EVALUATION OF REGRESSION COEFFICIENT ESTIMATES USING α-ACCEPTABILITY

by

George P. McCabe
Purdue University &
Princeton University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #492

EVALUATION OF REGRESSION COEFFICIENT ESTIMATES USING α-ACCEPTABILITY

George P. McCabe, Jr.

Purdue University & Princeton University

ABSTRACT

A framework, called α-acceptability, is proposed for evaluating alternatives to the usual least squares estimates of regression coefficients. An estimate is defined to be α-acceptable if it is in the usual (1-α) 100% confidence region. Estimates which are α-acceptable for large values of α, such as .99 (corresponding to a 1% confidence region) are viewed as statistically indistinguishable from the usual estimates and, in a sense, simply rounded off. Applications to subset selection, ridge regression, regression on principal components and a class of minimax procedures are discussed and illustrated with an example.

Key Words

Multiple Linear Regression
Subset Selection
Ridge Regression
Confidence Regions

1. INTRODUCTION

Under normality assumptions, the maximum likelihood estimators of regression coefficients in multiple regression problems are given by the principle of least squares. Although some numerical problems can arise in the presence of multicollinearity, these estimators are simple to evaluate and the associated distribution theory is straightforward. Computer programs which perform the necessary calculations are widely available. The use of least squares estimators is so widespread today that it could be properly described as standard statistical practice.

The advisability of using least squares estimators was seriously questioned by two key developments. On the theoretical side, the fundamental work of Stein [26] indicated that these estimators were inadmissible and thus could be improved upon. These results hold true whether or not multicollinearity is present. Numerous papers ([2], [5], [7], [8], [11], [27]) have appeared which extend and amplify this fundamental theme. On the applied side, Hoerl and Kennard [16], [17] have proposed ridge regression as a technique for coping with difficulties arising from multicollinearity. These ideas have also been extended and amplified by numerous authors (see Hocking [15] for references.)

Beaton, Rubin and Barone [6] have suggested that the numerically accurate least squares solution may be the right solution to the wrong problem. Using simulation methods, Dempster, Schatzoff and Wermuth [10] have compared 56 different alternatives to least squares. They found that substantial improvements can be obtained by such alternatives.

To the practioner, the message is clear: substantial benefits can be gained by deviating from least squares procedures. Some types of deviations from least squares are, in fact, quite common. Whenever one uses a subset selection procedure to obtain an equation with a reduced number of variables, one compromises the least squares principle by setting some coefficients equal to zero. Whether or not least squares is used for the reduced problem, a deviation from the least squares estimates for the full model is present.

If one desires to take advantage of the purported benefits to be gained by abandoning the standard least squares estimators, a suitable alternative from a large collection of procedures must be chosen. A variety of admissible classes of estimators exist and many ridge-type estimators are also available. There are, in addition, numerous methods for selecting variables, i.e. setting some coefficients equal to zero. Hocking [15] has given an excellent survey of this topic.

In this paper, a framework for evaluating estimates of regression coefficients is proposed. The idea is basically very

simple and uses the standard confidence regions for regression coefficients. Similar ideas have been suggested by McDonald [22], McDonald and Schwing [23] and Obenchain [25] for ridge regression and Aitkin [1] for subset selection. Cook [9] has proposed a related framework for detecting outliers. When applied in the subset selection context, interpretation in the framework proposed by Arvesen and McCabe [3], [4] and McCabe and Arvesen [20] is possible.

2. PRELIMINARIES

The statistical model underlying this discussion is the standard multiple regression model given by

$$Y = X\beta+\varepsilon \qquad (2.1)$$

where

$$Y' = (Y_1,\ldots,Y_n)$$

is an observable random variable,

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{m1} & & X_{np} \end{pmatrix}$$

is a full rank design matrix of known constants, and

$$\beta' = (\beta_0,\beta_1,\ldots,\beta_p)$$

is an unknown parameter vector of regression coefficients. The error vector

$$\varepsilon' = (\varepsilon_1, \ldots, \varepsilon_n)$$

is assumed to be normally distributed with zero mean and covariance matrix $\sigma^2 I$ where σ^2 is unknown, i.e.

$$\varepsilon \sim N(o, \sigma^2 I).$$

In short, we can write

$$Y \sim N(X\beta, \sigma^2 I). \qquad (2.2)$$

Note that the intercept term β_o is included in the model. Since the inferences to be made concern the entire parameter vector, it is inappropriate to standardize and neglect this parameter in the present context. For the rare situation where there is a fundamental a priori reason why this term should not be included in the model, trivial modifications to the present development can be made.

The least squares (maximum likelihood) estimator of β is given by

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y. \qquad (2.3)$$

It easily follows that

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2(X'X)^{-1}). \tag{2.4}$$

For this estimator, the sum of squared errors is given by

$$SSE(\hat{\beta}_{LS})=\Sigma_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \tag{2.5}$$

where

$$\hat{Y}_i= \hat{\beta}_o+\hat{\beta}_1 X_{1i}+\ldots+\hat{\beta}_p X_{pi}$$

It is easy to show that

$$SSE (\hat{\beta}_{LS}) = Y' [I-X(X'X)^{-1}X']Y. \tag{2.6}$$

The usual unbiased estimate of the parameter σ^2 is given by the sum of squared errors divided by its degrees of freedom:

$$s^2=SSE(\hat{\beta}_{LS}) / (n-p-1) \tag{2.7}$$

For any $\alpha \in (0,1)$, a $(1-\alpha)$ 100% confidence region for the parameter β is given by

$$S_\alpha=\{\beta: (\hat{\beta}_{LS}-\beta)' (X'X)(\hat{\beta}_{LS}-\beta)<s^2(p+1)F_{p+1,n-p-1;1-\alpha}\} \tag{2.8}$$

where $F_{p+1,n-p-1;1-\alpha}$ is the upper α quantile of the F distribution with p+1 and n-p-1 degrees of freedom. The set S_α is the interior of a hyperellipsoid in (p+1) - dimensional space.

3. ACCEPTABLE ESTIMATES

For any (p+1) dimensional vector β, let

$$D(\beta) = (\hat{\beta}_{LS} - \beta)' (X'X) (\hat{\beta}_{LS} - \beta). \qquad (3.1)$$

Clearly, $D(\beta)$ is a distance measure which indicates how far β is from $\hat{\beta}_{LS}$ in the appropriate metric. Since

$$S_\alpha = \{\beta: D(\beta) \leq d_\alpha\} \qquad (3.2)$$

where

$$d_\alpha = (p+1)\, s^2\, F_{p+1,n-p-1;1-\alpha},$$

the confidence region may be viewed as composed of those β's which are not too far away from $\hat{\beta}_{LS}$.

At this point, it is helpful to recall the connection between confidence regions and hypothesis tests. Let α be fixed and let β denote the true parameter vector. Then,

$$P_\beta(D(\beta) \leq d_\alpha) = 1-\alpha \qquad (3.3)$$

where P_β denotes probability calculated under the assumption that β is the true parameter vector. Values of β for which $D(\beta) > d_\alpha$ can therefore be rejected or viewed as unacceptable. Such values of β are too far from $\hat{\beta}_{LS}$ to be viewed as reasonable candidates for the true parameter.

