Ridge Regression in Practice*

DONALD W. MARQUARDT AND RONALD D. SNEE**

SUMMARY

The use of biased estimation in data analysis and model building is discussed. A review of the theory of ridge regression and its relation to generalized inverse regression is presented along with the results of a simulation experiment and three examples of the use of ridge regression in practice. Comments on variable selection procedures, model validation, and ridge and generalized inverse regression computation procedures are included. The examples studied here show that when the predictor variables are highly correlated, ridge regression produces coefficients which predict and extrapolate better than least squares and is a safe procedure for selecting variables.

I. Theory and Illustrative Examples

Part I of this paper focuses on the theory of biased estimation: not so much the algebraic details as the concepts involved. Several illustrative examples help clarify the concepts. While we emphasize ridge regression, we also discuss its relationship to generalized inverse regression. The second part of the paper discusses two larger examples where one can see ridge regression at work in data analysis in a realistic setting.

The types of data sets to which we are addressing ourselves can be messy for a variety of reasons. The predictor variables may be correlated because historical data were collected without the aid of an experimental design. Physical and mathematical constraints may necessitate correlated predictor variables, even when an experimental design is used. The presence of gross errors, missing values, correlated errors, split plotting, nonconstant variance, and other problems can create nonsense results, even when we employ sophisticated regression techniques to deal with the correlation problem. For purposes of this paper we assume that none of these other problems is present.

We emphasize that biased estimation is only one of the tools one uses in the analysis of a set of data. One starts by understanding the technical background of the problem and the candidate variable definitions. Next, the form of the model and the need for transformations are considered. Then, the data are examined for abnormal values, scatter plots are constructed to look for relationships, and subsequently the residuals are examined for randomness. If the variance inflation factors (VIF) [18, 22] of the least squares estimates are large, then it is appropriate to consider a biased estimation procedure such as ridge regression in order to minimize the effects of the predictor variable correlations and develop a set of stable coefficients. These initial steps often consume as much as 75% or more of the total effort expended in solving the problem. In this paper the emphasis will be on coefficient estimation; however, we do not want to leave the impression that we do not believe the other steps are important. In many instances these aspects are the most important part of the problem.

Comments on Some Common Practices

As we survey the literature and reflect upon the state of the art of regression analysis with large numbers of predictor variables, we have identified a number of practices about which we would like to comment before discussing ridge regression, *per se*.

One common practice we note is failure to remove nonessential ill conditioning through the use of standardized predictor variables. Standardizing of the predictors is appropriate whenever a constant term is present in the model. The ill conditioning that results from failure to standardize is all the more insidious because it is not due to any real defect in the data, but only to the arbitrary origins of the scales on which the predictor variables are expressed. In standardizing the predictor variables, the mean is subtracted from each variable ("centering") and then the centered variable is divided by its standard deviation ("scaling"). Centering removes the nonessential ill-conditioning, thus reducing the variance inflation in the coefficient estimates. In a linear model centering removes the correlation between the constant term and all linear terms. In addition, in a quadratic model centering reduces, and in certain situations completely removes, the correlation between the linear and quadratic terms. Scaling expresses the equation in a form that lends itself to more straight-forward interpretation and use.

The Acetylene data [7, 13] shown in Table 1 is the

Table 1

ACETYLENE DATA

x_1 Reactor Temperature (°C)	x_2 Ratio of H_2 to n-heptane (mole ratio)	x_3 Contact Time (sec)	y Conversion of n-heptane to Acetylene (%)
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	29.5

* This paper is based on a presentation by the authors at the University of Kentucky Conference on Regression with a Large Number of Predictor Variables, Lexington, Kentucky, October 1973.

** Engineering Dept., E. I. du Pont de Nemours & Co., Inc., Wilmington, DE 19898.

first example we will discuss. This is a typical set of response surface data for which a full quadratic model in x_1, x_2, x_3 is an appropriate candidate.

In Figure 1 two of the predictor variables are plotted against each other. Correlations like this cause inflation

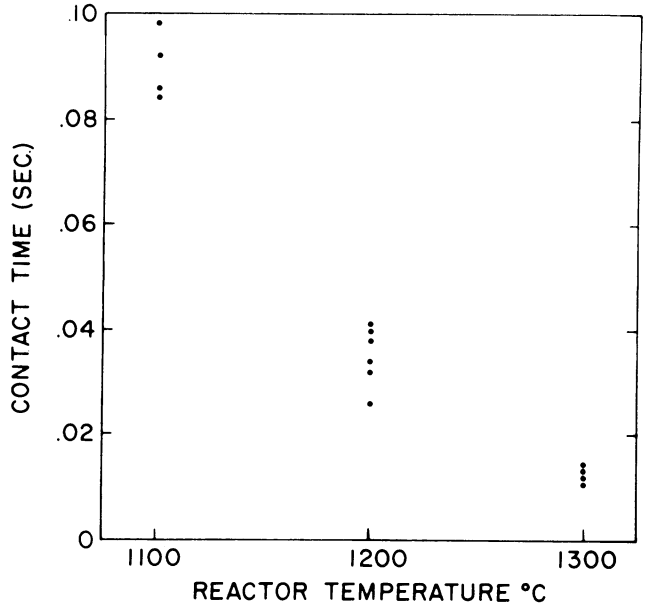

Figure 1.

of the variance of the estimated coefficients in the least squares model. The variance inflation factor [18, 22] for each term in the model measures the collective impact of these simple correlations on the variance of the coefficient of that term. The variance inflation factors are the diagonal elements of the inverse of the simple correlation matrix. As the multiple correlation of any predictor with the other predictors approaches unity, the corresponding VIF becomes infinite.

The variance inflation factors for the acetylene data are shown in Table 2. The maximum variance inflation factor is the best single measure of the conditioning of the data. For any predictor orthogonal to all other predictors, the inflation factor is 1.0. The inflation factors for the standardized model are in the second column. Scaling does not affect the VIF's, but centering does. A maximum factor of six thousand is horrible but two million is unthinkable and unnecessary. Now the question remains: Once we have standardized, how do we carry out a meaningful analysis of data with an inflation factor of over six thousand?

There are serious limitations of classical methodologies when they are employed for the analysis of ill-conditioned data. The first of these classic methodologies is least squares. It does not provide good estimators when the data are ill conditioned. In achieving optimum fit to the estimation data, least squares often destroys good prediction of new data (possibly outside the region covered by the estimation data).

The second classical methodology is variable selection as a technique for reducing the degree of ill conditioning. Variable selection implies a simplisitic two-valued classification logic wherein any predictor variable must either be important or unimportant. Large prediction biases can result from elimination of "non-significant" predictors. It is better to use a little bit of all the variables than all of some variables and none of the remaining ones. This is what biased estimators do. We will show how biased estimators can alleviate both of these limitations.

Finally, in our comments on current practices, we observe that most statisticians restrict themselves to models linear in the parameters. Frequently the known background of the problem suggests a function nonlinear in the parameters, one that may provide a simpler and more natural model. We will illustrate this with one of our examples.

Formulation of the Problem

In linear estimation one postulates a model of the

Table 2

ACETYLENE DATA REGRESSION RESULTS

Ten-Coefficient Quadratic Model

Term	Least Squares - VIF		Correlation Basis Coefficient			
				Ridge		
	Unstandardized	Standardized	Least Squares	k = .01	k = .05	Generalized Inverse (r=3.8)
x_1 = Temperature	2,856,748.93	375.25	.336	.589	.522	.507
x_2 = H_2 / n-Heptane	10,956.14	1.74	.233	.216	.209	.180
x_3 = Contact Time	2,017,162.52	680.28	-.676	-.327	-.379	-.414
x_1x_2	9,802.90	31.04	-.480	-.326	-.202	-.095
x_1x_3	1,428,091.88	6,563.35	-2.034	-.094	-.061	-.051
x_2x_3	240.36	35.61	-.266	-.083	.042	.123
x_1^2	2,501,944.59	1,762.58	-.835	.126	.125	.165
x_2^2	65.73	3.16	-.090	-.054	-.047	-.063
x_3^2	12,667.10	1,156.77	-1.001	-.069	-.024	-.053
Maximum VIF			6,563.35	12.38	2.63	.46
R_A^2			.994	.990	.983	.973

form

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon. \tag{1}$$

The $n \times p$ matrix $\mathbf{X}$ contains the values of p predictor variables at each of n data points, $\mathbf{Y}$ is the vector of observed values, β is the $p \times 1$ vector of population values of the parameters, and ε is an $n \times 1$ vector of experimental errors having the properties $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2\mathbf{I}_n$. For convenience, we assume that the x variables are scaled so that $\mathbf{X}'\mathbf{X}$ has the form of a correlation matrix.

The conventional estimator for β is the least squares estimator, $\hat{\beta}$, where $\hat{\beta}$ is chosen to minimize the sum of squares of residuals $\Phi(\hat{\beta})$:

$$\Phi(\hat{\beta}) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{g}$$

Note that $\mathbf{g}$ is the gradient of Φ.

The two key properties of $\hat{\beta}$ are that it is unbiased, that is, $E(\hat{\beta}) = \beta$, and that it has minimum variance among all linear unbiased estimators. The covariance matrix is

$$V(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

In their development of ridge regression [10, 11], Hoerl and Kennard focus attention on the eigenvalues of $(\mathbf{X}'\mathbf{X})$. A seriously non-orthogonal (or "ill-conditioned") problem is characterized by the fact that the smallest eigenvalue, $\lambda_{\min}$, is very much smaller than unity. For example, the smallest eigenvalues of the $\mathbf{X}'\mathbf{X}$ matrices for the acetylene data ($p = 9$ and 5), and the Laird and Cady ($p = 33$), and GC-ASTM ($p = 15$) examples to be discussed later are 0.00010, 0.01005, 0.00207, and .00027, respectively. It should also be noted that the variance inflation factor (VIF) mentioned earlier is a measure of how close the smallest eigenvalues are to zero [18, 22]. Hoerl and Kennard have summarized the dramatic inadequacy of least squares for nonorthogonal problems by noting that the expected squared length of the coefficient vector is

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \operatorname{Tr}(\mathbf{X}'\mathbf{X})^{-1}$$

$$> \beta'\beta + \sigma^2/\lambda_{\min}.$$

Thus, $\hat{\beta}$, the least squares coefficient vector, is much too long, on the average, for ill-conditioned data, since $\lambda_{\min} \ll 1$. The least squares solution yields coefficients whose absolute values are too large and whose signs may actually reverse with negligible changes in the data.

