TECHNOMETRICS© VOL. 15, No. 4 NOVEMBER 1973

Some Comments on C_P

C. L. MALLOWS

Bell Laboratories
Murray Hill, New Jersey

We discuss the interpretation of C_P-plots and show how they can be calibrated in several ways. We comment on the practice of using the display as a basis for formal selection of a subset-regression model, and extend the range of application of the device to encompass arbitrary linear estimates of the regression coefficients, for example Ridge estimates.

KEY WORDS

Linear Regression
Selection of Variables
Ridge Regression

1. INTRODUCTION

Suppose that we have data consisting of n observations on each of $k + 1$ variables, namely k independent variables $x_1, \cdots, x_k$ and one dependent variable, y. Write $x_0 = 1$, $\mathbf{x}(1 \times (k+1)) = (x_0, x_1, \cdots, x_k)$, $y(n \times 1) = (y_1, \cdots, y_n)^T$, $\mathbf{X}(n \times (k+1)) = (x_{ui})$. A model of the form

$$y_u = \eta(\mathbf{x}_u) + e_u \quad u = 1, 2, \cdots n \tag{1}$$

where

$$\eta(\mathbf{x}_u) = \beta_0 + \Sigma\beta_i x_{ui} = \mathbf{x}_u\boldsymbol{\beta}$$

is to be entertained,† with the residuals $e_1 \cdots e_n$ being regarded (tentatively) as being independent random variables with mean zero and unknown common variance σ^2. The x's are not to be regarded as being sampled randomly from some population, but rather are to be taken as fixed design variables. We suppose that the statistician is interested in choosing an estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \cdots, \hat{\beta}_k)$, with the idea that for any point $\mathbf{x}$ in the general vicinity of the data at hand, the value

$$\hat{y}(\mathbf{x}) = \hat{\beta}_0 + \Sigma\hat{\beta}_i x_i$$

will be a good estimate of $\eta(\mathbf{x})$. In particular he may be interested in choosing a "subset least-squares" estimate in which some components of $\hat{\boldsymbol{\beta}}$ are set at zero and the remainder estimated by least squares.

The C_P-plot is a graphical display device that helps the analyst to examine his data with this framework in mind. Consider a subset P of the set of indices $K^+ = \{0, 1, 2, \cdots, k\}$; let Q be the complementary subset. Suppose the number of elements in P, Q are $|P| = p$, $|Q| = q$, so that $p + q = k + 1$. Denote by $\boldsymbol{\beta}_P$ the vector of estimates that is obtained when the coefficients with subscripts in P

† If β_0 is absent the development is entirely similar. We assume throughout that $\mathbf{X}$ has rank $k + 1$.

Received June 1972; revised Oct. 1972

are estimated by least squares, the remaining coefficients being set equal to zero; i.e.

$$\hat{\beta}_P = \mathbf{X}_P^- y$$

where $\mathbf{X}_P^-$ is the (Moore–Penrose) generalized inverse of $\mathbf{X}_P$, which in turn is obtained from $\mathbf{X}$ replacing the columns having subscripts in Q by columns of zeroes. (Thus $\mathbf{X}_P^-$ has zeroes in the rows corresponding to Q, and the remaining rows contain the matrix $(\mathbf{Z}_P^T\mathbf{Z}_P)^{-1}\mathbf{Z}_P^T$ where $\mathbf{Z}_P$ is obtained from $\mathbf{X}$ by deleting the columns corresponding to Q). Let RSS_P denote the corresponding residual sum of squares, i.e.

$$\mathrm{RSS}_P = \sum (y_u - \mathbf{x}_u\hat{\beta}_P)^2$$

For any such estimate $\hat{\beta}_P$, a measure of adequacy for prediction is the "scaled sum of squared errors"

$$J_P = \frac{1}{\sigma^2}\sum_{u=1}^{n} (\mathbf{x}_u\hat{\beta}_P - \mathbf{x}_u\beta)^2 = \frac{1}{\sigma^2}(\hat{\beta}_P - \beta)^T\mathbf{X}^T\mathbf{X}(\hat{\beta}_P - \beta),$$

the expectation of which is easily found to be

$$E(J_P) = V_P + \frac{1}{\sigma^2}B_P$$

where V_P, B_P are respectively "variance" and "bias" contributions given by

$$\begin{aligned} V_P &= |P| = p \\ B_P &= \beta_Q^T\mathbf{X}^T(I - \mathbf{M}_P)\mathbf{X}\beta_Q \end{aligned} \tag{2}$$

and β_Q is β with the elements corresponding to P replaced by zeroes, and $\mathbf{M}_P = \mathbf{X}\mathbf{X}_P^- = \mathbf{X}_P\mathbf{X}_P^- = \mathbf{Z}_P(\mathbf{Z}_P^T\mathbf{Z}_P)^{-1}\mathbf{Z}_P^T$.

The C_P statistic is defined to be

$$C_P = \frac{1}{\hat{\sigma}^2}\mathrm{RSS}_P - n + 2p \tag{3}$$

where $\hat{\sigma}^2$ is an estimate of σ^2. Clearly (as has been remarked by Kennard (1971)), C_P is a simple function of RSS_P, as are the multiple correlation coefficient defined by $1 - R_P^2 = \mathrm{RSS}_P/\mathrm{TSS}$ (where TSS is the total sum of squares) and the "adjusted" version of this. However the form (3) has the advantage (as has been shown by Gorman and Toman (1966), Daniel and Wood (1971), and Godfrey (1972)) that since under the above assumptions

$$E(\mathrm{RSS}_P) = (n - p)\sigma^2 + B_P, \tag{4}$$

C_P is an estimate of $E(J_P)$, and is suitably standardized for graphical display, plotted against p. Graphical presentation of the various regression sums of squares themselves against p was advocated by Watts (1965). For k not too large it is feasible ([4], [5], [19]) to compute and display all the 2^{k+1} values of C_P; for larger values one can use algorithms of [2], [8], [14] to compute only the more interesting (smaller) values.

In section 2 we describe some of the configurations that can arise; in section 3 we provide some formal calibration for the display and in section 4 comment on the practice of using it as a basis for formal selection. The approach is extended in section 5 to handle arbitrary linear estimates of the regression coefficients.