Definition 1. Values of β for which

$$D(\beta) \leq d_\alpha$$

are called α-acceptable.

Note that S_α is the set of α-acceptable β's. Also, for $\alpha > \alpha'$, $S_\alpha \subset S_{\alpha'}$. Therefore if β is α-acceptable, it is also α'-acceptable for all $\alpha' < \alpha$.

Definition 2. For any β, the value of α for which

$$D(\beta) = d_\alpha$$

is called its acceptance level.

For any estimate $\hat{\beta}$, the corresponding value of $D(\hat{\beta})$, gives an idea of how far the estimate is from least squares. The acceptance level of the estimate gives a different quantification of this distance which can be readily interpreted. Roughly speaking, a practicioner should be very reluctant to use an estimate with an acceptance level of .001. Such an estimate is simply too far from $\hat{\beta}_{LS}$ to be compatible with the data. Alternatively, one could argue that if one were to hypothesize that this estimate represented the true value of β, the hypothesis could be soundly rejected at the .001 level of significance. On the other hand, consider an estimate with an acceptance level of .90. Such

an estimate is close to $\hat{\beta}_{LS}$ (in the appropriate metric). A null hypothesis that this estimate represented the true value of β could not be rejected unless one were willing to tolerate a 90% Type I error. In terms of the data, such a modification of the least squares estimate is simply a minor one when viewed relative to the natural variation in the problem. If the estimate has other (possibly) desirable properties such as being minimax or ridge or having a number of zero coefficients, then one can use it while being confident that it is really not very different (statistically) from the least squares estimate. Of course, undesirable properties can result by always taking some extreme point in the confidence region. Consider, for example, the simple location problem where only the term β_0 is present. If one always estimated β_0 by the upper 95% confidence bound, a substantial bias would result. However, such an estimate is still consistent and might be desirable under circumstances which could reasonably arise in practice.

The point of the present development is not to propose or justify alternative estimates to least squares. The literature abounds with such material. The aim of the present study is simply to provide a reasonable framework for using these alternative procedures.

4. SOME USEFUL FACTS

Many textbooks give the confidence region (3.2) and comment that this result is difficult to use in practice when the dimensions are large. While it is true that we cannot easily picture hyperellipsoids in high dimensional spaces, an alternative characterization of S_α is very useful in practice.

The following simple results will be used in subsequent sections to evaluate estimates. Let $\hat{\beta}$ denote any estimate of β and let

$$SSE(\hat{\beta}) = (Y-X\hat{\beta})' (Y-X\hat{\beta}). \tag{4.1}$$

This quantity is the sum of squared errors obtained by using $\hat{\beta}$ as an estimate of β.

Fact 1. $$D(\hat{\beta}) = SSE(\hat{\beta}) - SSE(\hat{\beta}_{LS}). \tag{4.2}$$

Recall that SSE $(\hat{\beta})$ is minimized by $\hat{\beta}_{LS}$. Fact 1, then, roughly means that $\hat{\beta}$ is fairly close to $\hat{\beta}_{LS}$ if it does not increase the error sum of squares by too much.

Proof of Fact 1. Be definition,

$$\begin{aligned} SSE(\hat{\beta}) &= (Y-X\hat{\beta})'(Y-X\hat{\beta}) \\ &= (Y-X\hat{\beta}_{LS}+X\hat{\beta}_{LS}-X\hat{\beta})' (Y-X\hat{\beta}_{LS}+X\hat{\beta}_{LS}-X\hat{\beta}) \\ &= (Y-X\hat{\beta}_{LS})'(Y-X\hat{\beta}_{LS})+2(Y-X\hat{\beta}_{LS})'(X\hat{\beta}_{LS}-X\hat{\beta})+(X\hat{\beta}_{LS}-X\hat{\beta})'(X\hat{\beta}_{LS}-X\hat{\beta}). \end{aligned}$$

The middle term is zero since

$$(Y-X\hat{\beta}_{LS})'(X\hat{\beta}_{LS}-X\hat{\beta})=(Y-X(X'X)^{-1}X'Y)'X(\hat{\beta}_{LS}-\hat{\beta})$$
$$= (X'Y-X'Y)'(\hat{\beta}_{LS}-\hat{\beta}).$$

Therefore,

$$SSE(\hat{\beta})=(Y-X\hat{\beta}_{LS})'(Y-X\hat{\beta}_{LS})+(\hat{\beta}_{LS}-\hat{\beta})'(X'X)(\hat{\beta}_{LS}-\hat{\beta})$$
$$=SSE(\hat{\beta}_{LS}) + D(\hat{\beta}),$$

and the result is proved.

For any $\hat{\beta}$, let

$$R^2(\hat{\beta}) = 1 - SSE(\hat{\beta})/\Sigma_{i=1}^{n} (Y_i-\bar{Y})^2. \qquad (4.3)$$

For least squares estimates R^2 is simply the usual squared multiple correlation coefficient.

Fact 2. For any $\alpha\varepsilon (0,1)$,

$$\hat{\beta} \varepsilon S_\alpha$$

if and only if

$$\frac{SSE(\hat{\beta})}{SSE(\hat{\beta}_{LS})} \leq 1 + \frac{p+1}{n-p-1} F_{p+1,\, n-p-1;\, 1-\alpha} \qquad (4.4)$$

or

$$R^2(\hat{\beta}) \geq R^2(\hat{\beta}_{LS}) - (1-R^2(\hat{\beta}_{LS})) \frac{p+1}{n-p-1} F_{p+1,\, n-p-1;\, 1-\alpha}. \qquad (4.5)$$

The proof of Fact 2 follows directly from the definition of S_α and Fact 1. Roughly speaking, Fact 2 indicates that the relative increase in the error sum of squares is fundamental. Also, the estimate $\hat{\beta}$ is close to $\hat{\beta}_{LS}$ if the R^2 for $\hat{\beta}$ is sufficiently large.

Fact 2 can easily be used to find the acceptance level for any estimate $\hat{\beta}$. Rearranging (4.5) gives

$$F_{p+1,n-p-1;\ 1-\alpha} = \frac{(n-p-1)\ [R^2(\hat{\beta}_{LS})-R^2(\hat{\beta})]}{(p+1)\ [1 - R^2(\hat{\beta}_{LS})]} \qquad (4.6)$$

Computer packages which report significance values for F statistics have routines which solve (4.5) for α. In addition, many desk and pocket calculators provide either short programs or built-in functions for this purpose.

5. APPLICATION TO SUBSET SELECTION

Some additional notation is required to discuss the subset selection problem. When considering a subset model, it can arbitrarily be assumed that the selected variables are the first k and the eliminated variables are the last $p-k$. Thus, the

design matrix X and the parameter vector β can be conformably partitioned so that the model (2.1) can be rewritten as

$$Y=X_1\Gamma_1 + X_2\Gamma_2 + \epsilon \qquad (5.1)$$

where

$$X=(X_1,X_2)$$

and

$$\beta = \begin{pmatrix}\Gamma_1\\ \Gamma_2\end{pmatrix}$$

Note that the constant term β_0 is included in Γ_1 and hence, X_1 is a $(k+1)Xp$ matrix. An estimate of β, say $\hat{\beta}=(\hat{\Gamma}_1',\hat{\Gamma}_2')'$, is an estimate for the subset problem if its last p-k components are zero, i.e. if $\hat{\Gamma}_2=0$. Let $\hat{\beta}_S$ denote any estimator for the subset problem. The following question naturally arises: which $\hat{\beta}_S$ is closest to $\hat{\beta}_{LS}$ in the sense defined in section 3? The answer, although not surprising, is interesting and useful.

