A Critique of Some Ridge Regression Methods

GARY SMITH and FRANK CAMPBELL*

Ridge estimates seem motivated by a belief that least squares estimates tend to be too large, particularly when there is multicollinearity. The ridge solution is to supplement the data by stochastically shrinking the estimates toward zero. Although flexibility is provided by the abstention from exact exclusion restrictions, ridge regression retains many weaknesses of similarly motivated procedures: a neglect of the basic fact that linear transformations should not change the implicit estimates of a model's coefficients, an incorrect labeling of nonorthogonal data as weak, and a loose representation of a priori beliefs and reliance at times on ad hoc pseudoinformation.

KEY WORDS: Ridge regression; Multicollinearity; Principal components.

1. INTRODUCTION

The familiar multivariate regression model has been found useful in a wide variety of practical applications. Researchers often find, however, that their data do not contain enough information to answer decisively the questions that the researchers have posed. Those with a well-defined specification may obtain confidence regions that are so large that the point estimates and forecasts are of little interest. Those who search for a model may find that the variation in a particular dependent variable can seemingly be explained equally well by an annoyingly wide variety of theoretically motivated and even randomly selected explanatory variables.

In response, a minority of researchers are content to shrug their shoulders and note the inadequacy of the data.[1] The purest (or laziest) simply estimate a single model and note that more precise estimates will require more information. Others believe that they have a good deal of outside information based on common sense or earlier studies. Such researchers commonly try to improve the reported estimates by changing their model. They may initially limit themselves to a small number of explanatory variables that are chosen in part for their high variances and relative orthogonality, or they may begin with a more complete model and then drop those variables with coefficients that are found to be incorrectly signed or statistically insignificant. The most ambitious researchers seem to toil endlessly for that elusive combination of variables that will yield statistically significant and plausibly signed parameter estimates.

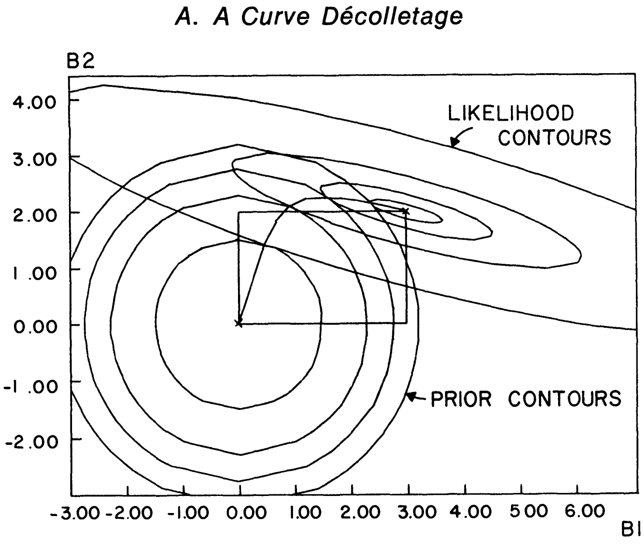
We enthusiastically advocate the use of a priori information (e.g., see Smith and Brainard 1976). In practice, however, this use too often involves only the iterative imposition of exact (typically exclusion) restrictions on individual parameters. Exact restrictions are discomforting since they force one into the delicate position of having to choose between omniscience and ignorance. In addition, a myopic parameter-by-parameter procedure neglects the opportunity for hedging provided by covariances in the data or the prior information. For example, Smith (1974) shows that parameter constraints that are individually more accurate than the corresponding unconstrained estimates may collectively worsen the constrained estimates of all the remaining coefficients. In particular, setting two "incorrectly signed" coefficients equal to zero may worsen the model's forecasting performance even when the two unconstrained estimates do in fact have the wrong signs. Thus, one should be cautious about mechanically shrinking individual parameter estimates toward what are believed to be more likely values.