The approach can also be extended to handle multivariate response data and to deal with an arbitrary weight function $w(\mathbf{x})$ in factor-space, describing a region

of interest different from that indicated by the configuration of the data currently to hand. In each case, the derivation is exactly parallel to that given above. In the former case, one obtains a matrix analog of C_P in the form $\hat{\Sigma}^{-1} \cdot \mathrm{RSS}_P - (n - 2p)I$ where $\hat{\Sigma}$ is an estimate of the residual covariance matrix, and RSS_P is $\Sigma(\mathbf{y}_u - \mathbf{x}_u\hat{\beta}_P)^T(\mathbf{y}_u - \mathbf{x}_u\hat{\beta}_P)$. One or more measures of the "size" of C_P (such as the trace, or largest eigenvalue) can be plotted against p. In the latter case, with the matrix $\mathbf{\Delta} = (\Delta_{ij})$ defined by $\Delta_{ij} = \int x_i x_j w(\mathbf{x})d\mathbf{x}$, one arrives at a statistic of the form

$$C_P^{\Delta} = \frac{1}{\hat{\sigma}^2}(\hat{\beta}_P - \hat{\beta})^T\mathbf{\Delta}(\hat{\beta}_P - \hat{\beta}) - V_{K^+}^{\Delta} + 2V_P^{\Delta} .$$

where $V_{K^+}^{\Delta} = \text{trace } (\mathbf{\Delta}(\mathbf{X}^T\mathbf{X})^{-1})$, $V_P^{\Delta} = \text{trace } (\mathbf{\Delta}(\mathbf{X}_P^T\mathbf{X}_P)^{-})$, and we can plot C_P^{Δ} against V_P^{Δ}. This reduces to the C_P-plot when $\mathbf{\Delta} = \mathbf{X}^T\mathbf{X}$. If interest is concentrated at a single point x, we have $\mathbf{\Delta} = \mathbf{x}\mathbf{x}^T$, and the statistic $\hat{\sigma}^2 C_P^{\Delta}$ is equivalent to that suggested by Allen (1971); his equation (9) $= \hat{\sigma}^2(C_P^{\Delta} - \mathbf{x}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^T)$.

2. Some Configurations on C_P-Plots

From (2), (3), (4) we see that if $\beta_Q = 0$, so that the P-subset model is in fact completely appropriate, then $\mathrm{RSS}_P \approx (n - p)\sigma^2$ and $C_P \approx p$. If $\hat{\sigma}^2$ is taken as $\mathrm{RSS}_{K^+/(n-k-1)}$, then $C_{K^+} = |K^+| = k + 1$ exactly. Notice that if P^* is a $(p + 1)$ element subset which contains P, then

$$C_{P^*} - C_P = 2 - \frac{SS}{\hat{\sigma}^2} \tag{5}$$

where SS is the one-d.f. contribution to the regression sum of squares due to the $(p + 1)$-th variable, so that $\mathrm{SS}/\hat{\sigma}^2$ is a t_1^2 statistic that could be used in a stepwise testing algorithm. If the additional variable is unimportant, i.e. if the bias contribution $B_P - B_{P^*}$ is small, then $E(\mathrm{SS}) \approx \hat{\sigma}^2$ and so

$$E(C_{P^*} - C_P) \approx 1.$$

Mantel (1970) has discussed the use of stepwise procedures, and how they behave in the face of various patterns of correlation amongst the independent variables. It is illuminating to consider how patterns similar to those he describes would show up on a C_P-plot.

First, suppose the independent variables are not highly correlated, that $\beta = \beta_P$, and that every non-zero element of β is large (relative to the standard error of its least-squares estimate). Then the C_P-plot will look something like Figure 1 (drawn for the case $p = k - 2$, $K^+ - P = \{1, 2, 3\}$). Notice the approximately linear diagonal configuration of points corresponding to the well-fitting subsets of variables.

Now, suppose x_1, x_2, x_3 are highly correlated with each other, with each being about equally correlated with y. Then any two of these variables, but not all three, can be deleted from the model without much effect. In this case the relevant points on the C_P-plot will look something like Figure 2a, if no other variables are of importance, or like Figure 2b if some other subset P is also needed. (In all these examples we are assuming that the constant term β_0 is always needed). Notice that now the diagonal pattern is incomplete. In an intermediate case, when x_1, x_2, x_3 have moderate correlations, a picture intermediate between Figures 1 and 2b will be obtained.

Thirdly, suppose x_1 , x_2 are individually unimportant but jointly are quite effective in reducing the residual sum of squares; suppose some further subset P of variables is also needed. Mantel gives an explicit example of this behavior. Figure 3 shows the resulting configuration in the case $|P| = k - 4$.

Notice that even if $C_{[P,1,2]}$ is the smallest C_P-value for subsets of size $p + 2$, there might be subsets P_1' , P_2' , (not containing P) with $|P_i'| = p$ or $p + 1$ that gave smaller values of C_P than those for P, $\{P, 1\}$, $\{P, 2\}$. In this case an upward stepwise testing algorithm might be led to include variables in these subsets and so not get to the subset $\{P, 1, 2\}$. Mantel describes a situation where this would happen.

3. Calibration

To derive bench marks for more formal interpretation of C_P-plots, we assume that the model (1) is in fact exactly appropriate, with the residuals $e_1 \cdots e_n$ being independent and Normal $(0, \sigma^2)$. Suppose $\hat{\sigma}^2$ is estimated by RSS_{K^+}/ν where $\nu = n - k - 1$, the residual degrees of freedom. We do not of course recommend that the following distributional results be used blindly without careful inspection of the empirical residuals $y_i - \hat{y}(x_i)$, $i = 1, \cdots, n$. However, they should give pause to workers who are tempted to assign significance to quantities of the magnitude of a few units or even fractions of a unit on the C_P scale.

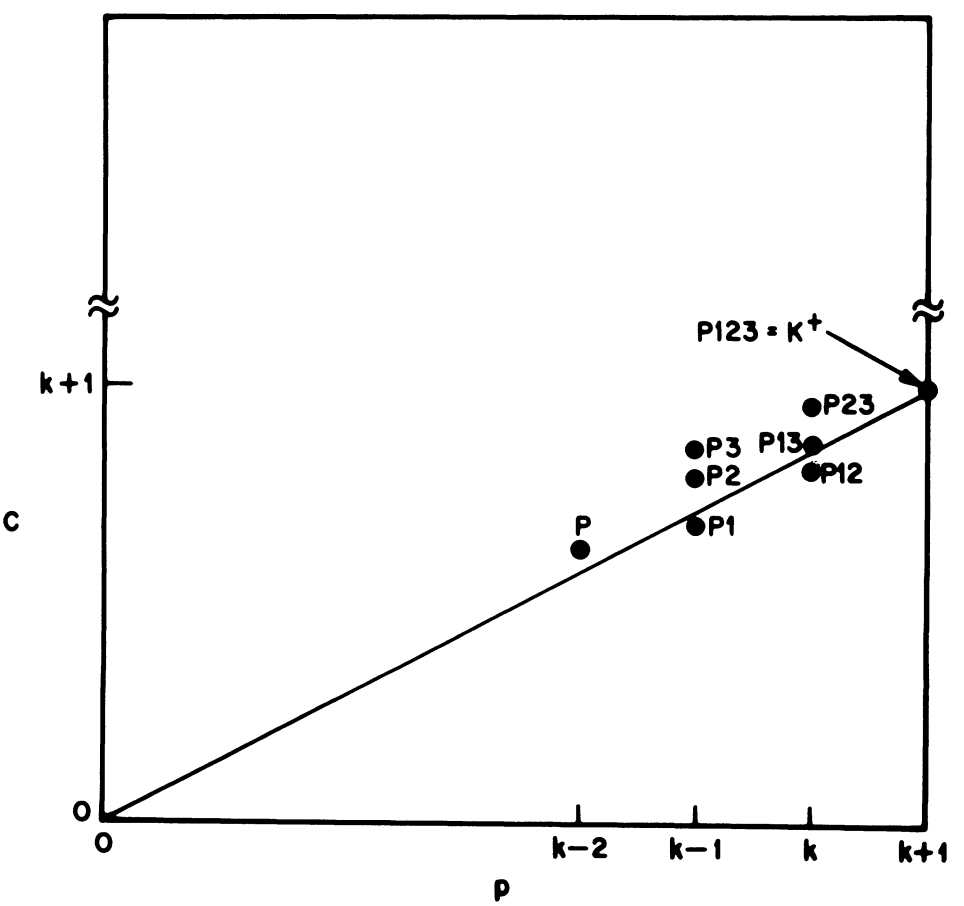

Figure 1—C_P-plot: P is an adequate subset.