Lemma. The $\hat{\beta}_S$ which minimizes $D(\hat{\beta}_S)$ is

$$\hat{\beta}_S=(\hat{\Gamma}_1',0')'$$

where

$$\hat{\Gamma}_1=(X_1' X_1)^{-1}X_1' Y. \qquad (5.2)$$

In other words, the subset estimate which is closest to the unrestricted least squares estimate is the one which is least squares for restricted problem.

Proof of Lemma. From Fact 1 of the previous section,

$$D(\hat{\beta}_S) = SSE(\hat{\beta}_S) - SSE(\hat{\beta}_{LS}).$$

Therefore, minimizing $D(\hat{\beta}_S)$ is equivalent to minimizing $SSE(\hat{\beta}_S)$. But,

$$SSE(\hat{\beta}_S) = (Y-X\hat{\beta}_S)'(Y-X\hat{\beta}_S)$$

$$=(Y-X_1\hat{\Gamma}_1-X_2\hat{\Gamma}_2)' (Y-X_1\hat{\Gamma}_1 - X_2\hat{\Gamma}_2).$$

The restriction implies $\hat{\Gamma}_2=0$, so

$$SSE(\hat{\beta}_S)=(Y-X_1\hat{\Gamma}_1)'(Y-X_1\hat{\Gamma}_1).$$

This expression is clearly minimized by

$$\hat{\Gamma}_1 = (X_1'X_1)^{-1} X_1' Y$$

and hence (5.2) follows.

As a result of this lemma, subset problems can be studied by considering least squares estimates for the subset problems. In terms of acceptance levels, it is apparent from Fact 2 and the lemma that the most acceptable (the one with the largest acceptance level) subset estimate for any given subset-size is the one with the largest R^2.

There has been some interest in calculating all possible values of R^2 or some subset of high values of these ([12],[13],[18]). The number of possible subsets is 2^p-1 which is very large even when p is moderate (1023 for p=10 and 1,048,575 for p=20.) Clearly, methods for summarizing and interpreting this type of potentially useful information are needed. One possible approach to this problem is through acceptance levels. Subsets can be ordered by R^2 values, irrespective of subset size, and the corresponding acceptance levels given. (This ordering is, of course, equivalent to ordering by acceptance level.) Rules for determining the amount of output can be constructed by bounding the number of subsets to be printed or the acceptance value or by some combination such as the minimum of the numbers given by these criteria. Subsets with large acceptance levels are statistically indistinguishable from the full model. If there are many such subsets then the data is not providing enough direct information for efficiently selecting a subset and additional considerations are required to arrive at a practical solution. If, on the other hand, only a few or perhaps even no subsets have large acceptance levels, then the data is providing a great deal of statistical information which can be used to select a subset. Of course, in this case also, other considerations may be sufficiently important to suggest selecting a subset with a moderate acceptance level. The important point is to present the statistical information as clearly as possible so that these higher level judgements can be made.

For each α, the set of α-acceptable subsets consists of all the subsets for which the corresponding restricted least squares estimate is in S_α. If we assume that the true model, (2.1), is a subset model, i.e. some of the β_j are zero, then the acceptance set approach has an additional interpretation which is presented in the following theorem. Let β_S denote a parameter with some coefficients equal to zero and $\hat{\beta}_S$ the corresponding restricted least squares estimate. Let P_β denote probabilities calculated under the assumption that β is the true value of the parameter.

Theorem 1. For any $\beta=\beta_S$,

$$P_{\beta_S}(\hat{\beta}_S \in S_\alpha) \geq 1-\alpha. \tag{5.3}$$

Proof. For any β,

$$P_\beta\ (\beta \in S_\alpha) = 1-\alpha$$

by the construction of the confidence region S_α. From 3.2 then, it follows that

$$P_\beta(D(\beta) \leq d_\alpha) = 1-\alpha \tag{5.4}$$

Now since $\hat{\beta}_S$ minimizes $D(\beta)$ subject to $\Gamma_2=0$, it follows that

$$D(\beta_S) > D(\hat{\beta}_S). \tag{5.5}$$

In other words, the true parameter β_S must be further away from $\hat{\beta}_{LS}$ than its estimate $\hat{\beta}_S$. Therefore, the estimate is more likely to be in S_α than the parameter. The conclusion (5.3) is implied by (5.4) and (5.5).

Arvesen and McCabe [3], [4] and McCabe and Arvesen [20] have applied the subset selection framework of Gupta and Sobel [14] to the problem of selecting subsets of variables in regression analysis. In the Gupta and Sobel context, a subset refers to a collection of solutions, in this case each solution is a subset of regression variables. The basic idea is to construct a collection of subsets which include the true (best) subset with a given prespecified probability. In Arvesen and McCabe [3], [4] and McCabe and Arvesen[20], a procedure for constructing such collections is given for the case in which the number of variables to be included in each subset is prespecified. This rule includes all subsets in the collection for which

$$\frac{SSE\ (\hat{\beta}_S)}{SSE(\hat{\beta}_S^*)} \leq c^{-1} \qquad (5.6)$$

where $SSE(\hat{\beta}_S)$ is as defined above, $SSE(\hat{\beta}_S^*)$ is the smallest of the $SSE(\hat{\beta}_S)$ values for the given subset size and c is a constant

which depends on n, p, the subset size, the design matrix X, and the bound for the probability of including the true subset in collection. The determination of c is based on asymptotic approximations and requires simulation for each different design matrix.

The theorem above can be used in the present context. Consider the following selection rule: include all subsets in the collection for which

$$\frac{SSE(\hat{\beta}_S)}{SSE(\hat{\beta}_{LS})} \leq 1 + \frac{p+1}{n-p-1} F_{p+1,n-p-1;1-\alpha} \quad (5.7)$$

This is, of course, equivalent to including all subsets for which the corresponding parameter estimates are in S_α.

Let CS denote the event that the true model is included in the collection. Thus,

$$P_{\beta_S}(CS) = P_{\beta_S}(\hat{\beta}_S \in S_\alpha).$$

The following corollary follows immediately from Theorem 1.

Corollary. If $\beta = \beta_S$, i.e. if a subset model is the true model, then

$$P(CS) \geq 1-\alpha. \quad (5.8)$$

It is interesting to contrast the rules given by (5.6) and (5.7). Advantages of the latter include applicability to all subset sizes simulataneously and the ease of computation for the critical values. On the other hand, since (5.6) restricts attention to a given subset size, the number of included subsets of a given size should be less with (5.6) than with (5.7). In other words, the results should be more precise. Since

$$SSE(\hat{\beta}_{LS}) \leq SSE(\hat{\beta}_S^*),$$

the left hand side of (5.6) will be less than or equal to that of (5.7). Thus, a direct comparison of the critical values given on the right-hand sides is not completely informative. In the particular examples which have been examined, one finds the expected result - the right hand side of (5.7) is larger than that of (5.6).