The "fly in the ointment" with least squares is its requirement of unbiasedness. Figure 2, top, illustrates the situation where an estimator $\hat{\beta}$ is unbiased but is plagued by a large variance. Typical confidence limits for this estimator would be nearly half the width of the figure. At the bottom is the corresponding frequency function for a biased estimator with much smaller variance. Statistical limits for this situation would be perhaps twenty percent of the width of the figure. Thus, it is meaningful to focus on the achievement of small

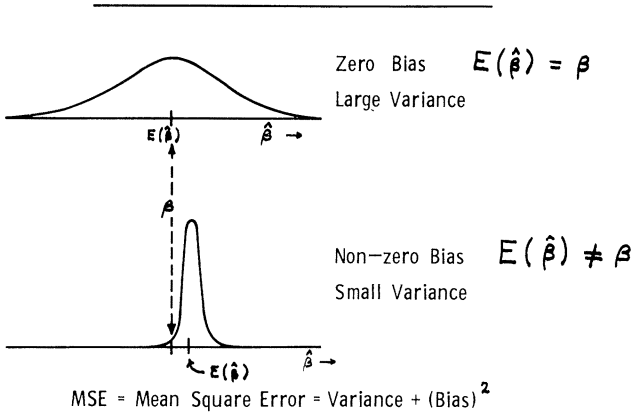

Figure 2.

mean square error as the relevant criterion, if a major reduction in variance can be obtained as a result of allowing a little bias. This is precisely what the ridge and generalized inverse solutions accomplish.

Ridge Solution

The ridge estimator is obtained by solving

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})\hat{\beta}^* = \mathbf{g}$$

to give

$$\hat{\beta}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{g}$$

for $k \geq 0$. Values of k between zero and 1.0 should be explored. In general, there is an "optimum" value of k for any problem, but it is desirable to examine the ridge solution for a range of admissible values of k. "Admissible" means having smaller mean square error in the parameters than the least squares solution. The mean square error in future predictions is also reduced correspondingly.

Hoerl gave the name "ridge regression" [9] to his procedure because of the similarity of its mathematics to methods he used earlier [8], i.e., "ridge analysis," for graphically depicting the characteristics of second order response surface equations in many predictor variables.

Key properties applicable to ridge regression are [10, 11, 18]:

- If $\hat{\beta}^*$ is the solution of $(\mathbf{X}'\mathbf{X} + k\mathbf{I})\hat{\beta}^* = \mathbf{g}$, then $\hat{\beta}^*$ minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length of $\hat{\beta}^*$. The sum of squares of residuals is an increasing function of k.
- The length of $\hat{\beta}^*$ is a decreasing function of k.
- The angle γ between the ridge solution $\hat{\beta}^*$ and the gradient vector $\mathbf{g}$ is a decreasing function of k.

In Figure 3 we illustrate the geometry of ridge regression for a hypothetical problem involving only two parameters β_1 and β_2. The point $\hat{\beta}$ at the center of the ellipses is the least squares solution. At $\hat{\beta}$ the sum of squares of residuals, Φ, achieves its minimum. The small ellipse is the locus of points in the β_1, β_2-plane where the sum of squares Φ is constant at some value

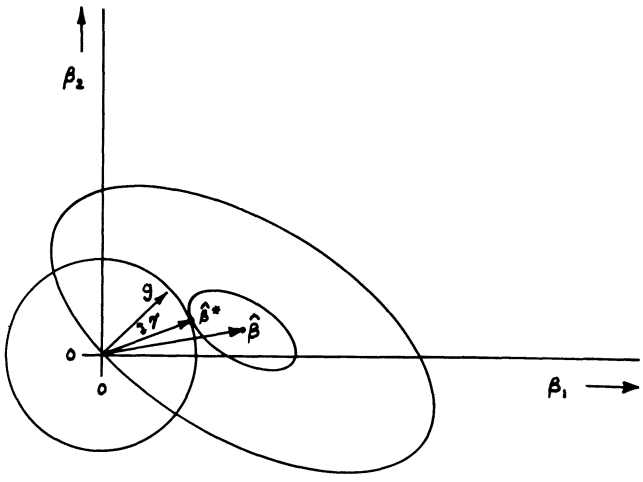

Figure 3.

larger than the minimum value. The circle about the origin is tangent to the small ellipse at $\hat{\beta}^*$. Note that the ridge estimate $\hat{\beta}^*$ is the shortest vector that will give a residual sum of squares as small as the Φ value anywhere on the small ellipse. Thus, the ridge estimate gives the smallest regression coefficients consistent with a given degree of increase in the residual sum of squares. The gradient $\mathbf{g}$ is perpendicular to the Φ-contour through the origin. The ridge estimate $\hat{\beta}^*$ is always between $\hat{\beta}$ and $\mathbf{g}$, the angle γ getting steadily smaller as the quantity k added to the diagonal increases.

Other key properties are:

- $\hat{\beta}^*$ is a linear transform of $\hat{\beta}$:

$$\hat{\beta}^* = \mathbf{Z}_k\hat{\beta} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\hat{\beta}$$

  Thus, $\hat{\beta}^*$ is a *biased estimator* $E(\hat{\beta}^*) = \mathbf{Z}_k\beta$

- $V(\hat{\beta}^*) = \sigma^2[\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}(\mathbf{X}'\mathbf{X})[\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}$

- $\text{ESD} = \text{Tr}[V(\hat{\beta}^*)] + \beta'(\mathbf{Z}_k - \mathbf{I})'(\mathbf{Z}_k - \mathbf{I})\beta$

  $= \text{Variance} + (\text{Bias})^2$

where ESD denotes the expected squared distance to β.

This has two crucially important corollaries:

1. The variance term is a decreasing function of k.
2. The bias term is an increasing function of k.

- If $\beta'\beta$ is bounded, there exists a $k > 0$ such that $\text{ESD}(\hat{\beta}^*) < \text{ESD}(\hat{\beta})$.

We note in passing that the ridge estimator also has an important Bayesian interpretation. In this connection, R. W. Kennard has frequently emphasized the fact that least squares implies an assumption of an unbounded uniform prior distribution on the coefficient vector. This unboundedness assumption can be used in place of the unbiasedness requirement in deriving least squares estimators. When selecting the amount of bias we work with predictor variables and response variable both scaled to correlation form. In this scaling, it is exceedingly rare for the population value of any regression coefficient to be larger than three in a real problem. In any case, the regression coefficient vector is surely finite. The ridge estimate is equivalent to placing mild boundedness requirements on the coefficient vector.

In a recent paper Theobald [23] generalizes the conditions under which ridge is known to produce a smaller expected squared distance than least squares. It is also known that the expected improvement of ridge over least squares depends on the orientation of the true regression vector relative to the principal axes defined by the eigenvectors of the $\mathbf{X}'\mathbf{X}$ matrix, the expected improvement being greatest when the orientation of β coincides with the eigenvector associated with the largest eigenvalue of $\mathbf{X}'\mathbf{X}$. Other results appear in References 1, 15, 16, 19, and 20.

A complete sequence of corresponding properties of the generalized inverse solution is developed in a previous paper [18]. An illustrative example in that paper demonstrates the close geometric similarity between the generalized inverse solution expressed as a function of the rank r assigned to the matrix $\mathbf{X}'\mathbf{X}$, and the ridge solution, expressed as a function of the bias parameter k added to the diagonal elements of $\mathbf{X}'\mathbf{X}$. In [18] it is emphasized that for this purpose the assigned rank is best defined as a piecewise continuous variable.

Analysis of the Acetylene Data

We return now to the Acetylene Data example. The standardized full quadratic model is

$$E(y) = \beta_0 + \sum_{j=1}^{3}\beta_j x_j + \sum_{1\le j<j'}^{3}\beta_{jj'}x_j x_{j'} + \sum_{j=1}^{3}\beta_{jj}x_j^2$$

where y = % conversion, x_1 = (Temperature − 1212.50)/80.623, x_2 = [H_2/(n-Heptane) − 12.44]/5.662, and x_3 = (Contact time − 0.0403)/0.03164.

We recommend that a polynomial model be reported to (and used by) the ultimate consumer in this form. Note that each predictor variable is standardized, but the expansion terms (squares and cross-products) are created directly from the standardized linear terms. The model is not standardized with respect to y. Numerical evaluation in this form is accurate, and interpretation of the coefficients is straightforward.