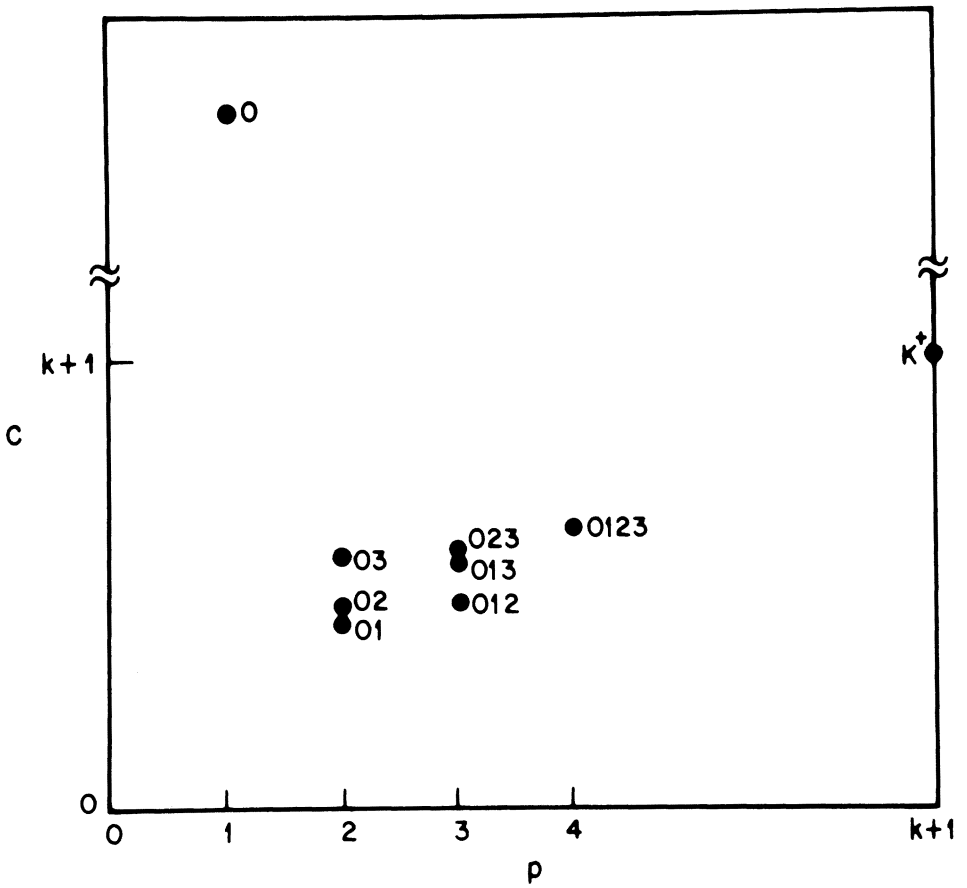

FIGURE 2a—C_P-plot: Variables 1, 2, 3, are highly explanatory also highly correlated.

First, notice that the increment $C_{P*} - C_P$ (in (5) above) is distributed as $2 - t_1^2$, where the t-statistic t_1 is central if $\beta = \beta_{P*}$. In this case this increment has mean and variance of approximately 1 and 2 respectively. Similarly,

$$C_{K^+} - C_P = k + 1 - C_P = q(2 - F_{q,\nu}) \tag{6}$$

where $q = k + 1 - p$ and the F statistic is central if $\beta = \beta_P$; thus if ν is large compared with q this increment has mean and variance approximately q and $2q$ respectively. The variance of the slope of the line joining the points (p, C_P), $(k + 1, k + 1)$ is thus $2/q$, so that the slope of a diagonal configuration such as is shown in Figure 1 will vary considerably about 45°. The following tables (derived from (6)) give values of $C_P - p$ that will be exceeded with probability α when the subset P is in fact adequate (i.e. when $\beta = \beta_P$ so that $\beta_Q = 0$), for the cases $\nu = n - k - 1 = 30, \infty$. The value tabulated is $q(F_{q,\nu}(\alpha) - 1)$.

For comparing two C_P-values corresponding to subsets P, P' with $P \cap P' = B$, $P = A \cup B$, $P' = A' \cup B$, it is straightforward to derive the results, valid under the null hypothesis that each of P and P' is an adequate subset,

$$E(C_P - C_{P'}) \approx |P| - |P'| = |A| - |A'|$$

$$\text{Var } (C_P - C_{P'}) \approx 2(|A| + |A'| - 2R^2)$$

where R^2 is the sum of squares of the canonical correlations between the sets of variables X_A and $X_{A'}$, after partialling out the variables X_B. (Thus if $|B| =$

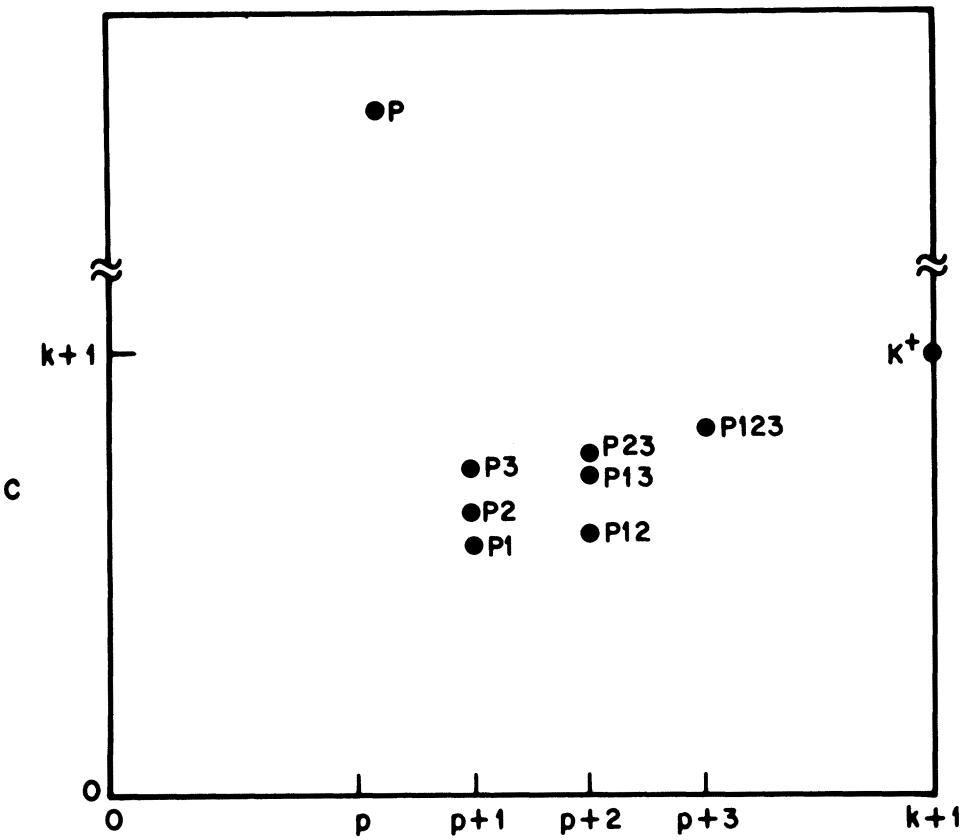

FIGURE 2b—C_P-plot: Same as 2a except that variables in P are also explanatory.