It should be noted that (5.7) (or equivalently (4.5)) is very close to Aitkin's [1] definition of an R^2-adequate(α) subset. He deines a subset to be R^2-adequate(α) if

$$\frac{SSE(\hat{\beta}_S)}{SSE(\hat{\beta}_{LS})} \leq 1 + \frac{p}{n-p-1} F_{p,n-p-1;\ 1-\alpha}. \qquad (5.9)$$

The degrees of freedom in the numerator of the F distribution, and hence the factor multiplying the F is p rather than p+1 as in (5.7). Hence, the collections of subsets obtained by (5.9) are not larger than those obtained using (5.7).

The theorem and corollary of this section provide a means for interpreting the concept of α-acceptability applied to the subset selection problem. It should be noted that an inference of the type contained in the corollary is not the primary purpose for introducing this concept. The inference in the corrollary pertains to which subsets of variables predict well rather than addressing the question of how well they predict with the particular estimate available. This consideration is essentially the basis for the preference for (5.7) over (5.9) in this context. By including the extra degree of freedom for the intercept term, the inference is extended to the entire parameter vector.

6. APPLICATION TO RIDGE REGRESSION

The standard ridge estimates proposed by Hoerl and Kennard [16] [17] are of the form

$$\hat{\beta}(k) = (X'X+kI)^{-1} X'Y \qquad (6.1)$$

where k is nonnegative. The case $k=0$ corresponds to the usual least squares estimate. This type of estimate may be viewed as the solution to the problem of minimizing $D(\beta)$ subject to the constraint that $\beta'\beta \leq B$ where B is some fixed positive number. In the solution, the ridge parameter k is a function of the bound B.

The relationship between k and $D(\hat{\beta}(k))$ is given by following.

Fact 3. $D(\hat{\beta}(k))$ is an increasing function of k.

Proof. From Fact 1,

$$D(\hat{\beta}(k)) = SSE(\hat{\beta}(k)) - SSE(\hat{\beta}_{LS}).$$

Since $SSE(\hat{\beta}_{LS})$ does not depend on k, it is sufficient to show that the derivative of $SSE(\hat{\beta}(k))$ with respect to k (which is the same as that of $SEE(\hat{\beta}(k))$) is positive for $k>0$.

Now,

$$SSE(\hat{\beta}(k)) = Y'[M(k)]^2Y$$

where

$$M(k)= I-X(X'X+kI)^{-1}X'.$$

Therefore,

$$\frac{\partial SSE(\hat{\beta}(k))}{\partial k} =Y' \left[M(k) \frac{\partial M(k)}{\partial k} + \frac{\partial M(k)}{\partial k} M(k) \right] Y.$$

But, $$\frac{\partial M(k)}{\partial k} = X(X'X+kI)^{-1} \left[\frac{\partial (X'X+kI)}{\partial k} \right] (X'X+kI)^{-1}X'$$

$$=X(X'X+kI)^{-2}X'.$$

Combining the above gives

$$\frac{\partial D(\hat{\beta}(k)}{\partial k} = 2kY'X(X'X+kI)^{-3}X'Y \qquad (6.2)$$

To show that the derivative is positive we first note that the matrix $(X'X+kI)$ is positive definite for $k>0$ since both $X'X$ and kI are positive definite. Finally, the derivative in (6.2) can be written as

$$Z'(X'X+kI)Z$$

with

$$Z=\sqrt{2k}\ (X'X+kI)^{-2}\ X'Y$$

Thus the derivative is positive for $k>0$ and the result is established.

Using Fact 3, the ridge estimates in the sets S_α are easily characterized as follows.

Fact 4. For each $\alpha\varepsilon(0,1)$ there is a positive number $k=k(\alpha)$ such that $\hat{\beta}(k)\ \varepsilon\ S_\alpha$ if and only if $k\leq k(\alpha)$

Proof. Fact 4 follows immediately from Fact 3 and the definition of S_α.

Fact 4 may be viewed as a means for setting a reasonable upper bound on the ridge parameter k. Alternatively, one may consider the ridge trace or the numerous other criteria which have been suggested as guides for determining k. If the acceptance level

is high, then the estimate may be used with the knowledge that it is really not very far from the usual estimate in a statistical sense. If the acceptance level is small, however, considerably more faith is required in the ridge technology to justify the chosen value of k.

It is important to note that the ridge procedure is generally performed on a standardized version of the regression problem. Hoerl and Kennard use X'X in correlation form and eliminate the β_0 term. Such modifications have no effect on Facts 3 and 4 since these results depend essentially only on $SSE(\hat{\beta}(k))$ which is invariant under this type of transformation.

Obenchain[25] has defined the associated probability of a ridge estimate as the value of α for which

$$D(\hat{\beta}(k)) = ps^2 F_{p,n-p-1;\ 1-\alpha}.$$

He also considers more general classes of ridge-type estimators. Note that the question of p+1 versus p degrees of freedom arises here, as in the case of subset selection. The same comments apply. Moreover, in many cases it would appear to be just as reasonable to apply ridge techniques to models including a β_0 term as to those without one. Problems with arbitrarily chosen dummy variables appear to be reasonable candidates for this type of approach, although some sort of standardization of the other variables may be desirable.

7. APPLICATION TO REGRESSION ON PRINCIPAL COMPONENTS

Various estimation procedures based on the eigenvector-eigenvalue structure of the regression variable correlations have been proposed. In some cases the structure for only the X variables is used. Webster, Gunst and Mason [28], on the other hand, recommend using this structure for all the variables including the one to be predicted.

Let

$$(X'X) = P\Lambda P', \tag{7.1}$$

Where the rows of P are the eigenvectors of (X'X) and the elements of the diagonal matrix Λ are the corresponding eigenvalues. The decomposition is unique up to permutations and it is customary to impose the following ordering on the elements of Λ: $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$. With this notation, the model (2.1) can be rewritten as

$$Y = W\Gamma + \varepsilon \tag{7.2}$$

where $W=XP$ and $\Gamma=P'\beta$.

Using this parameterization, the form of $D(\beta)$ can be simplified as follows:

$$D(\beta)=(\hat{\beta}_{LS}-\beta)'(X'X)(\hat{\beta}_{LS}-\beta)$$

$$=(\hat{\Gamma}_{LS}-\Gamma)'\Lambda(\hat{\Gamma}_{LS}-\Gamma)$$

where $\hat{\Gamma}_{LS}=P'\hat{\beta}_{LS}$. Since Λ is diagonal,

$$D(\beta)=\sum_{i=1}^{p}(\hat{\Gamma}_{iLS}-\Gamma_i)^2 \lambda_i \tag{7.3}$$

where $\hat{\Gamma}_{iLS}$ and Γ_i denote the i^{th} components of the vectors $\hat{\Gamma}_{LS}$ and Γ_i respectively.

First, consider the principal component estimators of Γ (and hence of $\beta=P\Gamma$) obtained by setting some components equal to $\hat{\Gamma}_{iLS}$ and others equal to zero. Let H denote the set of values of i corresponding to the zero coefficients. Then

$$D(\beta)=\Sigma_H \hat{\Gamma}^2_{iLS} \lambda_i \qquad (7.4)$$

From this expression, it would appear natural to order the components on the basis of $\hat{\Gamma}^2_{iLS} \lambda_i$ rather than λ_i. Suppose now that this ordering has been accomplished, i.e. $\hat{\Gamma}^2_{1LS} \lambda_1 \geq \hat{\Gamma}^2_{2LS} \lambda_2 \geq \ldots \geq \hat{\Gamma}^2_{pLS} \lambda_p$.

