variance [11].

6. Some special cases. (a) $p_1 = p_0$. This means that the regression equation $X_2'\beta_2$ is obtained from $X_1'\beta_1$ by adding $p_2 - p_1$ independent variables.

In this case the difference (5.5) is $\sum_{i=1}^{p_2-p_1} \theta_i^2$; the matrices Q_1 and Q_2 do not exist; all $p_2 - p_1$ nonzero solutions of (5.8) are equal to 1; and, the covariance matrix of $\hat{\theta}$ is the identity matrix times σ^2. As $\hat{\gamma}_1, \cdots, \hat{\gamma}_{p_2-p_1}$ we can use $\hat{\theta}_1, \cdots, \hat{\theta}_{p_2-p_1}$. Equation (3.2) is $\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2/(1 - \lambda)^2 = c^2$, with solutions

$$\lambda_{\min} = 1 - (\textstyle\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2)^{\frac{1}{2}}/c \quad \text{and} \quad \lambda_{\max} = 1 + (\textstyle\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2)^{\frac{1}{2}}/c .$$

Using Theorem 1 we find the confidence interval for $\sum_{i=1}^{p_2-p_1} \theta_i^2$ to be

$$(\max (0, (\textstyle\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2)^{\frac{1}{2}} - c))^2 \leqq \textstyle\sum_{i=1}^{p_2-p_1} \theta_i^2 \leqq ((\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2)^{\frac{1}{2}} + c)^2 .$$

The confidence interval does not cover 0 if and only if $\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2 > c^2 = s^2 p F_{\alpha;p,n-p}$. Only in that case we say that $X_2'\beta_2$ is better that $X_1'\beta_1$. In the particular case $p_2 = p_1 + 1$, we get the inequality $|\hat{\theta}_1| > s(pF_{\alpha;p,n-p})^{\frac{1}{2}}$. If we had used the usual t-test for judging whether the additional variable should be included, we would have included the new variable if and only if $|\hat{\theta}_1| > s(F_{\alpha;1,n-p})^{\frac{1}{2}}$. We shall discuss this further in Section 7. From (5.11) we have

$$\textstyle\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2 = SS(X_1') - SS(X_2') .$$

(b) $p_1 - p_0 = 1$. This case includes the situation where at a certain step we want to choose between several single additional independent variables. (Also, in that case, $p_2 - p_0 = 1$.) This special case where $p_1 - p_0 = 1$ is particularly important, since it refers to the kind of decision problem one has to face repeatedly with most stepwise regression methods. The vectors α_1 and α_2, which are all we need, are easily determined. The difference (5.5) is $\sum_{i=2}^{p_2-p_1+2} \theta_i^2 - \theta_2^2$, and the product $Q_1'Q_2$ is a scalar a, say. Equation (5.10) has two nonzero solutions $d_1 = -(1 - a^2)^{\frac{1}{2}}$ and $d_2 = (1 - a^2)^{\frac{1}{2}}$. Besides we have the root $d = 1$

of multiplicity $p_2 - p_1$.

(c) $Q_1'Q_2 = 0$. In this case the column vectors of X_1' and X_2' which do not belong to $C(X_1') \cap C(X_2')$ are orthogonal. This is a situation which is not likely to occur very often, but the computations are simple in this case. The covariance matrix (5.7) of $\hat{\theta}$ is $I\sigma^2$, and (5.8) has nonzero roots $d = 1$ with multiplicity $p_2 - p_0$, and $d = -1$ with multiplicity $p_1 - p_0$. As $\hat{\gamma}_1, \cdots, \hat{\gamma}_{p_1+p_2-2p_0}$ we can use $\hat{\theta}_1, \cdots, \hat{\theta}_{p_1+p_2-2p_0}$; equation (3.2) is

$$\sum_{i=1}^{p_1-p_0} \hat{\theta}_i^2/(1+\lambda)^2 + \sum_{i=p_1-p_0+1}^{p_1+p_2-2p_0} \hat{\theta}_i^2/(1-\lambda)^2 = c^2 .$$

7. An example. The following example is taken from [6]. The problem is to study the deflection d of a metal rod of length L under a constant load of 400 grams. The observations are

L	55	60	65	70	75	80	85	90 cm
d	1.165	1.518	1.948	2.428	2.965	3.610	4.242	5.010 mm.

First the data were analysed by using orthogonal polynomials. Only the first three terms in the polynominal were significant, and the resulting regression equation is

$$\hat{\delta} = (143.350 - 0.74896L + 0.12728L^2)10^{-2} , \tag{7.1}$$

where δ is the estimate of $E(d)$. The corresponding residual sum of squares is $SS(X_1') = 0.0011$. Thereafter the data were analysed by a stepwise procedure. As possible candidates were considered the constant term, L, L^2, L^3, L^4. The variable L^3 was taken into the regression equation first, since it gives the smallest residual sum of squares. After that, when L^3 is already in the regression equation, L^4 reduces the residual sum of squares most, and finally L^2 is taken into the equation. No one of the others gives additional significant contribution to the regression equation. The resulting equation is

$$\hat{\delta} = (-1.6659L^2 + 0.12059L^3 - 0.000372L^4)10^{-4} , \tag{7.2}$$

and the corresponding residual sum of squares is $SS(X_2') = 0.0011$.