$|P| - 1$, Var $(C_P - C_{P'}) = 4(1 - \rho^2)$ where ρ is the partial correlation coefficient $\rho_{AA'\cdot B}$).†

We now use the Scheffé confidence ellipsoid to derive a different kind of result. Let us write $\hat{\beta}^T = (\hat{\beta}_0 , \hat{\beta}_K^T)$ for the least-squares estimate of $\beta^T = (\beta_0 , \beta_K^T)$, and let

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \mathbf{m}^T \\ \mathbf{m} & \mathbf{D} \end{bmatrix}, \qquad \mathbf{D}_K = \mathbf{D} - \frac{1}{n}\mathbf{m}\mathbf{m}^T.$$

Then the Scheffé $100\alpha\%$ confidence ellipsoid for the elements of β_K is the region

$$S_\alpha = \{\beta_K : (\beta_K - \hat{\beta}_K)^T\mathbf{D}_K(\beta_K - \hat{\beta}_K) < k\hat{\sigma}^2F_\alpha\} \tag{7}$$

where F_α is the upper $100\alpha\%$ quantile of the F distribution on k, $n - k - 1$ degrees of freedom.

Notice that S_α can be written

$$S_\alpha = \left\{\beta_K : \frac{1}{\hat{\sigma}}(\beta_K - \hat{\beta}_K) \in S_\alpha^*\right\}$$

where S_α^* is a fixed ellipsoid centered at the origin:

$$S_\alpha^* = \{\gamma : \gamma^T\mathbf{D}_K\gamma < kF_\alpha\}.$$

† Srikantan (1970) has proposed the average, rather than the sum, of the squared canonical correlations as an overall measure of association. This measure has the property that its value is changed when a new variable, completely uncorrelated with all the previous ones, is added to one of the sets of variates.

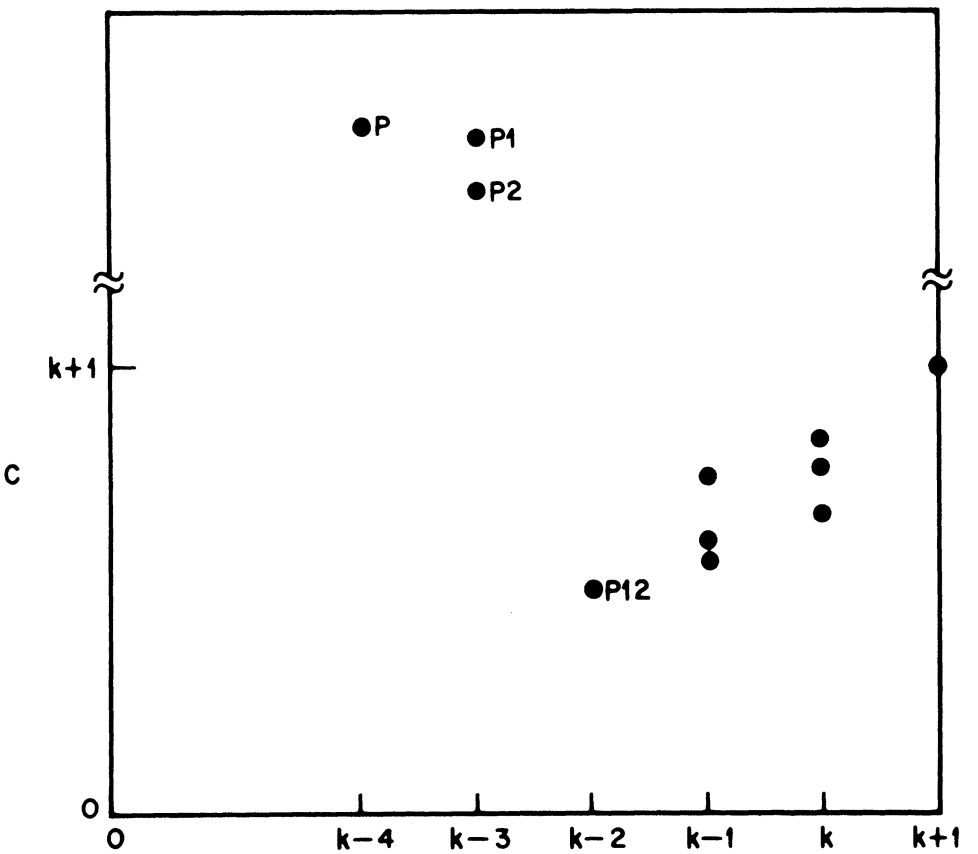

FIGURE 3—C_P-plot: Two variables that jointly are explanatory but separately are not.

Let P^-, Q be any complementary subsets of K, $P = \{0, P^-\}$.
The following lemma is proved in the Appendix.

Lemma The following statements are equivalent:
(i) The region S_α intersects the coordinate hyperplane $H_P = \{\beta_K : \beta_Q = 0\}$,
(ii) The projection of S_α onto the H_Q hyperplane contains the origin,
(iii) The subset least squares estimate $\hat{\beta}_P = (\hat{\beta}_0 , \hat{\beta}_{P^-})$ has $\hat{\beta}_{P^-}$ in S_α ,
(iv) $C_P < 2p - k - 1 + kF_\alpha$,
(v) $\mathrm{RSS}_P - \mathrm{RSS}_{K^+} < k\hat{\sigma}^2 F_\alpha$.

Now consider any hypothesis that specifies the value of β_K , and the corresponding $100\alpha\%$ acceptance region

$$T_{\beta_K} = \left\{\hat{\beta}_K : \frac{1}{\hat{\sigma}}(\hat{\beta}_K - \beta_K) \in S_\alpha^*\right\}. \tag{8}$$

(clearly $P(\hat{\beta}_K \in T_{\beta_k^\circ} \mid \beta_K^0)$ is in fact equal to α; this is just the confidence property of the Scheffé ellipsoid (7)). Starting from this family of acceptance regions for hypotheses that specify β_K completely, a natural acceptance region for a composite hypothesis of the form $\beta_Q = 0$ is given by the union of all regions $T_{\beta_k^\circ}$ for values of β_K^0 such that $\beta_Q^0 = 0$; the reasoning is that the hypothesis $\beta_Q = 0$ cannot be rejected if there is any β_K with $\beta_Q = 0$ that is acceptable according to the corresponding test in the family, i.e. if there is any β_K with $\beta_Q = 0$ lying within the confidence ellipsoid S_α . By the Lemma, the corresponding acceptable subsets

TABLE 1(a)

Values of C_P -p that are exceeded with probability α when $\beta = \beta_P$; $q = k + 1 - p$, $\nu = 30$.