However, for selection of the amount of bias it is necessary to examine the equation and its fit with all variables scaled to correlation form, including y, the linear predictors, and the expansion predictor variables. We refer to the regression coefficients so obtained as the "correlation basis" regression coefficients. In regular use of ridge regression we display the correlation basis regression coefficients in tables and/or graphs for about 25 values of k, spaced approximately logarithmically over the interval [0, 1]. Table 2 shows the correlation-basis regression coefficients by least squares, and by ridge with two values of k. Both $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$ are in

Table 3

ACETYLENE DATA REGRESSION RESULTS

Five-Coefficient Reduced Quadratic Model

	VIF	Correlation Basis Coefficient			
			Ridge		
	Least Squares	Least Squares	k =.01	k =.05	Generalized Inverse (r = 4.0)
x_1 = Temperature	43.11	.602	.557	.514	.518
x_2 = H_2 / n-Heptane	1.07	.194	.192	.187	.193
x_3 = Contact Time	53.52	-.323	-.368	-.391	-.417
x_1x_2	1.09	-.273	-.270	-.258	-.272
x_1^2	4.68	.173	.180	.169	.197
Maximum VIF		53.52	13.63	1.72	1.15
R_A^2		.991	.990	.989	

correlation form. Hence, the correlation-basis coefficient b is the expected change in y, measured in y-standard deviations, given an increase in x of one standard deviation (i.e. $b = 0.5$ implies that y increases 0.5 standard deviations when x is increased one standard deviation). In this problem the coefficients have stabilized by $k = .05$ (see the section on ridge trace interpretation), and the variance inflation factor is also reasonable there. Note that the large least squares coefficients for the x_1x_3 interaction and for x_3^2 have all but disappeared in the ridge model. Also, x_2x_3 and x_2^2 have small coefficients. Similar results are obtained with the generalized inverse model, shown for $r = 3.8$, for which the regression vector length is about the same as the ridge model. Suppose these four terms are eliminated. Variable selection is a safer strategy here, since the bias has removed most of the ill conditioning. For a nearly orthogonal model in the correlation-basis scaling, all coefficients will have nearly equal variances; hence, variable selection can be made on the basis of absolute values of the coefficients. The mild bias component of the mean square errors of the coefficients is ignored for this purpose. Table 3 shows that further biasing of the five-term model doesn't change the coefficients much.

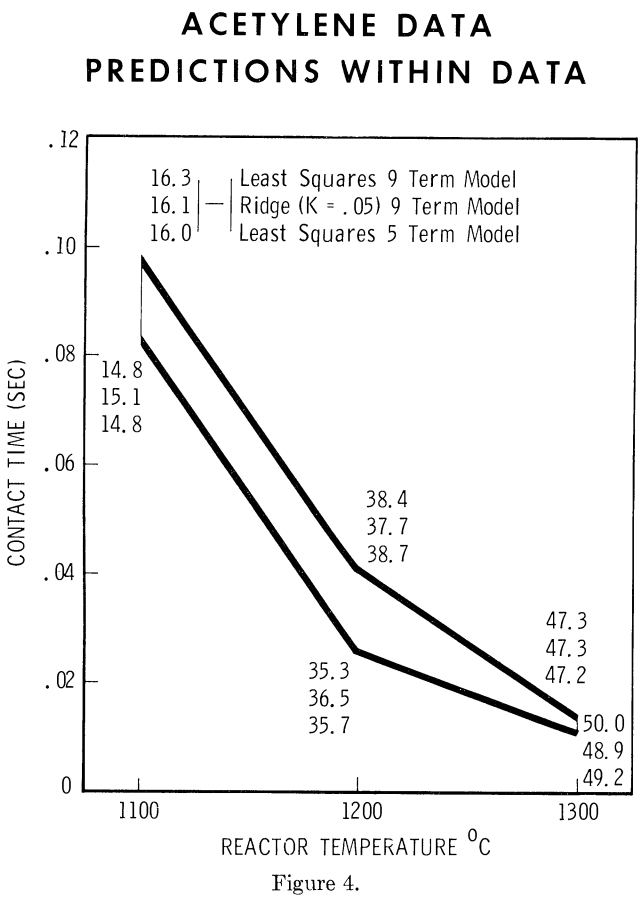

Figure 4.

Figure 4 shows the predictions for the least squares nine-term full quadratic, the ridge nine-term model, and the least squares five-term model. The prediction points shown are the extreme points of the data; they define the boundary of the region covered by the data. For practical purposes all three models predict equally well here.

Figure 5 shows predictions outside the data. Notice that we are not extrapolating beyond the ranges covered by the predictor variables individually. We are only extrapolating to the corners of the region. Consider the upper right corner. The least squares nine-term model predicts minus 86.2% conversion. This is physically impossible. The other models predict 32.9% and 38.0% conversion, a much more realistic prediction. A similar situation holds at the lower left corner.

What do we conclude from this example? Well, first of all, we conclude that the nine-term full quadratic given by least squares is not a good model, even though it is the one that fits the estimation data most closely.

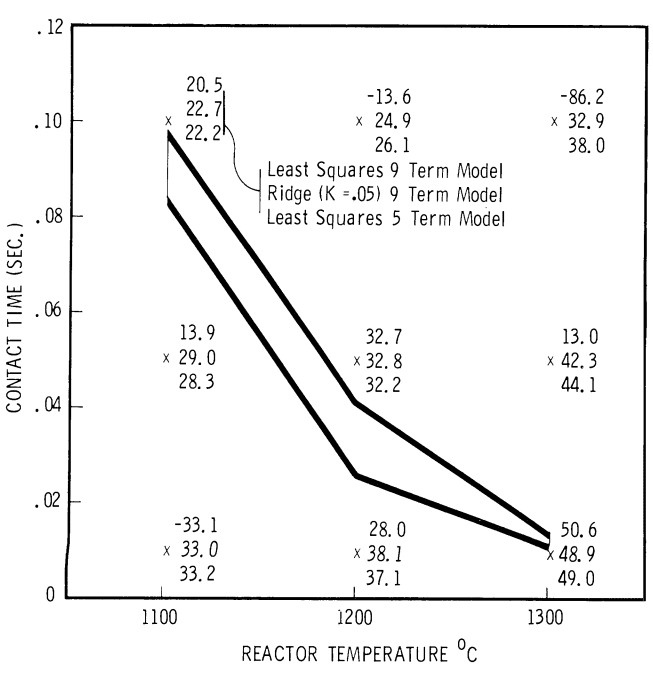

Figure 5.

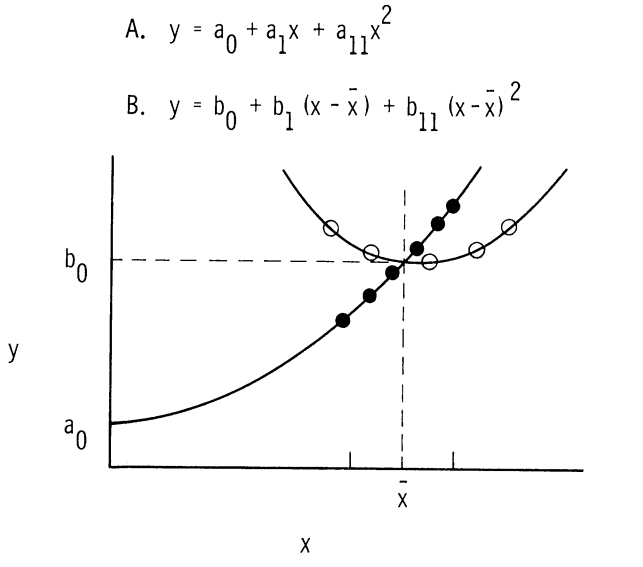

Figure 6.

Secondly, we find that biased regression, routinely applied, produces a much better model, still using all nine terms. Finally, we have been guided by the ridge results to a subset model that also does a good job. In this case, Snee [22] arrived at the same subset model by plotting the raw data.

When Is Variable Selection A Good Strategy?

At this point, it may be appropriate to put forth the conditions under which variable selection is, or is not, a good strategy. Variable selection is a good strategy when the candidate variables already are mutually orthogonal, or nearly so. It is also a good strategy when the variables have been made effectively orthogonal, or nearly so, by introducing bias in the estimator. The only other condition is when the selection of the candidate subsets can be strongly guided by knowledge of the background of the problem and the definitions/properties of the variables. This means information external to the numerical values of the data.

Variable selection is a poor strategy when the candidate variables are highly correlated. In essence, the extreme variance inflation completely unstabilizes all the criteria one can compute from the least squares estimates, leading to highly unstable subset selection. It is also a poor stragegy when the candidate variables include curvilinear effects (for example, squares) of other candidate variables. This is illustrated in Figure 6.

Consider the two candidate models A and B. Both can represent the quadratic curve that passes smoothly through the solid dot data, with $y = a_0$ at $x = 0$. Both models can also represent the open circle data, with minimum at $x = \bar{x}$, where $y = b_0$. Now, consider what happens if the linear terms are dropped or not selected by the subset procedure. Then if the data are like the open circles, model A is a total disaster, while model B does a great job. The converse holds for data like the solid dots. Thus, the operational behavior of all subset-selection procedures is not invariant with the arbitrary choice of computing origin.

Now, in most situations where quadratic or higher polynomials are applied, the model is really functioning like a series expansion of some function in the region of the data. There are normally only two appropriate points in the predictor variable space about which the expansion could be natural. These are the mean point of the predictor data and the origin of the predictor data space. This implies that any variable selection procedure should be done at least twice with curvilinear models, once computing about the mean point, and once computing about the origin in order to find out which, if either, gives a simple, well behaved model.

A Simulation Experiment

The final example of Part I is a synthetic one. It illustrates how both ridge and generalized inverse estimators do better than either least squares or any subset. It also illustrates the mechanics of using the biased estimation procedures.