By the theory of elasticity, however, the deflection should vary as L^3. If we use only L^3 we get the regression equation

$$\hat{\delta} = 0.69586.10^{-5}L^3 , \tag{7.3}$$

with residual sum of squares $SS(X_3') = 0.0113$.

The results of the above analysis are used in [6] as a warning against indiscriminate use of orthogonal polynomials. The discrepancy between (7.2) and (7.3) could be caused by imperfection in the experiment, or the fact that the physical law is not holding exactly.

Let us now compare the regression equation (7.3) with (7.1) and (7.2) by the method developed in this paper. In the example $n = 8$ and $p = 5$. We will use $\alpha = 0.05$. To compare (7.2) and (7.3) is particularly easy, since this is a

case with $p_1 = p_0 = 1$. We have $p_2 = 3$. By (a) of Section 6 we shall reject the hypothesis that the equations are equally good if $\sum_{i=1}^{p_2-p_1} \hat{\theta}_i^2 > s^2 p F_{\alpha,p,n-p}$. The right member of this inequality is $(0.0011/3) \cdot 5 \cdot 9.01 = 0.0165$, while from (5.11) the left member is $0.0113 - 0.0011 = 0.0102$. Hence we cannot claim that (7.2) is an improvement over (7.3).

When comparing (7.1) and (7.3) we have situation (b) in Section 6. The confidence interval is found to be $[-0.0098, 0.0537]$. It covers zero, hence it cannot be claimed that (7.1) is better than (7.3).

One of the defects of the stepwise procedure is well illustrated by the reasoning that led to (7.2). First L^3 was introduced into the regression equation. With L^3 in the regression equation it was determined which of the variables L^0, L, L^2, L^4 gave the largest reduction in the residual sum of squares. This reduction in the residual sum of squares, call it S_M, is then compared with the total residual mean square, s^2. The significance of the contribution is judged by comparing S_M/s^2 and the upper α-point of the F-distribution with 1 and $n - p$ (or $n - 2$ if one uses only the two variables introduced so far) d. f.. But S_M/s^2 does not have an F-distribution, because S_M is the maximum of four different sums of squares.

The weakness of the method presented in this paper is that it will give rather wide confidence intervals. The main reason for this is that we are constructing simultaneous confidence intervals for a very broad class of quadratic functions. The class is much larger than the class of quadratic functions we would ordinarily wish to consider. From what is found in Sections 3 and 6(a), the lengths of the intervals should be comparable to those obtained with Scheffé's S-method. The lengths of the intervals increase with p (or p^*). We should therefore be careful when selecting variables to be tried in the regression equation.

In the above example it seems reasonable a priori to exclude the constant term from the regression equation. Furthermore, from the theory of elasticity, L^3 should be a member of all regression equations. Hence we should work with $p = 4$ and $p^* = 3$. The additional reduction of the residual sum of squares obtained by introducing L^4 when L^3 already is included is 0.0079. This should be compared with $s^2 p^* F_{\alpha;p^*,n-p} = (0.0011/3)3 \cdot 6.59 = 0.0073$. Hence by this analysis L^4 is significant. But it is found that no other terms should be included.

Acknowledgment. The author wishes to thank Professor Henry Scheffé and Professor Erling Sverdrup for reading this paper and suggesting several improvements; the Associate Editor for drawing my attention to the paper by Forsythe and Golub (1965); and, the referees for various comments.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

[2] ANDERSON, T. W. (1962). The choice of the degree of a regression as a multiple decision problem. *Ann. Math. Statist.* **33** 255–265.

[3] BEALE, E. M. L., KENDALL, M. G. and MANN, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika* **54** 357–366.
[4] DRAPER, N. R. and SMITH, H. (1966). *Applied Regression Analysis*. Wiley, New York.
[5] FORSYTHE, G. E. and GOLUB, G. H. (1965). On the stationary values of a second-degree polynomial on the unit sphere. *SIAM J.* **13** 1050–1068.
[6] HAMAKER, H. C. (1962). On multiple regression analysis. *Statistica Neerlandica* **16** 31–56.
[7] HOTELLING, H. (1940). The selection of variates for use in prediction with some comments on the problem of nuisance parameters. *Ann. Math. Statist.* **11** 271–283.
[8] LAMOTTE, I. R. and HOCKING, R. R. (1970). Computational efficiency in the selection of regression variables. *Technometrics* **12** 83–93.
[9] MALLOWS, C. L. (1966). Choosing a subset regression. Bell Technical Report.
[10] SCHATZOFF, M., TSAO, R. and FIENBERG, S. (1968). Efficient calculation of all possible regressions. *Technometrics* **10** 769–779.
[11] SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.
[12] WILLIAMS, E. C. (1959). *Regression Analysis*. Wiley, New York.