A class of principal component estimators indexed by an integer h $(1\leq h\leq p)$ and a number q $(0\leq q<1)$ is described by

$$\hat{\Gamma}_i = \begin{cases} \hat{\Gamma}_{iLS} & \text{for } i=1,\ldots,h-1 \\ q\hat{\Gamma}_{hLS} & \text{for } i=h \\ 0 & \text{for } i=h+1,\ldots,p \end{cases}$$

Let $r=h+q$. Clearly, $0\leq r\leq p$ and given r, the numbers h and q can be determined. The estimate of β corresponding to $\hat{\Gamma}_i$ using h and q will be denoted by $\hat{\beta}(r)$. The distance function is given by

$$D(\hat{\beta}(r))=(1-q)^2\ \hat{\Gamma}^2_{hLS}\ \lambda_h+\Sigma^p_{i=h+1}\ \hat{\Gamma}^2_{iLS}\ \lambda_i. \qquad (7.6)$$

The following is a direct consequence of (7.6).

Fact 5. $D(\hat{\beta}(r))$ is a nondecreasing function of r. $(0\leq r\leq p)$.

Using Fact 5, the set of α-acceptable principal component estimates of the form (7.5) is easily characterized as follows.

Fact 6. For each $\alpha\varepsilon(0,1)$ there is a value of $r=r(\alpha)$ such that $\hat{\beta}(r)\varepsilon S_\alpha$ if and only if $r\leq r(\alpha)$.

Of course other procedures for setting components of Γ to zero are available. A rule which is of the form of (7.5) but with the ordering $\lambda_1\leq\lambda_2\leq\ldots\leq\lambda p$ is discussed by Hocking. Using 7.3, the distance from $\hat{\beta}_{LS}$ and hence the acceptance level are easily calculated for any estimate given in terms of this parameterization.

A procedure which uses components with large λ_i and discards those with small λ_i attempts to eleminate instabilities in the regression coefficients resulting from linear combinations of the X's which have small variances (λ_i). On the other hand it is possible that these particular linear combinations are good predictors of Y and should not be discarded. These ideas are

related to the concepts of predictive and nonpredictive near singularities discussed by Webster, Gunst and Mason [28]. By basing inclusion on the ordered values of $\hat{\tau}^2_{iLS}\lambda_i$, both the variance of the component and its predictive value are taken into account.

8. APPLICATION TO A MINIMAX FAMILY

The advisability of using the usual estimators for multivariate normal estimation problems was seriously questioned by the work of Stein [26]. This paper generated substantial interest in the problem and stimulated much research.

To see how the idea of α-acceptability can be used as a practical guide for applying theoretical results, attention will be focused on the family of minimax estimators of the mean of a normal distribution proposed by Baranchik [5].

Suppose the q-dimensional vector V is normal with mean Θ and covariance matrix $\sigma^2 I$; and s^2 is σ^2 times a χ^2 variable with ν degrees of freedom, independent of V. The loss function for estimating Θ is assumed to be

$$L(\hat{\Theta};\ \Theta,\sigma^2) = (\hat{\Theta}-\Theta)'\ (\hat{\Theta}-\Theta)/\sigma^2. \qquad (8.1)$$

Let $f=V'V/s^2$. Baranchik's result is the following.

Theorem [Baranchik]. The estimator

$$\hat{\Theta} = (1- r(f)/f)\ V \qquad (8.2)$$

is minimax relative to (8.1) if (i)$r(\cdot)$ is monotone nondecreasing and (ii)$0 \leq r(\cdot) \leq 2(q-2)/(\nu+2)$.

First we must transform this result into the regression context. From (2.4) recall that

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2(X'X)^{-1})$$

and from (7.1)

$$(X'X)=P\Lambda P'.$$

Now let $\delta=\Lambda^{\frac{1}{2}}P'\beta$ and similarly for $\hat{\delta}_{LS}$ and $\hat{\delta}$.
It follows that

$$\hat{\delta}_{LS} \sim N(\delta, \sigma^2 I) \qquad (8.3)$$

and we can apply Baranchik's results with $V=\hat{\delta}_{LS}$ and $\Theta=\delta$. The loss function is

$$L(\hat{\delta};\delta,\sigma^2)=(\hat{\delta}-\delta)'(\hat{\delta}-\delta)/\sigma^2$$

$$=(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)/\sigma^2.$$

Therefore, if an estimator $\hat{\delta}$ is minimax for δ under the loss function $L(\hat{\delta};\delta, \sigma^2)$ then the estimator

$$\hat{\beta}=P\Lambda^{-\frac{1}{2}}\ \hat{\delta} \qquad (8.4)$$

is minimax for β under the loss function

$$L(\hat{\beta};\beta,\sigma^2)=(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)/\sigma^2. \tag{8.5}$$

The random variable s^2 is given by (2.7) and the degrees of freedom correspondences are $q=p+1$ and $\nu=n-p-1$. The expression for f becomes

$$f=\hat{\beta}_{LS}'(X'X)\,\hat{\beta}_{LS}/s^2 \tag{8.6}$$

Note that $f=D(0)/s^2$. Now, suppose the function r satisfies Baranchik's conditions. Consider the minimax estimator

$$\hat{\delta}=(1-r(f)/f)\hat{\delta}_{LS}.$$

Since

$$\hat{\delta}_{LS}=\Lambda^{\frac{1}{2}}P'\hat{\beta}_{LS},$$

it follows that

$$\hat{\delta}=(1-r(f)/f)\Lambda^{\frac{1}{2}}P'\hat{\beta}_{LS}.$$

Using the transformation

$$\hat{\beta}=P\Lambda^{-\frac{1}{2}}\hat{\delta}$$

we obtain

$$\hat{\beta}=(1-r(f)/f)\hat{\beta}_{LS}, \tag{8.7}$$

as a minimax estimator under the loss function given by (8.5)

It will now be shown that the criterion of α acceptability can be used as a guide in selecting the function $r(\cdot)$. If $\hat{\beta}$ is of the form given by (8.7) then

$$D(\hat{\beta})=(\hat{\beta}_{LS}-\hat{\beta})'(X'X)(\hat{\beta}_{LS}-\hat{\beta})$$

$$=(r(f)/f)^2\hat{\beta}_{LS}'(X'X)\hat{\beta}_{LS}$$

So,

$$D(\hat{\beta})=s^2r^2(f)/f. \tag{8.8}$$

The estimate $\hat{\beta}$ will be α-acceptable if

$$D(\hat{\beta})\leq(p+1)s^2F_{p+1,n-p-1;1-\alpha}.$$

Equivalently,

$$r(f)\leq(f(p+1)F_{p+1,n-p-1;1-\alpha})^{\frac{1}{2}} \tag{8.9}$$

Since f is unbounded as a function of $\hat{\beta}$, the condition (8.9) is, in one sense, less restrictive than conditions (i) and (ii) of Baranchik's theorem. Thus, if $r(\cdot)$ is a nonnegative monotonic nondecreasing function satisfying

$$r(f)=\min(2(p-1)/(n-p+1),(f(p+1)F_{p+1,n-p-1;1-\alpha})^{\frac{1}{2}}), \tag{8.10}$$

then both the conditions of the theorem and (8.9) will be satisfied. To summarize these results, two additional definitions are needed.