For this "three-predictor example", the data structure is as shown in Figure 7. There are eight estimation data points shown by the open circles. If the parameter α is zero, the design is a classical 2^3 factorial. As α approaches 1.0, four of the points move, in the directions of the arrows. The correlation of x_1 and x_2 be-

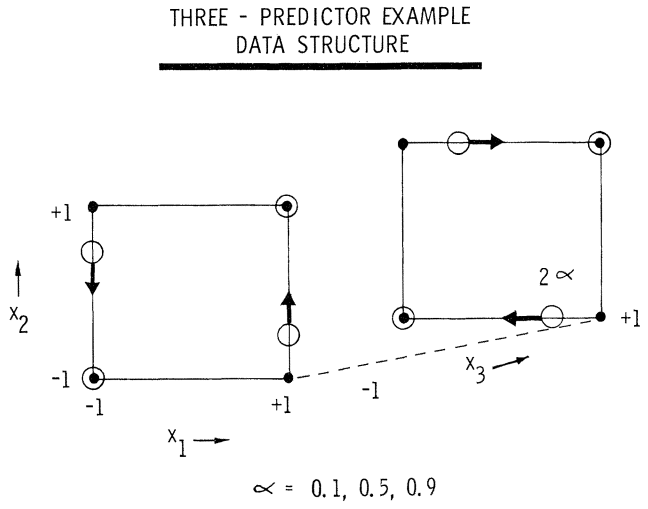

Figure 7.

comes progressively larger, with unit correlation as the limit.

We generated data at three values of α. Note that x_3 is orthogonal to x_1 and x_2 for all values of α. The solid dots are prediction points. The true model is $E(y) = x_1 + x_2 + x_3$; that is, all coefficients are unity, and the regression constant is zero. We have introduced additive errors selected randomly from a standard normal distribution with mean zero and standard deviation $\sigma = 0.8$.

The actual data are as follows:

i	x_1	x_2	x_3	$E(y)$	ϵ_i
1	-1	-1	-1	-3	-0.305
2	1	1	-1	1	-0.321
3	-1	-1	1	-1	1.900
4	1	1	1	3	-0.778
5	-1	$(1 - 2\alpha)$	-1	$-1 - 2\alpha$	0.617
6	1	$-(1 - 2\alpha)$	-1	$-1 + 2\alpha$	-1.430
7	$-(1 - 2\alpha)$	1	1	$1 + 2\alpha$	0.267
8	$(1 - 2\alpha)$	-1	1	$1 - 2\alpha$	0.978

The correlation between x_1 and x_2 is $r_{12} = \alpha/(1 - \alpha + \alpha^2)$. Thus $r_{12} = 0.110, 0.667, 0.989$ for $\alpha = 0.1, 0.5, 0.9$ respectively.

The regression model is $\bar{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$. In all cases we included a constant term on the hypothetical assumption that the data analyst does not know in advance that the true value of the regression constant is zero. The criterion by which we judge the quality of our regression models is the prediction standard error at the eight corners of the cube.

Table 4 shows the ridge regression results for $\sigma = 0.8$ for several values of k and for three values of α. The quantities tabulated are

S_e = residual standard error from the estimation data

R_A^2 = adjusted $R^2 = 1 - S_e^2/S_y^2$, where

S_y^2 = variance of y_i

$y_i = E(y_i) + \epsilon_i$

VIF = maximum variance inflation factor

S_p = prediction standard deviation at the eight prediction points

$$= \left(\sum_{i=1}^{8} [\hat{y}_i - E(y_i)]^2/8 \right)^{1/2}$$

Let us start by examining results for $\alpha = 0.9$. Note how the estimation residual error, which is 0.537 when $k = 0$, increases as the bias k is increased. As a consequence, the adjusted R^2 decreases. However, note the reduction in the variance inflation as k is increased. Finally, the proof of the pudding is the composite effect of these two opposing trends. This is seen in the prediction residual error S_p, which goes through a minimum at $k = 0.2$. This tells us that a ridge estimator with bias of $k = 0.2$ gives predictions at the corners of the cube with a standard error only 0.536 (circled in the table) compared with the least squares prediction residual error $S_p = 1.972$. Note that similar but less dramatic results occur for the smaller values of α.

At the bottom of the table is an extra line of numbers. These are the variance inflation factors for orthogonal predictors, as a function of k. Note that the minimum prediction residual error here occurs for values of k at or just beyond the value where the maximum variance inflation factor is about the same size as if the factors were orthogonal.

Table 5 shows the corresponding results using generalized inverse regression. Again focusing first on $\alpha = 0.9$, note how the regression residual gets larger as the assigned rank of the $\mathbf{X}'\mathbf{X}$ matrix is decreased. Again the adjusted R^2 decreases, and so does the maximum variance inflation factor. Again the prediction residual error goes through a minimum. The minimum occurs at assigned rank 1.5, which, note, is not an integer. In this example the minimum prediction residual by generalized inverse is smaller even than by ridge regression. We do not interpret this as a general result, but only as an indication that either type of biased estimator can give results substantially better than least squares.

Finally, for completeness, we show in Table 6 the corresponding results for all possible subset models. Focusing again on $\alpha = 0.9$, we note that the prediction residual error varies between 1.15 minimum and 2.18 maximum. The two best subset models are, not unexpectedly, the ones involving x_1 with x_3 and x_2 with x_3. Both do better than the full least squares model whose residual prediction error was 1.972; but all of these least squares models are much poorer than the ridge or generalized inverse models.

Model Validation

After we have developed a prediction equation, it is imperative that a measure of the accuracy of the model coefficients and predictions be obtained. One way to

Table 4

THREE-PREDICTOR EXAMPLE (σ = 0.80)

α		Ridge k: 0	0.1	0.2	0.4	0.8
0.1	S_e	.729	.759	.828	.995	1.286
	R^2_A	.856	.843	.813	.730	.550
	VIF	1.012	.833	.698	.511	.309
	S_p	.609	(.598)	.619	.702	.878
0.5	S_e	.591	.626	.701	.876	1.173
	R^2_A	.906	.895	.868	.794	.630
	VIF	1.800	1.155	.825	.510	.309
	S_p	.713	.651	(.636)	.674	.817
0.9	S_e	.537	.621	.699	.878	1.189
	R^2_A	.936	.914	.891	.829	.685
	VIF	45.751	.826	.694	.510	.309
	S_p	1.972	.560	(.536)	.583	.738
$(VIF)_o$		1.00	.826	.694	.510	.309

accomplish this is a study of the physical nature and theoretical basis of the system being studied. For example, in the acetylene data discussed earlier, negative percent conversion was predicted in some parts of the factor space by the 10-coefficient least squares quadratic model. Negative conversion is physically impossible and clearly indicates the associated model does not give an accurate description of the system which generated the data.

Another method of model validation is to collect additional data and see how well the model predicts the new data. This often is not possible. One way to simulate the collection of new data is to split the data in hand into two subsets. One subset, called the "estimation data", is used to estimate the coefficients in the model. The remaining subset, called the "prediction data", is used to measure the prediction accuracy of the model. When data are ordered with respect to time, some point in time can be used to split the data. For example, the Laird and Cady corn yield data [14], to be discussed later, were collected over a four-year period. Laird and Cady used the first three years of data as the estimation data and the fourth year as the prediction data. Kennard and Stone's CADEX Procedure [12] is another way of splitting the data and will be used in the analysis of the GC-ASTM Data.

Still another validation method is comparison with a theoretical model. This approach also is used with the GC-ASTM Data.

II. Use of Biased Estimation in Data Analysis

In the first part of this paper we reviewed the theory underlying ridge regression and biased estimation, presented the results of a little simulation experiment,

Table 5

THREE-PREDICTOR EXAMPLE (σ = 0.80)

α		Generalized Inverse r				
		3.0	2.5	2.0	1.5	1.0
0.1	S_e	.729	.741	.776	1.247	2.101
	R^2_A	.856	.851	.836	.577	.000
	VIF	1.012	1.000	1.000	.500	.450
	S_p	.609	.581	.571	.527	1.087
0.5	S_e	.591	.606	.649	1.172	2.057
	R^2_A	.906	.901	.887	.630	.000
	VIF	1.800	1.050	1.000	.500	.300
	S_p	.713	.634	.605	.563	1.105
0.9	S_e	.537	.553	.599	1.146	2.042
	R^2_A	.936	.932	.920	.708	.072
	VIF	45.751	23.001	1.000	.500	.251
	S_p	1.972	1.100	.563	.518	1.083

and illustrated the use of "Ridge" in developing a model for the acetylene data. The remainder of this paper will describe two data sets which illustrate the use of biased estimation in problems with a large number of variables.

Interpreting the Ridge Trace

When the predictor variable correlation matrix contains several large correlation coefficients, it is difficult to untangle the relationships among the predictor variables by inspection of the simple correlation coefficients. Some automatic procedures, such as stepwise, best subsets, and PRESS regression, attempt to untangle the variables by selecting some "best" subset of the predictors. However, these methods do not really give an insight into the structure of the factor space and the sensitivity of the results to the particular set of data at hand. In the Gorman and Toman ten-factor problem, Hoerl and Kennard [11] showed that a "best subsets" procedure does not necessarily reduce predictor variable correlations. The correlations may be greater than those among the original variables.

One of the big advantages of ridge regression is that a graphical display, called the "Ridge Trace", can help the analyst to see which coefficients are sensitive to the data. Thus, sensitivity analysis is an aim of ridge regression. The ridge trace is a plot of the value of each coefficient versus k. The trace will have one curve, or trace, per coefficient. For clarity we recommend that not more than ten coefficient traces be plotted on a given graph. It was noted earlier that the variance of a coefficient is a decreasing function of k and the bias is an increasing function of k. Thus, as k increases, the coefficient mean square error (variance plus squared bias) decreases to a minimum and then increases. The objective is to find a value of k which gives a set of coefficients with smaller MSE than the least squares solution. Of course, as k increases, the residual sum of squares will also increase. This should not be of great concern, because the objective is not to obtain the closest possible fit to the estimation data, but to develop a "stable" set of coefficients which will do a good job of predicting future observations. By stable we mean

Table 6

THREE-PREDICTOR EXAMPLE ($\sigma = 0.80$)

α		Subset Model					
		x_1	x_2	x_3	x_1, x_2	x_1, x_3	x_2, x_3
0.1	S_e	1.942	1.807	1.320	1.864	1.214	.934
	R^2	.000	.110	.525	.054	.598	.763
	VIF	1.00	1.00	1.00	1.01	1.00	1.00
	S_p	1.46	1.42	1.47	1.11	1.13	1.08
0.5	S_e	1.807	1.693	1.340	1.825	.933	.624
	R^2_A	.121	.229	.517	.104	.766	.895
	VIF	1.000	1.000	1.000	1.80	1.00	1.00
	S_p	1.42	1.43	1.47	1.17	1.07	1.09
0.9	S_e	1.687	1.656	1.643	1.811	.603	.492
	R^2_A	.367	.389	.399	.270	.919	.946
	VIF	1.00	1.00	1.00	45.75	1.00	1.00
	S_p	1.47	1.48	1.47	2.18	1.15	1.16

that the coefficients are not sensitive to small changes in the estimation data. If the predictor variables are highly correlated, the coefficients will change rapidly for small values of k and gradually stabilize (change little) at larger values of k. The value of k at which the coefficients have stabilized gives the desired set of coefficients. If the predictor variables are orthogonal, then the coefficients would change very little (i.e. the coefficients are already stable) indicating the least squares solution is a good set of coefficients.