This hedging arises in formal Bayesian procedures even for a two-parameter problem with orthonormal prior information. Consider, for example, a situation in which the explanatory variables are highly positively correlated, the second explanatory variable has a larger variance than the first, and the a priori parameter means are zero while both least squares estimates are positive. In this situation, the first parameter B1 is more receptive to prior information and, as indicated by the negative covariance for the least squares estimates, the data will resist reducing (or increasing) both estimates. Prior and likelihood contours of this type are displayed in Figure A. As one can see from the curve décolletage (Dickey 1974), unless the prior information is quite firm (low variances), the optimal procedure is to shrink the estimate of B1 toward zero while increasing the estimate of B2. A trial-and-error imposition of exact parameter restrictions is unlikely to stumble on the optimal estimates or to recognize them as such. In the more complicated problems that are usually encountered, it will be exceedingly difficult to perform the sophisticated mental gymnastics that are required to locate satisfactory estimates with a

* Gary Smith is Associate Professor, Department of Economics, University of Houston, Houston, TX 77004. Frank Campbell is employed by the Corporate Research Division, Traveler's Insurance Co., Hartford, CT 06115. This work was supported in part by grants from the National Science Foundation and from the Ford Foundation. The authors wish to thank the appropriate editors and referees for helpful suggestions.

[1] A conscientious referee informed us that the frequent lament, "One cannot make bricks without straw," is a common British idiom that is "recorded in English literature as early as 1658 in the memoirs of the Verney family ... More recently, Sherlock Holmes—who had a flair for misquotation—is reported to have said, 'Data! data! data! I can't make bricks without clay.'" (J.H. Watson, M.D., "The Adventure of the Copper Beeches")

© Journal of the American Statistical Association
March 1980, Volume 75, Number 369
Theory and Methods Section



simple search procedure. These considerations argue for an explicitly Bayesian approach that accurately and efficiently incorporates one's a priori beliefs.

There is a third category of researchers who search over various combinations of explanatory variables with little regard for a priori information about the parameters. These specification searches can be viewed as the iterative imposition of exact parameter restrictions on a more general model. If these implicit restrictions have no a priori weight behind them, then the final reported estimates and statistics will have little meaning. By emphasizing the estimated variances but neglecting the biases that have been introduced (but cannot be measured), the researcher does little more than disguise the imprecision of the estimates. Techniques such as stepwise regression, generalized inverses, and principal components involve the formal imposition of wholly ad hoc parameter restrictions and will be successful only by fortuitous accident (see Smith 1974). It is very difficult to be comfortable with mechanical data manipulation that is insensitive to the particular phenomena being modeled and to information about the coefficients.

There has recently been some interest in ridge regression as a method for coping with inadequately informative data. This approach seems to blend the popular practices that we have here disparaged, in that it is often motivated by a priori information that it does not accurately describe and can degenerate into ad hoc data manipulation.

Ridge estimates seem to be motivated by the belief that least squares estimates tend to be too large, particularly when there is a multicollinearity problem. The ridge solution is to supplement the data by stochastically shrinking the estimates toward zero. The abstention from exact exclusion restrictions contributes to the flexibility and attractiveness of the procedure. Unfortunately, ridge regression retains many of the weaknesses of earlier procedures.

One characteristic of ridge regression is the neglect of the basic fact that a linear transformation of a model does not change the model and should not change the estimates of a model. A second characteristic is the incorrect labeling of nonorthogonal data as inadequately informative data. Multicollinearity can be one source of weak data, but the strength of the data cannot be measured solely by the orthogonality of the data. A third characteristic is the use of a loose representation of a priori beliefs and the reliance at times on ad hoc pseudoinformation.

In the following sections we will discuss each of these points in turn and illustrate them by referring to an article by Marquardt and Snee (1975). For simplicity, we will discuss only their first example, which deals with acetylene data.

2. WHICH PARAMETERS SHOULD BE EXPLICITLY ESTIMATED?

The standard linear regression model is

$$Y = X\beta + \epsilon \ , \quad \epsilon \sim N[0, \sigma_\epsilon^2 I] \ , \tag{2.1}$$

where Y is a $(\tau \times 1)$ vector of observations of the dependent variable, X is a $(\tau \times n)$ matrix of observations of the n independent variables, β is an $(n \times 1)$ vector of the unobserved parameters, and ϵ is a $(\tau \times 1)$ vector of the unobserved disturbance term. The elements ϵ_i of ϵ are assumed to be independent and normally distributed with zero means and constant variance σ_ϵ^2.