Definition 3. An estimator is α-acceptable if the maximum acceptance level for any realization is greater than or equal to α.

Definition 4. The acceptance level of an estimator is the infimium of the values of α for which it is α-acceptable.

Note that Definitions 1 and 2, given in section 3 were for particular numerical values of β and could be applied to estimates whereas Definitions 3 and 4 apply to estimators which are random variables.

The results above are summarized in the following theorem.

Theorem 2. Let $\alpha(0\leq\alpha\leq 1)$ be fixed. The estimator

$$\hat{\beta}=(1-r(f)/f)\hat{\beta}_{LS}$$

where

$$f=\hat{\beta}_{LS}'(X'X)\hat{\beta}_{LS}/s^2$$

and $r(f)$ is a nonnegative monotonic nondecreasing function satisfying

$$r(f)\leq\min(2(p-1)/(n-p+1),\ (f(p+1)F_{p+1,n-p-1;1-\alpha})^{\frac{1}{2}})$$

has acceptance level α and is minimax relative to the loss function

$$L(\hat{\beta};\beta,\sigma^2)=(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)/\sigma^2.$$

Now let $\hat{\beta}(\alpha)$ be defined by letting the inequality in the theorem be an equality, i.e.

$$r_\alpha(f)=\min(2(p-1)/(n-p+1),(f(p+1)F_{p+1,n-p-1;1-\alpha})^{\frac{1}{2}}),$$

and

$$\hat{\beta}(\alpha)=(1-r_\alpha(f)/f)\hat{\beta}_{LS}. \qquad (8.11)$$

Since $r_\alpha(\cdot)$ clearly satisfies the conditions of the theorem, the following corollary is evident.

Corollary. Let α, $(0\leq\alpha<1)$ be fixed. The estimator $\hat{\beta}(\alpha)$ given by (8.11) has acceptance level α and is minimax with respect to (8.5).

9. PREDICTED VALUES

Let x denote a (p+1) dimensional vector with first component one. Each row of the design matrix X is the transpose of such a vector. To construct an estimate of the expected value of Y for a configuration of predictor values corresponding to x, we

simply take the appropriate linear combination of the estimated regression coefficients. These estimated expected values are called predicted values. Thus, for the estimate $\hat{\beta}$ and the vector x_0, the predicted value is

$$\hat{Y}_0 = x_0' \hat{\beta}. \tag{9.1}$$

The vector of predicted values for the rows of X is

$$\hat{Y} = X\hat{\beta}. \tag{9.2}$$

Components of $\hat{Y}$ are denoted by $\hat{Y}_i$ for i=1,...n. When the least squares estimate $\hat{\beta}_{LS}$ is used, the corresponding predicted values are denoted by $\hat{Y}_{OLS}$ and $\hat{Y}_{LS}$.

Since regression equations are often used for predicition it is important to consider the effects of different choices of $\hat{\beta}$ upon (9.1) and (9.2). First, the effect of $\hat{\beta}$ upon the predicted values in the original data is considered. Second, a type of simulataneous view is developed for treating all x_0.

From the definition of $D(\hat{\beta})$.

$$\begin{aligned} D(\hat{\beta}) &= (\hat{\beta}_{LS} - \hat{\beta})'(X'X)(\hat{\beta}_{LS} - \hat{\beta}) \\ &= (X\hat{\beta}_{LS} - X\hat{\beta})'(X\hat{\beta}_{LS} - X\hat{\beta}) \\ &= (\hat{Y}_{LS} - \hat{Y})'(\hat{Y}_{LS} - \hat{Y}). \\ &= \Sigma_{i=1}^{n} (\hat{Y}_{iLS} - \hat{Y}_i)^2 \end{aligned}$$

Recall that

$$\hat{\beta} \varepsilon S_\alpha$$

iff

$$D(\hat{\beta}) \leq d_\alpha$$

where

$$d_\alpha = (p+1)s^2 F_{p+1,n-p-1;1-\alpha}$$

Therefore, $\hat{\beta} \varepsilon S_\alpha$ if and only if the sum of squares of the differences between the least squares and the alternative predicted values is sufficiently small, i.e. if

$$\Sigma_{i=1}^{n} (\hat{Y}_{iLS} - \hat{Y}_i)^2 \leq d_\alpha \qquad (9.3)$$

In terms of the root mean squared deviation, (9.3) becomes

$$\sqrt{\frac{1}{n}\Sigma_{i=1}^{n} (\hat{Y}_{iLS} - \hat{Y}_i)^2} \leq s h_\alpha / n^{\frac{1}{2}} \qquad (9.4)$$

where

$$h_\alpha = ((p+1)F_{p+1,n-p-1;1-\alpha})^{\frac{1}{2}}.$$

On one hand, an expression such as (9.4) may be viewed as providing insight and aiding the interpretation of α-acceptability. If α is preselected, the right hand side of (9.4) provides a measure of how much deviation from $\hat{Y}_{LS}$ is being allowed by

limiting the search for $\hat{\beta}$ to S_α. The inequality (9.4) can be relativized by dividing through by s. In this form, the right hand side can be calculated before any data is examined.

On the other hand, an expression such as (9.4) can be used as a guide in choosing a reasonable for α. If one decides upon a reasonable bound for the relative root mean square deviation, then a value of α which will satisfy this goal is easily found. Attention can then be focused on S_α for alternatives to $\hat{\beta}_{LS}$.

Suppose now that α has been fixed. What can be said about the predicted values $x_0' \hat{\beta}$ for a given value of x_0 when $\hat{\beta}$ is restricted to be in S_α. The following fact provides an answer. Note that the former restriction that the first component of x_0 be equal to one is irrelevant in the remainder of this section.

Fact 7. If $\hat{\beta} \epsilon S_\alpha$ then

$$x_0'\hat{\beta}_{LS} - sh_\alpha (x_0'(X'X)^{-1}x_0)^{\frac{1}{2}} \leq x_0'\hat{\beta} \leq x_0'\hat{\beta}_{LS} + sh_\alpha (x_0'(X'X)^{-1}x_0)^{\frac{1}{2}} \quad (9.5)$$

for any x_0.

Proof. The extrema of $x_0' \hat{\beta}$ subject to the constraint

$$D(\hat{\beta}) = d_\alpha \quad (9.6)$$

are needed. This problem is easily solved using Lagrange multipliers.