Many statisticians have expressed concern about the selection of k. It is the authors' experience that this is not a problem in practice. As will be pointed out later in the examples, the plot of prediction standard deviation of new data versus k usually has a flat minimum; hence, there is a range of k-values which give equivalent results from a practical point of view. We have also observed that the ridge trace is easy for experimenters to interpret. They respond quickly to graphical output, and after observing two or three examples they can usually interpret the trace readily.

The k selected via the ridge trace is, technically speaking, a random variable. Even though that fact is not a practical concern in selecting the regression estimates, it does complicate the theory of confidence limits and hypothesis tests, due to the introduction of bias. Because of the bias, the mean square errors are all mildly dependent upon the true coefficient vector β, which is unknown. This is a fertile area for continued research.

On the practical side it should be noted that models with no constant term ($\beta_0 = 0$) typically require a smaller value of k (often $\leq .01$) than models with a constant term ($\beta_0 \neq 0$). Also, models with low R_A^2 statistics usually require larger values of k than models with high R_A^2. Increasing k indefinitely will ultimately drive all coefficients to zero, but for smaller values of k it is not uncommon to see a coefficient (perhaps after an initial sign change) increase in absolute value as k increases. In this situation we have often found that good results are obtained by using a value of k about where the coefficient passes through the maximum absolute value. This procedure was used to select the value of k for the GC-ASTM model to be discussed later.

Plots of the generalized inverse coefficients can be

Table 7

LAIRD AND CADY CORN YIELD DATA
33-TERM MODEL COEFFICIENTS

Rank	Variable	Ridge Regression (k = .3)	Generalized Inverse (r = 9.5)	Rank	Variable	Ridge Regression	Generalized Inverse
1	*N	.249	.179	18	A^2	.035	.040
2	BN	.188	.166	19	JN	.033	-.009
3	AN	.182	.181	20	BL	.031	.000
4	*J	-.119	-.113	21	AL	.027	.017
5	*AC	-.112	-.096	22	A	.026	.042
6	AJ	-.101	-.108	23	G^2	.025	.039
7	AH	-.099	-.069	24	BH	-.017	-.078
8	*DN	-.096	.028	25	HN	.017	.008
9	BJ	-.091	-.109	26	LN	.015	.086
10	$*N^2$	.091	.175	27	$*B^2$	-.014	.011
11	C	-.082	-.095	28	D	-.013	-.062
12	*H	-.074	-.069	29	CN	-.007	-.006
13	*F	-.072	-.066	30	L	-.005	.006
14	BD	-.069	-.057	31	AD	-.005	-.053
15	*AB	.056	.050	32	E	.003	.054
16	BC	-.050	-.088	33	B	.002	.013
17	G	.043	.036	Vector Length		.239	.236

* Variables selected by PRESS

constructed and interpreted in the same manner as the ridge trace by using the assigned rank r as the abscissa of the trace in place of k.

Laird and Cady Corn Yield Data

Our first example of a large data set is the corn yield data published by Laird and Cady [14]. Cady and Allen [3] later used these data to illustrate the PRESS procedure. The response is the corn yield, in tons per hectar, measured at each of four applied nitrogen levels in each of 72 experimental sites. This resulted in 288 data points over a four-year period.

There are 11 predictor variables—

Applied Nitrogen	N	Soil Slope	F
Soil Nitrogen	A	Soil Texture	G
Previous Crop	B	Hail	H
Excess Moisture	C	Blight	J
Drought	D	Weeds	L
Rooting Zone Depth	E		

Four of these are measured (N, A, E, F); one is expressed as an index (D), and the remaining variables have been assigned a value on a subjective scale. The nonnitrogen variables will be referred to as "site variables".

The model used by Cady and Allen to describe these data was a subset of the full quadratic containing a constant term, 11 linear terms, 18 cross-product or interaction terms, and 4 squared terms for a total of 33 terms (Table 7). We will note here for later reference that 13 out of the 18 interaction terms involve the two nitrogen variables (N and A).

Cady and Allen chose to divide the data into two sets. The 228 data points collected in the first three years were used to estimate the coefficients in the model. The 60 observations obtained in the fourth year are used to test the predictability of the model. These data sets will be called the "estimation data" ($n = 228$) and "prediction data" ($n = 60$), respectively.

Cady and Allen found the residual standard deviation of the fits of the full, stepwise, and PRESS models to be 0.59, 0.62, and 0.65 for the estimation data and 1.03, 0.84, 0.72 for the prediction data. The PRESS model did the best job of predicting. This might be expected since the variables in the full and stepwise models are highly correlated, with maximum VIF of 180 and 122, respectively, while the variables in the PRESS model are less correlated, with maximum VIF 12.

What does ridge do with this problem? We found that the ridge trace computed from the first three years of data stabilized around $k = 0.3$.

The prediction standard deviation for the fourth year of data is plotted in Figure 8 versus the value of k used in developing the ridge model. As shown there, the prediction standard deviation decreases as k increases, reaching a very flat minimum of 0.71 at $k = 0.6$. A bias of only $k = 0.01$ reduced the prediction standard deviation from 1.03 to 0.82. At $k = 0.3$, which we selected from the ridge trace, the prediction standard deviation is 0.72, which is identical to the prediction

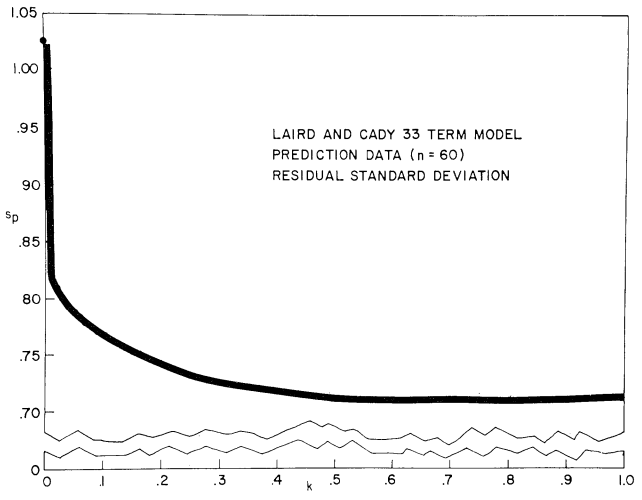

Figure 8.

standard deviation of the PRESS model. The $k = 0.3$ ridge coefficients ranked in order of absolute value are shown in Table 7.

The three largest coefficients are N, BN and AN, all involving applied nitrogen (N). The nine terms in the PRESS model, denoted by an (*), rank from the largest to 27th out of 33—8 being in the top 15 and the ninth, B^2, ranking 27th. It is even more interesting that out of the top 15 ridge coefficients, nine are interactions and 9 out of 15 involve either applied nitrogen (N) or soil nitrogen (A).

What is this telling us? Applied nitrogen and soil nitrogen ("nitrogen" for short) *is the dominant variable.* When nitrogen is absent (i.e. = 0) there cannot be any corn yield. This would suggest a zero intercept with respect to nitrogen. However, the other variables do not have a zero intercept. From this we conjectured that a multiplicative model would be most natural for these data. This is consistent with the dominance of the many interaction terms involving nitrogen in the 33-term model. Our postulated multiplicative model is

$$E(y) = (\text{applied nitrogen} + \text{soil nitrogen}) \times (\text{site variables})$$
$$= Z_1 Z_2,$$

where $Z_1 = (N + \beta_2 A) + \beta_{12}(N + \beta_2 A)^2$,

and $Z_2 = \beta_1 + \beta_3 B + \beta_4 C + \cdots + \beta_{11} L$.

Schematically, corn yield is modeled as a product of two factors. The first is applied nitrogen plus soil nitrogen, which we denote by Z_1. The second factor contains the site variables, denoted by Z_2. Factor Z_1 is a quadratic function of $(N + \beta_2 A)$ and Z_2 is a linear function of the nonnitrogen site variables. The β_2 coefficient is needed because applied nitrogen (N) and soil nitrogen (A) are measured in different units. The model contains twelve coefficients, as compared to ten coefficients in the PRESS model.

The ten coefficient PRESS model fit the estimation and prediction data with residual standard deviations of 0.65 and 0.72 respectively (Table 8). The twelve coefficient multiplicative model, fitted by nonlinear least squares, [17], had residual standard deviations of 0.72 and 0.75 for the estimation and prediction data which are in the same ballpark as the PRESS model.

An examination of the coefficient confidence limits suggested the three coefficients corresponding to previous crop (B), root depth (E), and weeds (L) were not significant. When these variables were deleted, the resulting nine-coefficient multiplicative model fit the estimation and prediction data with residual standard deviations of 0.73 and 0.64, respectively, giving a somewhat better fit to the prediction data than the PRESS model (Table 8).