$q = k + 1 - p$	1	2	3	4	5	6	7	8	9	10	15	20	30
α = .10	1.88	2.98	3.83	4.57	5.25	5.88	6.49	7.07	7.64	8.20	10.83	13.35	18.19
.05	3.17	4.63	5.77	6.76	7.67	8.52	9.34	10.13	10.90	11.65	15.22	18.63	25.23
.01	6.56	8.78	10.53	12.07	13.50	14.84	16.13	17.38	18.60	19.79	25.20	30.97	41.58

TABLE 1(b)

Values of C_P -p that are exceeded with probability α when $\beta = \beta_P$; $q = k + 1 - p$, $\nu = \infty$.

$q = k + 1 - p$	1	2	3	4	5	6	7	8	9	10	15	20	30
α = .10	1.71	2.61	3.25	3.78	4.24	4.65	5.02	5.36	5.68	5.99	7.31	8.41	10.26
.05	2.84	3.99	4.82	5.49	6.07	6.59	7.07	7.51	7.92	8.31	10.00	11.41	13.77
.01	5.63	7.21	9.34	9.28	10.09	10.81	11.48	12.09	12.67	13.21	15.58	17.57	20.89

$\{0, P^-\}$ are just those that have

$$C_P < 2p - k - 1 + kF_\alpha . \tag{9}$$

We state the property formally:

A subset $P = \{0, P^-\}$ satisfies (9) if and only if there is some vector of coefficients β having $\beta_Q = \mathbf{0}$ that lies within the Scheffé ellipsoid (7), i.e. if and only if there is some vector of this form that is accepted by the corresponding test with acceptance region of the form (8).

As an example, consider the 10-variable data studied by Gorman and Toman (1966). Taking $\alpha = 0.10$, $k = 10$, $\nu = 25$, we find that among the 58 subsets for which Gorman and Toman computed C_P-values, there are 39 that satisfy (9), in number 7, 13, 9, 10 with $p = 7, 8, 9, 10$ respectively. This result gives little support to the view that this set of data is sending a clear message regarding the relative importance of the variables under consideration.

Notice that if the true coefficient vector β^* has $\beta^*_{\bar{Q}} = \mathbf{0}$, then Pr {for all P containing P^*, $C_P \leq 2p - k - 1 + kF_\alpha\} \geq \alpha$, with equality only if $p^* = 1$ (i.e. $P^* = \{0\}$). This property of the procedure is not completely satisfying since it is not an equality; also the form of the boundary in the C_P-plot is inflexible. In theory, one way of getting a better result is the following. Given any subset P^* and a sequence of constants $c_1, c_2, \cdots, c_k$ (and the matrix $\mathbf{D}_K$) one could compute the probability Pr {for all P containing P^*, $C_P < c_p$}; this probability depends on $c_1, \cdots, c_k$, P^* and $\mathbf{D}_K$, but not on any other parameters. One could then adjust $c_1, \cdots, c_k$ so as to make the minimum of this probability over all choices of P^* (or possibly only over all choices with $p^* >$ some p_0) equal to some desired level α. The computation would presumably be done by simulation.

Starting from the Scheffé ellipsoid, Spjøtvoll (1972) has developed a multiple-comparison approach that provides confidence intervals for arbitrary quadratic functions of the unknown regression parameters, for example $B_P - B_{P'}$ for two subsets P, P'.

4. Formal Selection of Subset Regressions

Many authors have studied the problem of giving formal rules for the selection of predictors; Kennedy and Bancroft (1971) give many references. Lindley (1968) presents a Bayesian formulation of the problem. The discussion in section 3 above does not lend any support to the practice of taking the lowest point on a C_P-plot as defining a "best" subset of terms. The present author feels that the greatest value of the device is that it helps the statistician to examine some aspects of the structure of his data and helps him to recognize the ambiguities that confront him. The device cannot be expected to provide a single "best" equation when the data are intrinsically inadequate to support such a strong inference.

To make these remarks more precise and objective, we shall compute (in a special case) a measure of the performance to be expected of the rule "choose the subset that minimizes C_P, and fit it by least-squares". We shall use as a figure of merit of an arbitrary estimator $\hat{\eta}(x)$ the same quantity as was used in setting up the C_P-plot, namely the sum of predictive squared errors

$$J_{\hat{\eta}} = \frac{1}{\sigma^2} \sum_{u=1}^{n} (\hat{\eta}(\mathbf{x}_u) - \mathbf{x}_u\beta)^2.$$

We can handle in detail only the case of orthogonal regressors, and so now assume $\mathbf{X}^T\mathbf{X} = nI$. In this case we see from (5) that C_P is minimized when P contains

just those terms for which $t_j^2 > 2$, where $t_j = \sqrt{n}\, b_j/\hat{\sigma}$ is the t-statistic for the j-th regression coefficient, and b_j is the least squares estimate, $b_j = \Sigma x_{uj} y_u/n$. Thus in this case the "minimize C_P" rule is equivalent to a stepwise regression algorithm in which all critical t-values are set at $\sqrt{2}$ and $\hat{\sigma}^2$ is kept at the full-equation value throughout.

Now let us assume that n is sufficiently large that variation in $\hat{\sigma}$ can be ignored; then $t_0, t_1, \cdots, t_k$ will be independent Normal variables with unit variances and with means $\tau_0, \cdots, \tau_k$ where $\tau_i = \sqrt{n}\, \beta_i/\hat{\sigma}$. Let $d(t)$ be the function that equals 0 for $|t| \leq \sqrt{2}$, and equals 1 otherwise, then J for the "minimum-C_P subset least squares" estimate can be written

$$J_{\text{Min } C_P} = \frac{1}{\sigma^2} \sum_{u=1}^{n} \left(\sum_{j=0}^{k} x_{uj}(b_j\, d(t_j) - \beta_j) \right)^2$$

which reduces to

$$J_{\text{Min } C_P} = \sum_{j=0}^{k} (t_j\, d(t_j) - \tau_j)^2.$$

Hence

$$E(J_{\text{Min } C_P}) = \sum_{j=0}^{k} m(|\tau_j|)$$

where $m(\tau) = E((u + \tau)d(u + \tau) - \tau)^2$ (where u is a standard Normal variable), and is the function displayed in Figure 4 (labelled "16%", since Pr $\{|u| > \sqrt{2}\}$ =