Estimates of the n elements of β form a basis for the implicit estimates of any linear combination of these parameters. Indeed, the forecasts of Y are an application of these implicit estimates. The model (2.1) can always be rewritten as

$$Y = (XA)(A^{-1}\beta) + \epsilon = Z\gamma + \epsilon \ , \tag{2.2}$$

using any nonsingular matrix A. The parameters β and γ are uniquely related by $\beta = A\gamma$ and their estimates should be also. It should make no difference whether β is estimated explicitly or implicitly from $\beta = A\gamma$.

A nonsingular linear transformation of a model does not change a model and should not change the implicit estimates of the model's parameters. Similarly, the minimization of a well-defined loss function should not depend on whether certain parameters are estimated directly or indirectly. If the estimates do vary, then the loss function has been inadvertently altered by the estimation procedure. There is no theoretical reason to prefer the representation of (2.1) to (2.2), and thus the decision to estimate explicitly β or γ should be entirely arbitrary. Practitioners should be unsettled by an estimation procedure that is affected by this arbitrary choice.

Consider, for example, a consumption function

$$C = b_0 + b_1 Y_P + b_2 Y_T + b_3 r_S + b_4 r_L + b_5 S_{-1} + b_6 L_{-1} + b_7 S_{-2} + b_8 L_{-2} + \epsilon \ ,$$

where Y_P and Y_T are the permanent and transitory components of income, $Y = Y_P + Y_T$; r_S and r_L are the yields on short-term and long-term assets; and S and L are stocks of short-term and long-term assets. One could just as easily have written the identical model in a wide variety of equivalent ways: instead of Y_P and Y_T, use Y and Y_P or Y and Y_T; instead of a consumption function, explain saving, $S \equiv Y - C$; instead of r_S and r_L, use r_S and $r_L - r_S$ or r_L and $r_L - r_S$; instead of S_{-1} and L_{-1}, use $S_{-1} + L_{-1}$ and S_{-1} or $S_{-1} + L_{-1}$ and L_{-1}; instead of S_{-1} and S_{-2}, use S_{-2} and $S_{-1} - S_{-2}$ or S_{-1} and $S_{-1} - S_{-2}$. In every case, there is a simple unique linear equivalence between both the variables and the coefficients. None of these representations is superior to the others nor to the infinite variety of less obvious ones (such as using Y and $3Y_P - 2Y_T$). They are all fully equivalent, and one surely wants to avoid the anxiety that should accompany the use of an estimation procedure that depends on such choices. The point is obvious, but there are a number of instances and even formal estimation procedures that ignore it. Ridge regression is one of these procedures.

3. ARE THE DATA INFORMATIVE?

The starting point for ridge regression (and similarly motivated techniques) is the decision that the data are inadequate and need to be augmented. The traditional yardstick (e.g., see Farrar and Glauber 1967) is to measure the intercorrelations among the explanatory variables by calculating either the eigenvalues or the off-diagonal elements of the moment matrix or its inverse. This is clearly inadequate since the model can always be rewritten as from (2.1) to (2.2) so that $Z'Z$ is an identity matrix with no off-diagonal elements and all eigenvalues equal to one. Because the implicit least squares estimates are unchanged, the model is no more informative than it was before. High intercorrelations among variables have simply been transformed into low variances on linear combinations of variables.

Consider, for example, the earlier consumption function and the data assumptions that Y_P and Y_T are uncorrelated, with respective variances of 9 and 1. With Y_P and Y_T as explanatory variables there is no collinearity problem; with Y_P and Y there is, since the correlation coefficient between them is .95. In the first case the high variance on the estimate of b_2 is attributed to the low variance of Y_T. In the second case, it is attributed to the high correlation between Y and Y_P. These are, of course, equivalent descriptions. We surely do not want to use a measure of informational content that depends on whether we use Y_P and Y_T or Y_P and $Y_P + Y_T$ as explanatory variables.