There are other differences between the PRESS and multiplicative models which should be noted:

(i) The PRESS model contains a constant term—the multiplicative model does not.

(ii) Previous crop (B) is included in the PRESS model but not in the multiplicative, and

Table 8

LAIRD AND CADY CORN YIELD DATA

FIT OF PRESS AND MULTIPLICATIVE MODELS

Model	Number of Coefficients	Residual Standard Deviation Estimation (n = 228)	Residual Standard Deviation Pred. (n = 60)
PRESS	10	.65	.72
Multiplicative	12	.72	.75
Multiplicative	9	.73	.64

Table 9

MODEL FOR 158° ASTM DATA

GC CUT	TEMPERATURE RANGE °F	COEFFICIENTS ESTIMATED FROM DATA: LEAST SQUARES	RIDGE (k = .006)	GENERALIZED INVERSE (r = 7)	THEORETICAL DISTILLATION LEAST SQUARES COEFFICIENTS
1	0 - 15	224	126	120	125
2	15 - 40	87	94	105	132
3	40 - 87	110	104	109	102
4	87 - 126	116	75	74	84
5	126 - 145	-92	45	46	42
6	145 - 175	80	26	45	20
7	175 - 198	96	38	21	0
8	198 - 220	54	14	-8	-9
9	220 - 237	-125	-62	-45	-15
10	237 - 285	-30	-15	-20	-21
11	285 - 300	-65	-19	1	-29
12	300 - 333	21	6	-4	-25
13	333 - 376	181	36	5	-25
14	376 - 414	-217	25	24	-25
15	414 - 487	22	-40	-10	-24
Vector Length			.36	.37	
Prediction Data Std. Dev.		1.28	1.01	.96	

(iii) Soil texture (G) is included in the multiplicative model and not in the PRESS model.

To summarize this problem, we feel that by keeping all the terms in the model and reducing the variable correlations with Ridge we were able to

(i) Obtain a model which predicts well, and
(ii) Learn more about the roles of all the variables in the model.

In this case the ridge results suggested a nonlinear alternative model. While this may not be the ultimate model, it is consistent with the physical background of the problem as described in the Laird and Cady paper, and gives the scientist a different way to think about the mechanism under study.

GC-ASTM Model

The second example concerns the relationship between the ASTM and gas chromatograph, or GC for short, distillation of a gasoline sample. One of the properties which determines the quality of a gasoline is volatility as measured by the percent of the blend evaporated at various temperatures (°F). The standard method of measuring volatility is an ASTM distillation, in which the gasoline sample is heated and the vapors pass through an ice bath and condense. The cumulative percent evaporated at various temperatures is recorded. In the ASTM distillation some of the higher boiling components "hold back" the lower boiling components. The GC distillation is much more accurate and each component "comes off" at its true boiling point.

ASTM and GC distillation curves are not identical, and gasoline volatility specifications are written in terms of ASTM. In order for a refinery to use the GC for on-line control of volatility, a model is needed to predict the ASTM distillation of a blend from a GC distillation of the blend. This example is typical of those situations where a property of a material is measured by a series of points which form a curve. It is important to use a sufficient number of points to describe the curve, but one should be careful not to overdefine the curve. The use of too many points results in redundant information. In this example, the points on the GC curve were selected by the engineer to give a proper description of the curve. The points on the ASTM curve to be predicted corresponded to specifications of various gasolines.

The GC temperature range was divided into fifteen cuts: 0-15, 15-40, . . . , 414-478°F (Table 9). The fifteen predictor variables in the model, $x_1, \ldots, x_{15}$, are the volume fraction of the blend evaporated in each of the cuts. The responses to be predicted, $y_1, \ldots, y_{14}$, are the cumulative percent of the blend evaporated at each of the 14 ASTM temperatures. While a model was developed for each of the 14 ASTM temperatures, we will concentrate our attention on the three most important specifications: y_4 = ASTM 158, y_6 = ASTM 212, and y_{10} = ASTM 302.

The postulated model is

$$E(y) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{15} x_{15},$$

where y is the cumulative percent evaporated at a given ASTM temperature and x_j is the fraction evaporated in the j^{th} GC cut. The constant term (β_0) has been deleted because the x_j's sum to 1.0, and produce a singular $\mathbf{X}'\mathbf{X}$ matrix if β_0 were included.

The main uses of the model would be prediction of gasoline blend ASTM distillations, for which the prediction standard deviation had to be $\leq 1.5\%$, and as input to linear programming calculations to determine optimum volatility blending procedures. Hence, it was imperative that the ASTM predictions also be responsive to changes to the GC curve in any temperature range, and that the estimated coefficients in the model be "realistic" in light of the available engineering knowledge. Part of the prior history on this problem was that coefficients developed by least squares and stepwise regression were unacceptable from a physical viewpoint.

GC and ASTM distillation data were available on 59 blends. It was felt that it would be advantageous to have an independent estimate of the model prediction standard deviation. We used Kennard and Stone's CADEX algorithm [12] to split the data into two sets—29 estimation blends and 30 prediction blends. While we do not claim that this is necessarily the best way to split the data, we are describing the analysis the way it was actually conducted. As in the Laird and Cady corn yield example, the coefficients will be estimated from the 29 estimation blends. The 30 prediction blends will be used to obtain a measure of the model prediction standard deviation.

The ridge trace for y_4 = ASTM 158 is shown in Figure 9 where the x axis is $k \times 10^{-1}$. The traces for the first 10 coefficients are shown on the top and the traces

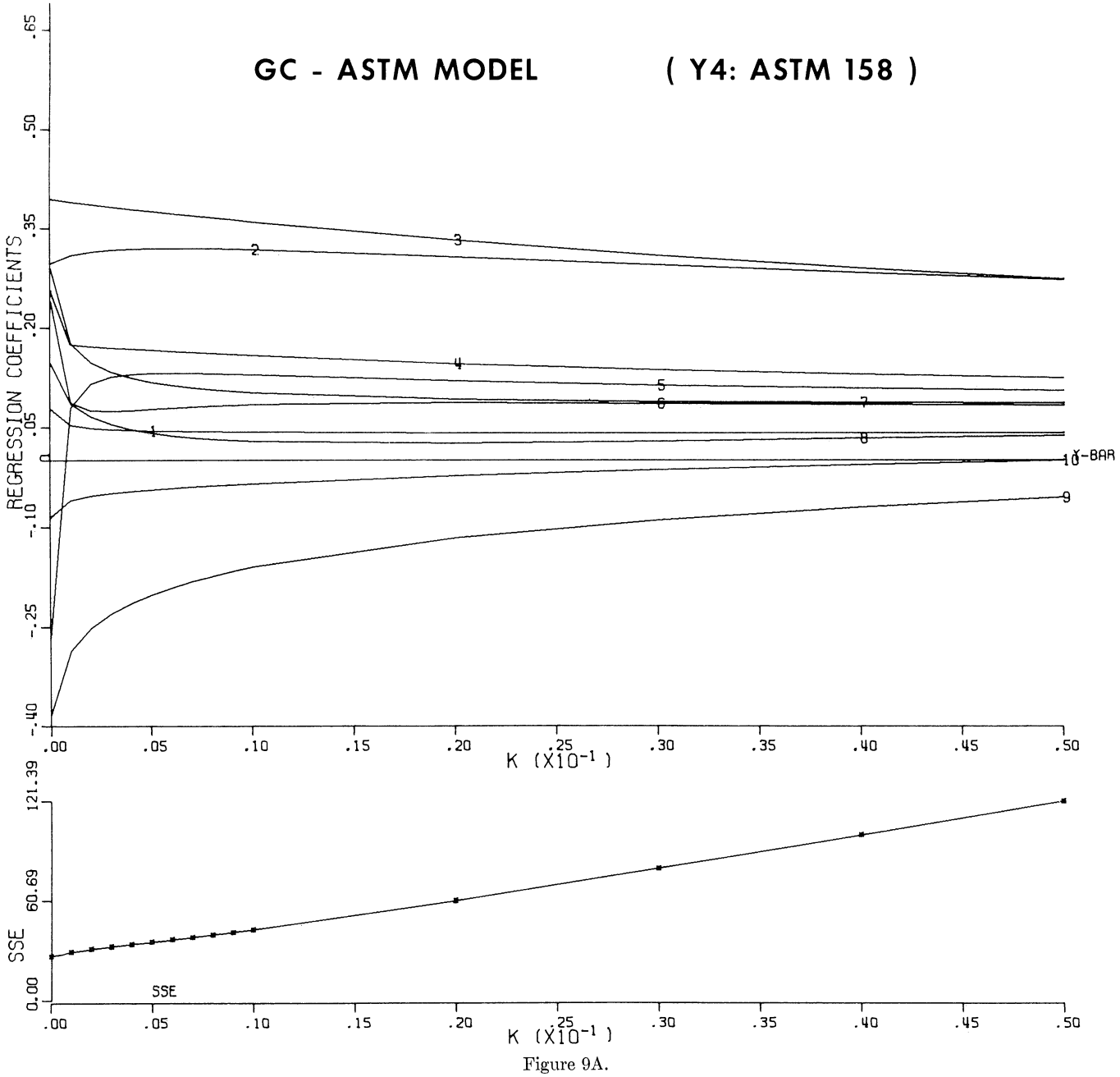

Figure 9A.

for the last five coefficients are shown on the bottom. The system stabilizes quickly around $k = .005$ or .01. We decided to use the coefficients at $k = .006$. Before studying these coefficients further, let us look at a plot of the prediction standard deviation of the 30 prediction blends versus k shown in Figure 10.