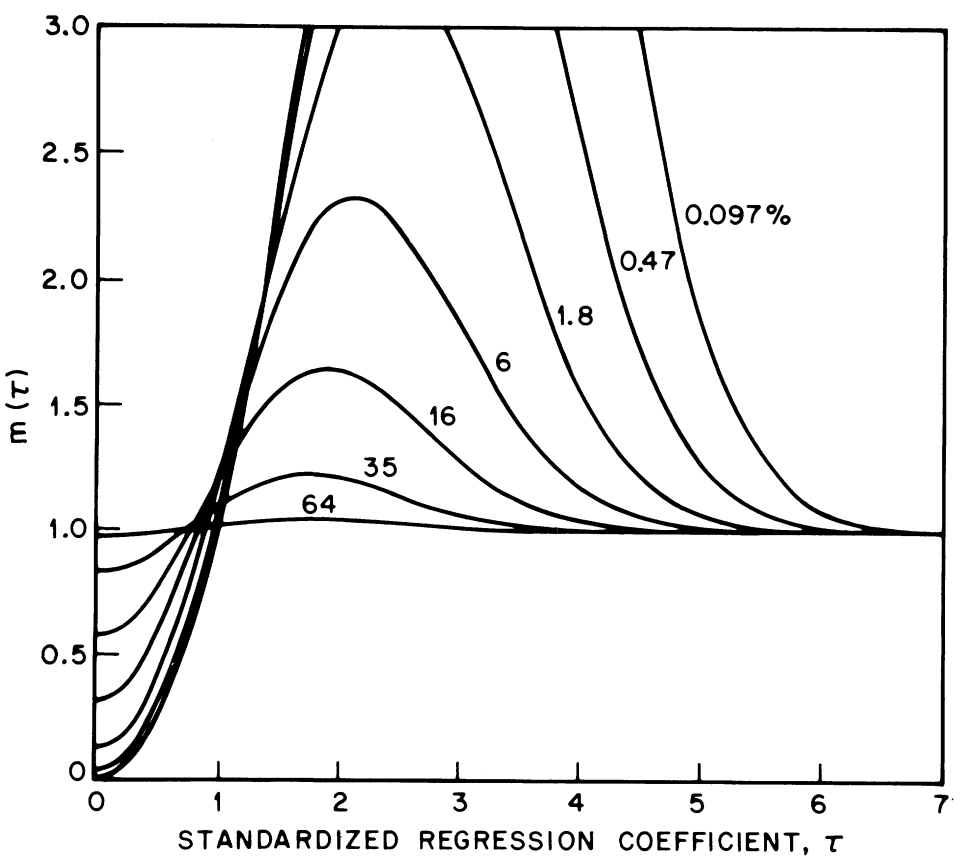

FIGURE 4—m-functions

.1573). If the constant term is always to be included in the selected subset, the corresponding result is

$$E(J_{\text{Min } C(0,P-1)}) = 1 + \sum_{j=1}^{k} m(|\tau_j|).$$

Notice that the function $m(\tau)$ is less than 1 only for $|\tau| < .78$, and rises to a maximum value of 1.65 at $|\tau| = 1.88$. It exceeds 1.25 for $1.05 < |\tau| < 3.05$.

We reiterate that in this case of orthogonal regressors with n very large, the "minimum C_P" rule is equivalent to a stepwise regression algorithm with all critical levels set at 15.73%. Also shown in Figure 4 are the m-functions corresponding to several other critical levels; when all $k + 1$ terms are infallibly included (the "full-1.s." rule), $m(\tau) = 1$ for all τ, so that $E(J_{\text{full 1.s.}}) = k + 1$. We see that the "minimum C_P" rule will give a smaller value for $E(J)$ than the "full-1.s." rule only when rather more of the true regression coefficients satisfy $|\tau| < .78$ than satisfy $|\tau| > 1$; in the worst case with $|\tau_j| = 1.88$ for $j = 1, \cdots, k$, $E(J)$ for the "minimize C_P" rule is 165% of that for the "full-1.s." rule. Similarly for rejection rules with other critical levels; in particular, a rule with a nominal level of 5% (two tailed) gives an $E(J)$ at worst 246% of that of the "full-1.s." rule.

Thus using the "minimum C_P" rule to select a subset of terms for least-squares fitting cannot be recommended universally. Notice however that by examining the C_P-plot in the light of the distributional results of the previous section one can see whether or not a single best subset is uniquely indicated; the ambiguous cases where the "minimum C_P" rule will give bad results are exactly those where a large number of subsets are close competitors for the honor. With such data no selection rule can be expected to perform reliably.

5. C_L-Plots

We now extend the C_P-plot device to handle general linear estimators. With the same set-up as in the Introduction, consider an estimate of the form

$$\boldsymbol{\beta}_L = \begin{pmatrix} \bar{y} \\ \mathbf{L}y \end{pmatrix}$$

where $\bar{y}$ is the mean $\bar{y} = \Sigma y_u/n$, and L is a $k \times n$ matrix of constants. We shall assume that $\mathbf{L}\mathbf{1}_n = \mathbf{0}$ (where $\mathbf{1}_n^T = (1, 1, \cdots, 1)$) so that a change in the origin of measurement of y affects only $\hat{\beta}_0$ and not $\hat{\beta}_1, \cdots, \hat{\beta}_k$. Examples of estimators of this class are: full least-squares; subset-least-squares; and Bayes estimates under multinormal specifications with a multinormal prior, a special case of which is the class of "Ridge" estimates advocated by Hoerl and Kennard (1970a, b), (see also Theil (1963), section 2.3):

$$\mathbf{L} = \mathbf{L}_f = (\mathbf{X}^T\mathbf{X} + f\mathbf{I})^{-1}\mathbf{X}^T \tag{11}$$

where f is a (small) scalar parameter (Hoerl and Kennard used k), and in this section we are writing $\mathbf{X}$ for the $n \times k$ matrix of independent variables, which we are now assuming have been standardized to have zero means and unit variances. Thus $\mathbf{1}_n^T\mathbf{X} = \mathbf{0}$, diag $(\mathbf{X}^T\mathbf{X}) = \mathbf{I}$.

As a measure of adequacy for prediction we again use the scaled summed mean square error, which in the present notation is

$$J_L = \frac{1}{\sigma^2}\left(||\mathbf{X}\hat{\boldsymbol{\beta}}_L - \mathbf{X}\boldsymbol{\beta}||^2 + n(\bar{y} - \beta_0)^2\right)$$

and which has expectation

$$E(J_{\mathbf{L}}) = V_{\mathbf{L}} + \frac{1}{\sigma^2} B_{\mathbf{L}}$$

where

$$V_{\mathbf{L}} = 1 + \operatorname{tr}(\mathbf{X}^T\mathbf{X}\mathbf{L}\mathbf{L}^T)$$

$$B_{\mathbf{L}} = \beta_k^T(\mathbf{L}\mathbf{X} - I)^T\mathbf{X}^T\mathbf{X}(\mathbf{L}\mathbf{X} - I)\beta_K .$$

The sum of squares about the fitted regression is

$$\mathrm{RSS}_{\mathbf{L}} = ||\mathbf{y} - \bar{y}\mathbf{1}_n - \mathbf{X}\hat{\beta}_{\mathbf{L}}||^2$$

which has expectation

$$E(\mathrm{RSS}_{\mathbf{L}}) = \hat{\sigma}^2 V_{\mathbf{L}}^* + B_{\mathbf{L}}$$

where

$$V_L^* = n - 1 - 2\operatorname{tr}(\mathbf{X}\mathbf{L}) + \operatorname{tr}(\mathbf{X}^T\mathbf{X}\mathbf{L}\mathbf{L}^T).$$

Thus we have an estimator of $E(J_{\mathbf{L}})$, namely

$$C_{\mathbf{L}} = \frac{1}{\hat{\sigma}^2}\mathrm{RSS}_L - n + 2 + 2\operatorname{tr}(\mathbf{X}\mathbf{L}). \tag{12}$$

By analogy with the C_P development, we propose that values of $C_{\mathbf{L}}$ (for various choices of $\mathbf{L}$) should be plotted against values of $V_{\mathbf{L}}$. Notice that when $\mathbf{L}$ is a matrix corresponding to subset least squares, $C_{\mathbf{L}}$, $V_{\mathbf{L}}$ reduce to C_P , p respectively.