Marquardt and Snee (1975) use such a measure and advocate rearranging the explanatory variables to reduce the readings on their yardstick:

> 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 straightforward interpretation and use. (p. 3)

This standardization applies a unique nonsingular linear transformation to the variables and consequently has no effect on the model, the least squares forecasts, or the implicit least squares estimates of any of the coefficients. Consider specifically the acetylene data example of Marquardt and Snee. The unstandardized model

$$Y = b_0 + \sum_{i=1}^{3} b_i X_i + \sum_{1 \le i \le j}^{3} b_{ij} X_i X_j + \epsilon \tag{3.1}$$

can be rewritten in the equivalent standardized form

$$\begin{aligned} Y &= [b_0 + \sum_{i=1}^{3} b_i \bar{X}_i + \sum_{1 \le i \le j}^{3} b_{ij} \bar{X}_i \bar{X}_j] \\ &\quad + \sum_{i=1}^{3} [b_i + b_{ii}\bar{X}_i + \sum_{i=1}^{3} b_{ij}\bar{X}_j] S_i (X_i - \bar{X}_i)/S_i \\ &\quad + \sum_{1 \le i \le j}^{3} b_{ij} S_i S_j (X_i - \bar{X}_i)(X_j - \bar{X}_j)/S_i S_j + \epsilon \\ &= \beta_0 + \sum_{i=1}^{3} \beta_i Z_i + \sum_{1 \le i \le j}^{3} \beta_{ij} Z_i Z_j + \epsilon \end{aligned} \tag{3.2}$$

(where $Z_i = (X_i - \bar{X}_i)/S_i$ and the scaling has been by $S_i = [\sum (X_i - \bar{X}_i)^2/(n-1)]^{\frac{1}{2}}$).

Because rewriting the model in the form (3.2) does not affect any of the implicit estimates, it has no effect on the amount of information contained in the data. Nonetheless, Marquardt and Snee consider the representation (3.2) to be preferable to (3.1) because of the reduced variance inflation factors (VIF), which they use to measure ill conditioning and to indicate whether ridge regression methods should be used.

In the general model (2.1), the variance of the least squares estimate of β_k is given by

$$\operatorname{var}(\hat{\beta}_k) = \frac{\sigma_\epsilon^2}{(n-1)S_k^2}\left(\frac{1}{1 - R_k^2}\right) ,$$

where R_k^2 is the squared multiple correlation coefficient between the kth variable and the remaining explanatory variables. The variance inflation factor is defined as

$$\operatorname{VIF}(\hat{\beta}_k) = 1/(1 - R_k^2) ,$$

which can be interpreted as the ratio of the variance of $\hat{\beta}_k$ to what that variance would be if X_k were uncorrelated with the remaining X_i.

The inadequacy of this measure is again illustrated by the cosmetic extreme of orthogonalizing the data so that all the VIF's are equal to one. Because orthogonalization does not improve the model or the estimates, VIF's are an inadequate measure of ill conditioning. Again, any multicollinearity problem can be equivalently described as a problem of low variation, and it can be misleading to measure one and neglect the other.

When the acetylene model is written in form (3.1), there are some very high intercorrelations among the variables, due in part to the fact that there are only 16 observations on 9 variables, with 6 of the variables constructed as products of the other 3. One of the more dramatic and yet easily understood facets of this example is the term $b_{11}X_1^2$. The only observations on X_1 are 6 at 1,300, 6 at 1,200, and 4 at 1,100, which gives a simple correlation between X_1 and X_1^2 of .99967. Overall, the squared correlation between X_1^2 and the remaining variables is .9999996, which gives b_{11} a VIF of 2.5 million. This means that the variance of b_{11} will be large unless the variance of X_1^2 is large or the variance of the disturbance term is small. In this example, the former is 3.77×10^{10} and the latter is estimated to be .81258, so that

$$\operatorname{var}(\hat{b}_{11}) = \frac{.81258}{15(3.77 \times 10^{10})}(2.5 \times 10^6) = .36 \times 10^{-5} .$$

Is this large or small? Obviously one cannot say without knowing something about b_{11} and the use that will be made of its estimate. Marquardt and Snee state that a VIF of 2 million "is unthinkable and unnecessary," since the model can be written in the standardized form (3.2), in which the coefficient of $(X_1 - \bar{X}_1)^2/S_1^2$ has a VIF of less than 2,000, as the squared correlation of this variable with the remaining variables in (3.2) is "only" .99943. Because the coefficient of this variable is $b_{11}S_1^2$, the implicit estimate

$$\hat{b}_{11} = \widehat{(b_{11}S_1^2)}/S_1^2$$

will be identical to the estimate that is directly obtained from (3.1). As they note, scaling does not affect the VIF (since it doesn't affect the correlation coefficients), so that in form (3.2),