For y_4 = ASTM 158, the prediction standard deviation has a value of 1.28 at the least squares solution, decreases as k increases, and reaches a minimum near $k = .006$, the value of k we chose from the ridge trace! The ASTM 212 curve follows a similar pattern, reaching a minimum around $k = .003$. At the 302 and 375 temperatures the experimental error is smaller. The prediction standard deviation increases as k increases, although the ASTM 375 curve is flat until $k = .006$. We now turn our attention to the coefficients in the ASTM 158 model shown in Table 9.

First, note the standard error of prediction at the bottom of the columns. As we noted previously, the ridge regression model is a better predictor. It has a standard error of 1.01% compared to a standard error of 1.28% for the least squares model. Again, these two numbers were computed from the 30 blends not used in computing the coefficients. An examination of the models reveals that the coefficients in the least squares model are in general larger than the coefficients in the ridge regression model. In the least squares model, the largest coefficients are around 200 in absolute value (in cuts 1, 13 and 14), while the largest coefficients in the ridge regression model are around 100 in absolute value (in cuts 1 and 3). In addition, the least squares coefficients are not well behaved with respect to sign. These points can be seen easily when the models are compared graphically.

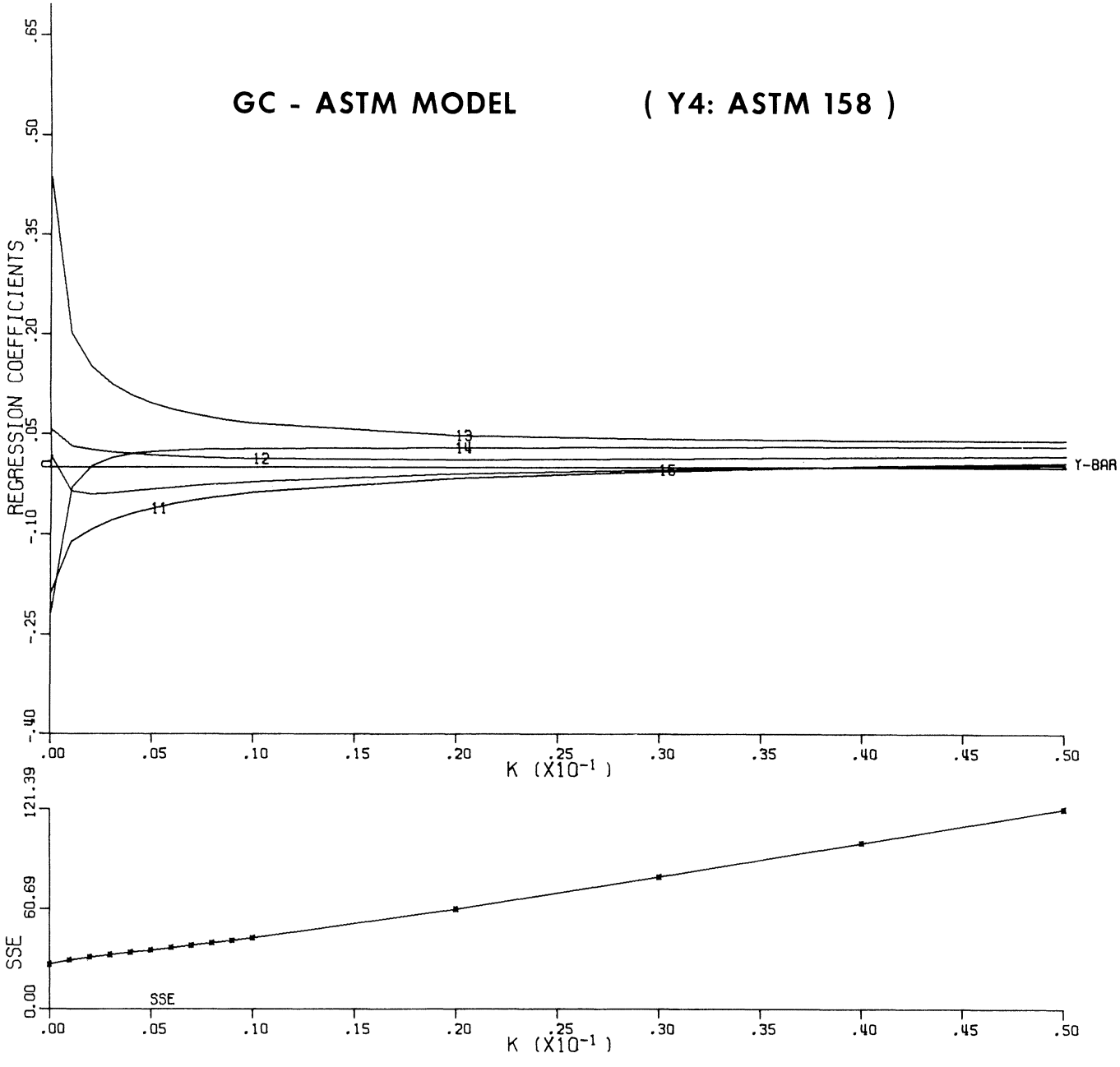

Figure 9B.

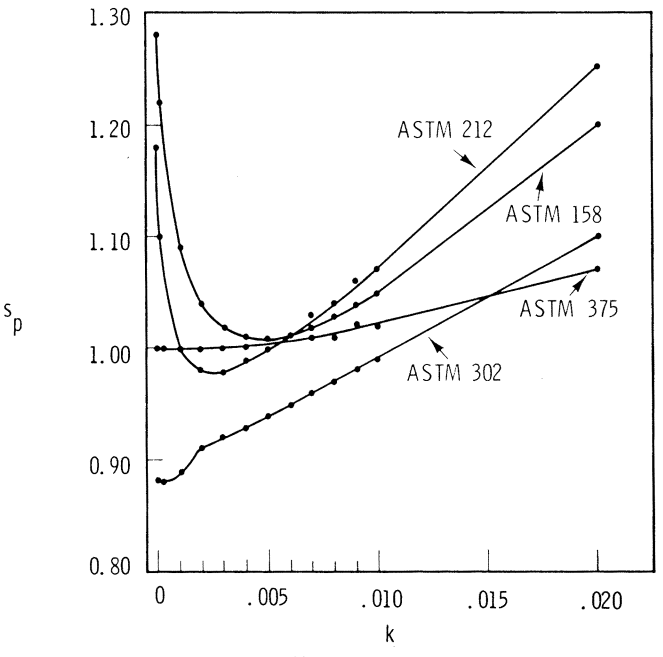

Figure 10.

In Figure 11 the coefficients are plotted versus GC cut temperature. The top graph shows the least squares coefficients, and the second graph shows the ridge regression coefficients. Note that coefficients 5, 14 and 15 have changed sign and coefficients 1, 4, 9, and 13 are considerably smaller.

Previous knowledge indicated that all the coefficients at the higher temperatures should be negative. To gain greater insight into this problem, we designed a theoretical distillation experiment centered around the 59 blends. The amounts of each of the 15 GC cuts were varied according to a pseudocomponent simplex design for mixtures comprised of 15 pure component and 105 binary blends. The distillations for these 120 blends were calculated using Raoult's Law with activity coefficients of unity and atmospheric pressure. The 15 GC cuts were treated as pure hydrocarbons, having true boiling points at the midpoint of the cut. This theoretical distllation is similar to the ASTM distillation and can provide corollary information concerning the relationship between the model coefficients and GC temperature. The models developed for the theoretical distillation data confirmed our previous theories. At the bottom of Figure 11 we see that the theoretical distillation coefficients decrease in size with increasing temperature and finally go negative at the higher temperatures. It is immediately obvious from this graph that the ridge coefficients bear a closer resemblance to the theoretical distillation coefficients than the least squares coefficients. The ridge regression coefficients and theoretical distillation coefficients follow a smooth pattern as the GC temperature increases; however, the least squares coefficients do not follow this nice relationship.

Ridge regression gave equally meaningful coefficients in the models at the other ASTM temperatures as evidenced by the coefficients in the ASTM 212, 302, and 375°F models shown in Figure 12. As the scientific background of the problem indicated, the number of large positive coefficients in the model increases as the ASTM temperature increases.

We can summarize this example by saying that the project goals were met and on-line process control of volatility went into operation on schedule. Both the ridge model and the theoretical distillation model worked in practice, whereas the least-squares and stepwise regression models had not.

Generalized Inverse Results

In our discussion of the Laird and Cady Corn Yield Data and the GC-ASTM Distillation Model, we have focused on the ridge estimators. The corresponding results for the generalized inverse estimators are shown similtaneously in Tables 7 and 9. The generalized inverse results shown correspond to a value of r for which the regression vector length is approximately equal to the length for the selected ridge bias k. Detailed ex-

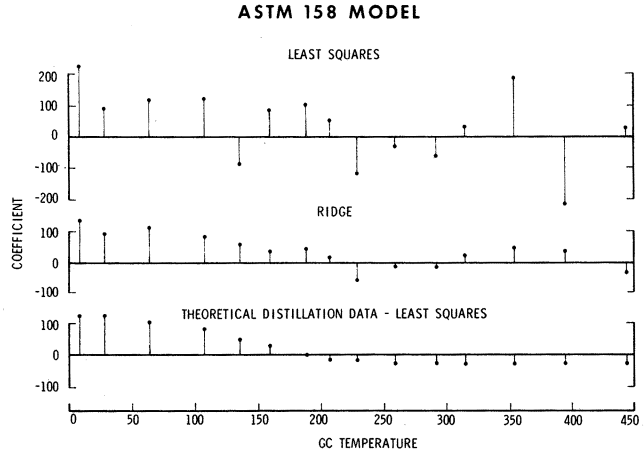

Figure 11.