For computing values of $C_{\mathbf{L}}$, $V_{\mathbf{L}}$ for Ridge estimates (11), the following steps can be taken. First, find $\mathbf{H}$ (orthogonal) and $\mathbf{\Lambda}$ = diagonal $(\lambda_1 , \lambda_2 , \cdots , \lambda_k)$ so that $\mathbf{X}^T\mathbf{X} = \mathbf{H}^T\mathbf{\Lambda}\mathbf{H}$. Compute $\mathbf{z} = \mathbf{H}\mathbf{X}^T\mathbf{y}$. Then

$$\begin{aligned} V_{\mathbf{L}_f} &= 1 - \sum_{i=1}^{k}\left(\frac{\lambda_i}{f + \lambda_i}\right)^2 \\ \operatorname{tr}(\mathbf{X}\mathbf{L}) &= \sum_{i=1}^{k}\frac{\lambda_i}{f + \lambda_i} \\ \mathrm{RSS}_{\mathbf{L}_f} - \mathrm{RSS}_{\mathbf{L}_0} &= \sum_{i=1}^{k}\frac{f^2 z_i^2}{\lambda_i(f + \lambda_i)^2}. \end{aligned} \tag{13}$$

Figure 5 gives the resulting plot for the set of 10-variable data analyzed by Gorman and Toman (1966) and by Hoerl and Kennard (1970b). Shown are (p, C_P) points corresponding to various subset-least-squares estimates and a continuous arc of $(V_{\mathbf{L}} , C_{\mathbf{L}})$ points corresponding to Ridge estimates with values of f varying from zero at (11, 11) and increasing to the left. For this example, Hoerl and Kennard (1970b) suggested that a value of f in the interval (0.2, 0.3) would "undoubtedly" give estimated coefficients "closer to β and more stable for prediction than the least-squares coefficients or some subset of them". On the other hand from Figure 5 one would be inclined to suggest a value of f nearer to .02 than to 0.2.

One obvious suggestion is to choose f to minimize $C_{\mathbf{L}_f}$. Some insight into the effect of this choice can be gained as follows. First consider the case of orthogonal regressors, and now assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. Notice that in this case our risk function $E(J)$ is equivalent to the quantity $\sum_{i=1}^{k} E(\hat{\beta}_i - \beta_i)^2$ used by Hoerl and Kennard (1970a). We may take $\mathbf{H} = \mathbf{I}$, so that $z_i = \hat{\beta}_i$, the least-squares estimate of β_i .

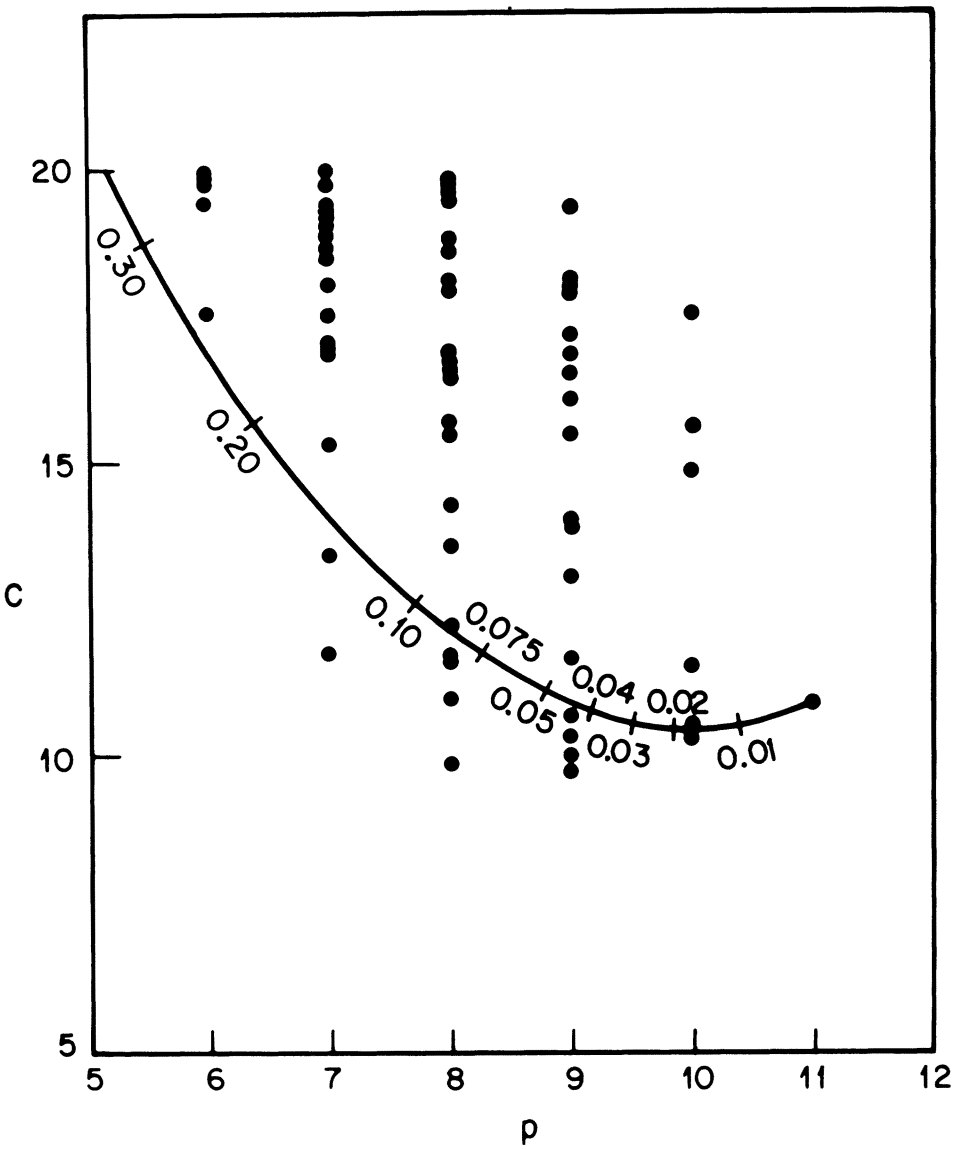

FIGURE 5—C_L plot for Gorman-Toman data; subset (C_P) values and Ridge (C_{L_f}) values.

From (12) and (13) we see that C_{L_f} is a minimum when f satisfies

$$(1 + f)/f = \sum_{i=1}^{k} \beta_i^2/k\sigma^2;$$

the adjusted estimates are then given by

$$\beta_i^* = \left(1 - \frac{k\sigma^2}{\Sigma\hat{\beta}_i^2}\right)\hat{\beta}_i \qquad i = 1, \cdots, k. \tag{14}$$

It is interesting that this set of estimates is of the form suggested by Stein (1960) for the problem of estimating regression coefficients in a multivariate Normal distribution. James and Stein (1961) showed that for $k \geq 3$ the vector of estimates $\hat{\beta}^{**}$ obtained by replacing the multiplier k in (14) by any number between 0 and $2k - 4$ has the property that $E(J^{**})$ is less than the full-least-squares value

$k + 1$ (see (10)), for all values of the true regression coefficients. Thus our "minimize C_{L_f}" rule dominates full least-squares for $k \geq 4$. This result stands in interesting contrast to the disappointing result found above for the "minimize C_P" rule.