Let

$$f(\hat{\beta})=x_0'\hat{\beta}+\lambda(D(\hat{\beta})-d_\alpha),$$

then

$$\partial f(\hat{\beta})/\partial\hat{\beta}=x_0-2\lambda(X'X)(\hat{\beta}_{LS}-\hat{\beta}).$$

Setting this expression equal to zero gives

$$(\hat{\beta}_{LS}-\hat{\beta})=(X'X)^{-1}x_0/2\lambda,$$

which when substituted into the definition of $D(\hat{\beta})$ and combined with (9.6) yields

$$\lambda=\pm(x_0'(X'X)^{-1}x_0/4d_\alpha)^{\frac{1}{2}}$$

Thus, the values of $\hat{\beta}$ giving extrema are

$$\hat{\beta}=\hat{\beta}_{LS}\pm(X'X)^{-1}x_0(d_\alpha/(x_0'(X'X)^{-1}x_0))^{\frac{1}{2}}$$

and the extreme values of $x_0'\hat{\beta}$ are

$$x_0'\hat{\beta}_{LS}\pm sh_\alpha(x_0'(X'X)^{-1}x_0),^{\frac{1}{2}}$$

since

$$s^2h_\alpha^2=d_\alpha.$$

The result is proved by noting that h_α is a monotonic decreasing function of α.

In the sense that S_α is viewed as a confidence region for β, regions of the type given in (9.5) can be viewed as simultaneous confidence intervals for all values of x_0. This idea is made precise in the following fact.

Fact 8. Let α and β be fixed. Then

$$P_\beta(x_0'\beta \epsilon H(x_0) \text{ for all } x_0)=1-\alpha \tag{9.7}$$

where

$$H(x_0)=[x_0'\hat{\beta}_{LS}-sh_\alpha(x_0'(X'X)^{-1}x_0)^{\frac{1}{2}}, x_0'\hat{\beta}_{LS}+sh_\alpha(x_0'(X'X)^{-1}x_0)] \tag{9.8}$$

Proof. Since

$$P_\beta(\beta \epsilon S_\alpha)=1-\alpha$$

it suffices to show that the events

$$\beta \epsilon S_\alpha \tag{9.9}$$

and

$$x_0'\beta \epsilon H(x_0) \text{ for all } x_0 \tag{9.10}$$

are equivalent. Since Fact 7 states that (9.9) implies (9.10) it remains only to show that (9.10) implies (9.9).

Suppose (9.10) holds. Then it holds for

$$x_0=(X'X)(\hat{\beta}_{LS}-\beta)$$

which gives

$$x_0'(\hat{\beta}_{LS}-\beta)= D(\beta)$$

and also,

$$x_0'(X'X)^{-1}\, x_0=D(\beta)$$

Therefore, from (9.10) and (9.8), it follows that

$$-sh_\alpha(D(\beta))^{\frac{1}{2}} \leq D(\beta) \leq sh_\alpha(D(\beta))^{\frac{1}{2}}$$

Hence,

$$D(\beta) \leq s^2h_\alpha^2$$

$$=d_\alpha,$$

which is equivalent to (9.9).

The intervals $H(x_0)$ are, of course, substantially larger than the conventional intergals which use the critical value $t_{n-p-1;1-\alpha/2}$ in place of h_α. With the larger intervals, however, the confidence statement is much stronger since it pertains simultaneously to all x_0.

10. EXAMPLE

To illustrate the ideas presented in the preceeding sections, the air pollution data from McDonald and Schwing [23] is examined. This data is also discussed in Hocking's [15] paper. Sixty

observations (n=60) were taken on total mortality and fifteen potential predictors (p=15). It is important to keep in mind that although 60 seems like a large number of observations, the regression coefficient estimate depends on estimated values for 16 means, 16 variances and 120 covariances. Thus, a total of 152 parameter estimates are required.

The R^2 for the unrestricted least squares estimate is 0.764. There are a very large number of subset estimates which are very close to $\hat{\beta}_{LS}$. For example, the 50th best subset of size 10 has an R^2 of .749 with an acceptance level of 0.9997. In other words, this estimate is in the 0.03% confidence region around $\hat{\beta}_{LS}$; rejection of this value as a hypothesized value of β would require a Type I error rate of 0.9997. Similarly, the 50th best subsets of sizes 6,7,8, and 9 are 80%, 94%, 98% and 99% acceptable respectively.

These results should not be viewed as negative or discouraging. With 152 parameters to estimate from 60 observations, these results are not even particularly surprising. The point is that there are a large number of subset models which fit the data almost as well as the unrestricted least squares fit. From a statistical point of view these alternatives are essentially indistinguishable from $\hat{\beta}_{LS}$.

If selection of a subset model is desirable, additional criteria need to be imposed upon the problem. In some cases, it may be desirable to minimize the number of variables included. For this problem, one might be willing to consider subsets of size six. Even with this restriction, however, there are 14 subsets which are 90% acceptable with R^2's ranging from .717 to .735. Another alternative would be to impose some cost structure on the predictor variables as has been discussed by Lindley[19] and McCabe and Ross [21].

McDonald and Schwing use several methods which lead to consideration of models with variables (1,2,3,6,9,14) and (1,2,6,8,9,14). These fits have R^2's of .735 and .724 and are 98% and 94% acceptable, respectively. From a classical viewpoint, these estimates are very acceptable.

McDonald and Schwing and Hocking have also investigated this data using ridge methods. The first authors suggest a value of k=.2 for the full model and the subset models (1,2,3,6,9,14) and (1,2,6,8,9,14). The R^2 values are .724, .711 and .708. These fits are 94%, 84% and 81% acceptable. (Note that the definition of R^2 used here is given by (4.3) which agrees with that used by Hocking. McDonald and Schwing use a different definition and so the values given here differ from those given in their paper. As given by (4.3), the quantity R^2 will be negative for a $\hat{\beta}$ which does worse than $\bar{Y}$). Since the acceptance levels for k =.2 are only moderately large, some confidence in the ridge procedure which led

to this choice is necessary if one is to accept these deviations from $\hat{\beta}_{LS}$.

On the other hand, Hocking gives a ridge estimate for the subset (1,2,3,4,5,6,8,9,12,13,14) with k=.06. For this estimate R^2=.734 and it is 98% acceptable. Since this estimate is so close to $\hat{\beta}_{LS}$, one might consider using it even if one were very skeptical about the validity of ridge procedures. The possibility of obtaining the purported benefits of the ridge procedure can be purchased for a very small price - a relatively small deviation from $\hat{\beta}_{LS}$.

Two principal component-type estimates are also given by Hocking. For one component, R^2=.742; for two R^2=.743. These estimates are both 99% acceptable.

The Baranchik-type minimax estimator given by (8.11) is .9957 $\hat{\beta}_{LS}$. This is, of course, 99% acceptable also. It should be noted that if calculations are done with the correlation matrix, then $\bar{Y}$ should also be multiplied by the shrinkage factor to obtain the correct transformed estimate.

Inspection of the above estimates has led the author to consider the following as candidates for reasonable regression equation estimates:

$$\hat{Y}=.25(X_1-X_2-X_3-X_4-X_5-X_6) + .6X_9-X_{12}+X_{13}-.1X_{14}$$

and

$$\hat{Y}=.25(X_1-X_2+X_{14})-.1(X_3+X_4+X_5+X_6-X_8+X_{12}-X_{13})+.6X_9.$$

For these equations $R^2=.742$ and .725. They are 99% and 95% acceptable; respectively. The equations given are in standardized form and hence the estimates are directly comparable to the estimates given in the McDonald and Schwing and Hocking papers. In many problems, useful linear combinations of variables which may be fitted with a single coefficient can be constructed either before or after studying the data. Mosteller and Tukey [24] have discussed this idea and use the term "judgement composites" to describe these combinations.