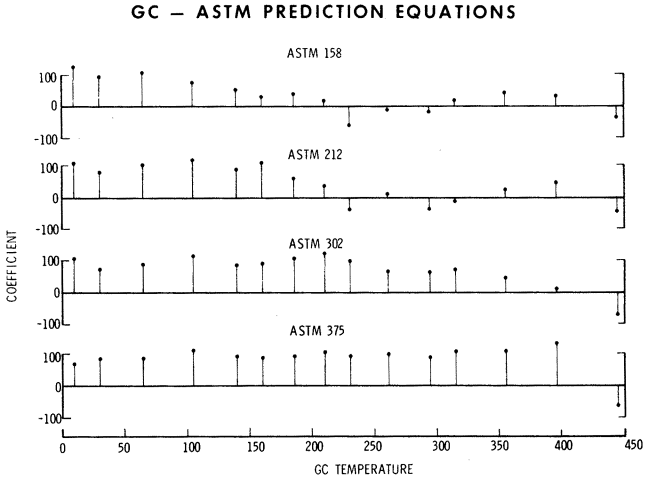

Figure 12.

Table 10
CORRELATION BETWEEN RIDGE AND GENERALIZED INVERSE REGRESSION COEFFICIENTS

Data Set	Number of Coefficients	Ridge		Generalized Inverse		Correlation Coefficient *
		Bias (k)	Vector Length ($\hat{\beta}'\hat{\beta}$)	Bias (r)	Vector Length ($\hat{\beta}'\hat{\beta}$)	
Acetylene Data (7)	9	.05	.524	3.8	.522	.98
Gorman & Toman (6)	10	.26	.373	6.6	.384	.92
Laird & Cady (14)	33	.30	.239	9.5	.236	.89
GC - ASTM 158°	15	.006	.358	7.0	.374	.96
Steam Data (5)	8	.30	.329	3.5	.318	.96
Rocket Engine (5)	13	.10	.817	9.0	.811	.90
McDonald & Schwing (21)	15	.18	.376	8.5	.384	.91
Liver Cirrhosis (2)	4	.30	.239	1.0	.262	.98

* Linear Correlation Coefficient

amination of the coefficients shows that the ridge and generalized inverse coefficients are remarkably similar. Table 9 shows that the generalized inverse model achieves a reduction of the prediction standard deviation comparable to the ridge model.

Table 10 shows the correlation coefficients between ridge regression coefficients and generalized inverse coefficients (chosen to have the same approximate vector length) for eight sets of data, including the three sets discussed in this paper. The correlation coefficients are all very high.

Computing Ridge Regression and Generalized Inverse Coefficients

One of the advantages of the ridge and generalized inverse estimators is ease of computation. The $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$ matrices are computed once and scaled to form the correlation matrix. For ridge regression, 10-30 inversions of $(\mathbf{X}'\mathbf{X} + k\mathbf{I})$, one for each value of k, are usually sufficient to determine where the ridge trace stabilizes. The generalized inverse coefficients are computed from the eigenvalues and eigenvectors of the correlation matrix. Numerical analysis is not a problem, nor is it a problem in the case of the ridge estimates where the addition of k to the diagonal of the correlation matrix reduces the nonorthogonality and thereby improves the numerical analysis. Furthermore, a separate matrix inversion is calculated for each value of k and roundoff errors cannot accumulate as in some stepwise algorithms. With the biased estimators, the same inverse matrix, for a given k or rank r, can be used to calculate the coefficients in the models for all responses. The best subset and stepwise algorithms require a separate computer run for each response.

Acknowledgments

We are grateful to R. W. Kennard for suggestions concerning the presentation of this material. The theoretical distillation results were generated from a model developed by J. B. Jones. M. H. Sarner and R. W. McGill did other computer programming required for this study. Appreciation is expressed to F. B. Cady for sending us the raw data associated with corn yield example. Insightful comments from the referees were most helpful in improving the presentation.

REFERENCES

[1] Allen, D. M. (1974): The relationship between variable selection and data augmentation and a method for prediction, *Technometrics, 16*, 125–127.
[2] Brownlee, K. A. (1965): *Statistical Methodology in Science and Engineering*, Second Edition, John Wiley & Sons, Inc., New York, N.Y.
[3] Cady, F. B., and Allen, D. M. (1972): Combining experiments to predict future yield data, *Agronomy Journal, 64*, 211–214.
[4] Daniel, C., and Wood, F. S. (1971): *Fitting Equations to Data*, Wiley Interscience, New York, N.Y.
[5] Draper, N. R., and Smith, H. (1966): *Applied Regression Analysis*, John Wiley & Sons, Inc., New York, N.Y.
[6] Gorman, J. W., and Toman, R. J. (1966): Selection of variables for fitting equations to data, *Technometrics, 8*, 27–51.
[7] Himmelblau, D. M. (1970): *Process Analysis by Statistical Methods*, John Wiley & Sons, Inc., New York, N.Y.
[8] Hoerl, A. E. (1959): Optimum solution of many variable equations, *Chemical Engineering Progress, 55*, 69 ff.
[9] Hoerl, A. E. (1962): Application of ridge analysis to regression problems, *Chemical Engineering Progress, 58*, 54–59.
[10] Hoerl, A. E., and Kennard, R. W. (1970): Ridge regression: biased estimation for nonorthogonal problems, *Technometrics, 12*, 55–67.
[11] Hoerl, A. E., and Kennard, R. W. (1970): Ridge regression: Applications to nonorthogonal problems, *Technometrics, 12*, 69–82.
[12] Kennard, R. W., and Stone, L. (1969): Computer aided design of experiments, *Technometrics, 11*, 137–148.
[13] Kunugi, T., Tamura, T., and Naito, T. (1961): New acetylene process uses hydrogen dilution, *Chemical Engineering Progress, 57*, 43–49.
[14] Laird, R. J., and Cady, F. B. (1969): Combined analysis of

yield data from fertilizer experiments, *Agronomy Journal, 61,* 829–834.
[15] Lindley, D. V., and Smith, A. F. M. (1972): Bayes estimates for the linear model (with discussion), *J. Royal Statistical Society, Series B, 34,* 1–41.
[16] Mallows, C. L. (1973): Some comments on Cp, *Technometrics, 15,* 661–675.
[17] Marquardt, D. W. (1963): An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Indust. Appl. Math., 11,* 431–441.
[18] Marquardt, D. W. (1970): Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics, 12,* 591–612.
[19] Marquardt, D. W. (1974): Discussion of "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," by A. E. Beaton and J. W. Tukey, *Technometrics, 16,* 189–192.
[20] Mayer, L. S. and Willke, T. A. (1973): On biased estimation in linear models, *Technometrics, 15,* 497–508.
[21] McDonald, G. C. and Schwing, R. C. (1973): Instabilities of regression estimates relating air pollution to mortality, *Technometrics, 15,* 463–481.
[22] Snee, R. D. (1973): Some aspects of nonorthogonal data analysis. Part I. Developing prediction equations, *J. Qual. Technol., 5,* 67–79.
[23] Theobald, C. M. (1974): Generalizations of mean square error applied to ridge regression. *J. Royal Statistical Society, Series B, 36,* 103–106.

P-values: Interpretation and Methodology*

JEAN D. GIBBONS** AND JOHN W. PRATT***

1. *Introduction*

The most common traditional method of carrying out any hypothesis test is to select a region for rejection and form a rejection rule such that the probability of committing a Type I error does not exceed some preselected number called the level of the test. Then the investigator reports whether or not the observations are "significant" at the chosen level. This procedure probably stems from the use of the Neyman-Pearson theory in classical statistics, where the decision function for the test is determined such that the probability of a Type II error is a minimum subject to the conditions imposed by the level selected. This method of test construction circumvents the problem of interrelationship between the probabilities of the Type I and Type II error. However, in many cases the choice of a significance level is completely arbitrary. In nonparametric statistics particularly, but also in parametric statistics when the null distribution is discrete, the chosen level may not even be attainable. Further, in nonparametric statistics, there is usually not sufficient information about alternative distributions so that the probability of a Type II error can even be discussed in general. Rather, the decision function is selected by logical reasoning, or according to the research hypothesis, or sometimes even by the data.

Another approach to hypothesis testing is currently attaining wide acceptance. This is the practice of reporting the smallest level at which the observations are significant in a particular direction. This quantity, which is herein called the *P-value*, is sometimes called the "critical level" or "significance level" (e.g., in Birnbaum, [3, p. 289]), the "observed level of significance" (e.g., in Kraft and Van Eeden, [8, p. 63]), the "prob-value" (e.g., in Wonnacott and Wonnacott, [12, p. 190]), or the "associated probability" (e.g. in Siegel, [11, p. 11]). Many elementary textbooks are now introducing this procedure, in addition to or instead of the more traditional one, for one sided tests based on both parametric and nonparametric methods. However, little attention has been paid to the proper interpretation of a *P*-value, nor to the inherent problem of defining *P*-values for two sided tests, particularly when the null distribution is not symmetric. These questions will be discussed in this paper, along with some comments about the need for making a clear distinction between statistical significance and practical significance in decision making.

2. *Methodology and Advantages of One Sided P-values*

Consider any hypothesis testing situation where the appropriate critical region for the test clearly lies in one particular tail of the sampling distribution of the test statistic. Then the observed value of the test criterion can be used to compute a tail probability which we call the *P*-value. The *P*-value is defined as the probability under null distributions of a sample outcome equal to or more extreme than that observed. In well-behaved problems, which include almost all one sided tests commonly used, the possible outcomes can be ordered according to how "extreme" they are in one direction relative to the outcome expected under the null hypothesis, and the values of the test statistic are also ordered in a corresponding manner. Then the *P*-value is a well defined quantity, because the meaning of extreme is clear.

* This paper was written by Gibbons, but its content overlaps parts of Chapter 1 of a forthcoming book, *Concepts of Nonparametric Theory*, written by both authors. The first draft of Chapter 1 was prepared by Pratt.

** Dept. of Statistics, Univ. of Alabama, University, AL 35486.
*** Grad. School of Bus. Admin., Harvard Univ., Boston, MA 02163.