Now, consider the case of equi-correlated regressors, with $\mathbf{X}^T\mathbf{X} = \mathbf{I} + \rho(\mathbf{1}\mathbf{1}^T - \mathbf{I})$. In this case the least-squares estimate $\hat{\bar{\beta}}$ of $\bar{\beta} = \Sigma\beta_i/k$ has variance $1/k(1 - \rho + k\rho)$, and the vector of deviations $(\hat{\beta}_i - \hat{\bar{\beta}})$ has covariance matrix $(I - k^{-1}\mathbf{1}\mathbf{1}^T)/(1 - \rho)$. Thus when ρ is large, these deviations become very unstable.

It is found that for ρ near unity, C_{L_f} is minimized when f is near $(1 - \rho)g$, where

$$\frac{1 + g}{g} = \frac{(1 - \rho)}{(k - 1)\hat{\sigma}^2} \sum_{i=1}^{k} (\hat{\beta}_i - \hat{\bar{\beta}})^2.$$

The adjusted estimates are given approximately by

$$\hat{\beta}_i^* = \hat{\bar{\beta}} + \left(1 - \frac{g}{1 + g}\right)(\hat{\beta}_i - \hat{\bar{\beta}}).$$

Thus here the "minimize C_{L_f}" rule leads to shrinking the least-squares estimates towards their average. While the details have not been fully worked out, one expects that this rule will dominate full least-squares for $k \geq 5$.

6. Acknowledgements

It is a great personal pleasure to recall that the idea for the C_P-plot arose in the course of some discussions with Cuthbert Daniel around Christmas 1963. The use of the letter C is intended to do him honor. The device was described publicly in 1964 [16] and again in 1966 [17] (with the extensions described at the end of section 1 above) and has appeared in several unpublished manuscripts. Impetus for preparing the present exposition was gained in the course of delivering a series of lectures at the University of California at Berkeley in February 1972; their support is gratefully acknowledged.

7. Appendix

Proof of the Lemma

The key to these results is the identity, true for any subset P that includes 0, i.e. $P = \{0, P^-\}$,

$$\mathrm{RSS}_{P^-} - \mathrm{RSS}_{K^+} = (\hat{\beta}_{P^-} - \hat{\beta}_K)^T\mathbf{D}_K(\hat{\beta}_{P^-} - \hat{\beta}_K)$$

where $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_K^T)$ is the vector of least-squares estimates of all the coefficients in the model, and $(\hat{\beta}_0^-, \hat{\beta}_{P^-}^T)$ is the vector of subset-least-squares estimates. From the form of S_α (7) it now follows that (iii) $\hat{\beta}_{P^-}$ is in S_α if and only if (v) $\mathrm{RSS}_{P^-} - \mathrm{RSS}_{K^+} < k\hat{\sigma}^2F_\alpha$, which is directly equivalent (if $\hat{\sigma}^2 = \mathrm{RSS}_{K^+}/n - k - 1$) to (iv) $C_P < 2p - k - 1 + kF_\alpha$. Clearly (iii) implies (i); to prove the converse we remark that for any vector $\beta^T = (\beta_0, \beta_K^T)$ with β_K in the hyperplane $H_P = \{\beta_K : \beta_Q = \mathbf{0}\}$, we have

$$||\mathbf{X}\hat{\beta} - \mathbf{X}\beta||^2 = ||\mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_P||^2 + ||\mathbf{X}\hat{\beta}_P - \mathbf{X}\beta||^2,$$

the cross-product term vanishing by definition of β_P. Thus if any point of H_P is in S_α, β_P must be. Finally, (i) is directly equivalent to (ii) by a simple geometrical argument.

To handle the case of non-orthogonal regressors, Sclove (1968) has suggested transforming to orthogonality before applying a shrinkage factor. A composite procedure with much intuitive appeal for this writer would be to use the C_P plot or some similar device to identify the terms that should certainly be included (since they appear in all subsets that give reasonably good fits to the data), to fit these by least squares, and to adjust the remaining estimates by orthogonalizing and shrinking towards zero as in (14).

References

[1] Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics, 13*, 469–475.

[2] Beale, E. M. L., Kendall, M. G. and Mann, D. W. (1967). The discarding of variables in multivariate analysis, *Biometrika 54*, 357–366.

[3] Daniel, C. and Wood, F. S. *Fitting Equations to Data* New York: Wiley-Interscience, 1971.

[4] Furnival, G. M. (1971). All possible regressions with less computation. *Technometrics 13*, 403–408.

[5] Garside, M. J. (1965). The best subset in multiple regression analysis. *Applied Statistics 14*, 196–200.

[6] Godfrey, M. B. (1972). Relations between C_p, RSS, and mean square residual. Submitted to *Technometrics*.

[7] Gorman, J. W. and Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics 8*, 27–51.

[8] Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics 9*, 531–540.

[9] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics 12*, 55–67.

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

[11] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp.* 1, 361–379, Univ. of California Press.

[12] Kennard, R. W. (1971). A note on the C_p statistic. *Technometrics 13*, 899–900.

[13] Kennedy, W. J. and Bancroft, T. A. (1971). Model-building for prediction in regression based on repeated significance tests. *Ann. Math. Statist. 42*, 1273–1284.

[14] La Motte, L. R. and Hocking, R. R. (1970). Computational efficiency in the selection of regression variables. *Technometrics 12*, 83–93.

[15] Lindley, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. B 30*, 31–53, (Discussion, 54–66).

[16] Mallows, C. L. (1964). Choosing variables in a linear regression: a graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7–9.

[17] Mallows, C. L. (1966). Choosing a subset regression. Presented at the Annual Meeting of the American Statistical Association, Los Angeles, August 15–19, 1966.

[18] Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics 12*, 621–625.

[19] Schatzoff, M., Tsao, R. and Fienberg, S. (1968). Efficient calculation of all possible regressions. *Technometrics 10*, 769–779.

[20] Sclove, S. L. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Statist. Assoc. 63*, 596–606.

[21] Spjøtvoll, E. (1972). Multiple comparison of regression functions. *Ann. Math. Statist. 43*, 1076–1088.

[22] Srikantan, K. S. (1970). Canonical association between nominal measurements. *J. Amer. Statist. Assoc. 65*, 284–292.

[23] Stein, C. (1960). Multiple regression. *Contributions to Probability and Statistics*, ed. I. Olkin, Stanford University Press, 424–443.

[24] Theil, H. (1963). On the use of incomplete prior information in regression analysis. *J. Amer. Statist. Assoc. 58*, 401–414.

[25] Watts, H. W. (1965). The Test-o-Gram; a pedagogical and presentational device. *Amer. Stat. 19*, No. 4, 22–28.