Finally, the quantity $h_\alpha/n^{\frac{1}{2}}$ used in (9.4) is 0.30 for $\alpha=.99$ and 0.35 for $\alpha=.95$. Thus, the average (in the root mean square sense) change in predicted value obtained by using any 99%-acceptable estimate is not more than about .3 standard deviations. For 95%-acceptable estimates the corresponding value is .35 standard deviations.

11. IMPLEMENTATION

The idea of α-acceptability has been presented as a tool for evaluating various alternatives to least squared regression coefficient estimates. Several special topics relating to the construction and evaluation of these alternatives are discussed in this section.

In the sections on subset selection and ridge regression the question regarding p vs p+1 arose. This question actually arises when considering most alternatives to least squares and essentially, is equivalent to the choice of working with a correlation matrix or with the full X'X matrix. When working with a correlation matrix, there is usually an implicit assumption about the estimate of the intercept term. Let the standardized form of the prediction equation be given by

$$\frac{\hat{Y}_j - \bar{Y}}{s_Y} = \Sigma_{i=1}^{p} \hat{\alpha}_i \left(\frac{X_{ij} - \bar{X}_i}{s_i} \right)$$

Then, in terms of the unstandardized form,

$$\hat{\beta}_i = \hat{\alpha}_i \, s_Y / s_i, \quad i = 1, \ldots, p.$$

For the two forms to match, the intercept is estimated by

$$\hat{\beta}_0 = \bar{Y} - \Sigma_{i=1}^{p} \hat{\alpha}_i \, \bar{X}_i / s_i. \tag{11.1}$$

Most of the least squares alternatives focus upon $(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)$ and neglect $\bar{Y}$. Thus, the dimensionality is reduced from p+1 to p. Although this reduction may be appropriate under certain circumstances, there does not seem to be any a priori justification for always neglecting $\bar{Y}$. The minimax procedure in section 8 corresponds to multiplying the α's and $\bar{Y}$ by a specified quantity.

The estimation constraint given by (11.1) may be viewed in another way. With this constraint, all estimates will give the same predicted value ($\overline{Y}$) at ($\overline{X}_1,\overline{X}_2,\ldots,\overline{X}_p$). Again, this may or may not be a reasonable condition to impose. Alternatively, one could impose the condition that all estimates give the same predicted value at some other (hopefully meaningful) point.

The choice of an appropriate α is a complex problem. For large α, such as α=.99, the entire process of selecting an alternative may be viewed as a special kind of rounding off. Consideration of the case p=0 can provide some insight. For α=.992, the set of α-acceptable estimates is the 0.8% confidence interval $\overline{Y} + s_{\overline{Y}}/100$. Relative to the statistical variation represented by $s_{\overline{Y}}$, all values in this interval are somewhat equivalent.

Lower values of α give estimates which may be further from $\hat{\beta}_{LS}$ and therefore, require some external justification. Small values of α give estimates which border on being incompatible with the data.

Estimation of regression coefficients by integers or simple fractions and grouping terms together, as suggested in the example, deserves further study. In most cases, the use of some un-standardized form (which was not done in the example) is most

likely to be profitable. Ideally, such estimates should be constructed by an individual who is familiar with the variables and can give some interpretation to the results. An interactive computing environment would be ideal for this purpose.

The determination of a useful set of estimated regression coefficients for problems with correlated predictors and a moderate amount of data requires a substantial amount of statistical manipulation coupled with sound judgement. In many situations, a combination of methods, such as Hocking's [15] ridge-select procedure are appropriate. By restricting attention to estimates in an α-acceptable set, reasonable constraints on the search for useful estimates are imposed.

REFERENCES

[1] Aitkin, M.A. (1974). Simultaneous inference and the choice variable subsets in multiple regression, Technometrics, 16, 221-227.

[2] Alam, K. (1973). A family of minimax estimators of the mean of a multivariate normal distribution. Ann. Statist., 1, 517-525.

[3] Arvesen, J.N. and McCabe, G.P.,Jr. (1973). Variable selection in regression analysis. Proc. Univ. of Kentucky Conference on Regression with a Large Number of Predictor Variables. Thompson, W.O. and Cady, F.B. (eds.) Dept. of Statist., Univ. of Kentucky, Lexington, Kentucky, 136-148.

[4] Arvesen, J.N. and McCabe, G.P., Jr. (1975). Subset selection problems for variances with applications to regression analysis. J. Amer. Statist. Assoc., 70, 166-170.

[5] Baranchik, A.J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. Ann. Math. Statist., 41, 642-45.

[6] Beaton, A.E., Rubin, D.B. and Barone, J.L. (1976). The acceptability of regression solutions: another look at computational accuracy. J. Amer. Statist. Assoc. 71, 158-68.

[7] Berger, J.O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. Ann. Statist., 4, 223-26.

[8] Bock, M.E. (1975). Minimax estimators of the mean of a multivariate normal distribution. Ann. Statist., 3, 209-18.

[9] Cook, R.D. (1977). Detection of influential obserbation in linear regression. Technometrics, 19, 15-18.

[10] Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. J. Amer. Statist. Assoc., 72, 77-91.

[11] Effron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. J. Amer. Statist. Assoc., 68, 117-30.

[12] Furnival, G.M. (1971). All possible regressions with less computation. Technometrics, 13, 403-8.

[13] Furnival, G.M. and Wilson, R.W., Jr. (1974). Regression by leaps and bounds. Technometrics, 16, 499-512.

[14] Gupta, S.S. and Sobel, M. (1962). On selecting a subset containing the population with the smalles variance. Biometrika, 49, 495-507.

[15] Hocking, R.R. (1976). The analysis and selection of variables in linear regression, Biometrics, 32, 1-49.

[16] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression; biased estimation for non-orthogonal problems. Technometrics, 12, 55-67.

[17] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: applications to non-orthogonal problems. Technometrics, 12, 69-82.

[18] LaMotte, L.R. and Hocking, R.R. (1970). Computational efficiency in the selection of regression variables. Technometrics, 12, 83-94.

[19] Lindley, D.V. (1968). The choice of variables in multiple regression. J. Roy. Statist. Soc. B, 30, 31-53.

[20] McCabe, G.P., Jr. and Arvesen, J.N. (1974). A subset selection procedure for regression variables. J. Statist. Comput. Simul., 3, 137-146.

[21] McCabe, G.P., Jr. and Ross, M.A. (1975). A stepwise algorithm for selecting regression variables using cost criteria. Proc. of Computer Science and Statistics: 8th Annual Symposium on the Interface, 228-232.

[22] McDonald, G.C. (1975). Discussion of ridge analysis following a preliminary test of the shruken hypothesis. Technometrics, 17, 443-445.

[23] McDonald, G.C. and Schwing, R.C. (1973). Instabilities of regression estimates relating air pollution to mortality. Technometrics, 15, 463-481.

[24] Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression. Reading, Massachusetts: Addison-Wesley.

[25] Obenchain, R.L. (1977). Classical F-tests and confidence regions for ridge regression. Unpublished manuscript.

[26] Stein, C. (1955). Inadmissability of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Prob. 1, 197-206.

[27] Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. Ann. Math. Statis., 42, 358-88.

[28] Webster, J.T., Gunst, R.F. and Mason, R.L. (1974). Latent root regression analysis. Technometrics, 16, 513-22.