$$\operatorname{VIF}(\hat{b}_{11}) = 1/(1 - .99943) = 1{,}762.58 ,$$

while

$$\operatorname{var}(\hat{b}_{11}) = \left[\frac{\sigma_\epsilon{}^2}{(n-1)\sigma^2{}_{(X_1-\bar{X}_1)^2}}\right]\operatorname{VIF}(\widehat{b_{11}S_1^2})$$

$$= \frac{.81258}{(15)(2.656 \times 10^7)}1{,}762.58 = .36 \times 10^{-5} .$$

Thus, the use of form (3.2) has no effect on the estimate of b_{11} or on the precision of this estimate. The VIF has been reduced by a factor of 1,000 but so has the variance of the associated variable, so that the imprecision has simply been relabeled a problem of low variation rather than one of high covariation. A similar analysis could be carried out for any of the estimable coefficients.

Thus, although some may find the VIF helpful in describing the sources of imprecision, it does not measure the amount of imprecision and cannot be used to justify the reliance on weakly held supplementary information, nor can it be used to motivate linear transformations of the variables.

The essential problem with VIF and similar measures is that they ignore the parameters while trying to assess the information given by the data. Clearly, an evaluation of the strength of the data depends on the scale and nature of the parameters. One cannot label a variance or a confidence interval (or, even worse, a part of the variance) as large or small without knowing what the parameter is and how much precision is required in the estimate of that parameter. In particular, a seemingly large variance may be quite satisfactory if the parameter is very large, if one has strong a priori information about the parameter, or if the parameter is uninteresting (perhaps because the associated variable will be constant during the forecast period). A meaningful assessment will require a well-defined loss function that must necessarily depend on the particular problem being examined.

4. WHAT A PRIORI INFORMATION SHOULD BE USED?

For estimation purposes, Marquardt and Snee (1975) prefer to rewrite the model in correlation form:

$$\frac{Y - \bar{Y}}{(n-1)^{\frac{1}{2}}S_Y} = \sum_{i=1}^{3}\left(\frac{\beta_i}{S_Y}\right)\left[\frac{Z_i}{(n-1)^{\frac{1}{2}}}\right] + \sum_{1 \le i \le j}^{3}\left(\beta_{ij}\frac{S_{ij}}{S_Y}\right)\left[\frac{Z_iZ_j - \bar{Z}_i\bar{Z}_j}{(n-1)^{\frac{1}{2}}S_{ij}}\right] + \frac{\epsilon - \bar{\epsilon}}{(n-1)^{\frac{1}{2}}S_Y} \tag{4.1}$$

(which is again a transformation that does not alter the model). They then argue that "the 'fly in the ointment' with least squares is its requirement of unbiasedness. . . . Thus, it is meaningful to focus on the achievement of small 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" (p. 5).

A mean squared error comparison (MSE) for these suggested estimators is always ambiguous, however, since the size of the bias depends on the unknown population values of the parameters. Indeed, most non-Bayesians believe that a major advantage of unbiased estimators is that the mean squared errors do not depend on the actual values of the parameters.

Alternatively, least squares can be justified on likelihood grounds or as the mode of the posterior with improper uniform priors. In response to these justifications, Marquardt and Snee argue that a reasonable person would have bounded priors and that in correlation form "it is exceedingly rare for the population value of any regression coefficient to be larger than three in a real problem" (p. 6). This claim is consistent with the interpretation of ridge estimators as a method for introducing auxiliary information. The problem is that the particular prior information implicit in a ridge regression is inadequately justified.

Consider specifically the supplementary information to the model (2.1),

$$\beta = b + u$$
$$u \sim N(0, \Sigma) \ .$$

Theil and Goldberger's (1961) mixed estimation is a classical approach that views β as fixed and b and Y as random and applies generalized least squares to the two sets of data to obtain

$$\beta^* = [X'X + \sigma_\epsilon^2\Sigma^{-1}]^{-1}[X'Y + \sigma_\epsilon^2\Sigma^{-1}b] \ .$$

Chipman's (1964) partially Bayesian analysis takes b and Y as fixed and β as random and obtains β^* as the linear minimum mean squared error estimator. In a fully Bayesian approach with known σ_ϵ^2, β^* is the posterior mean.

Since the ridge estimator is

$$\hat{\beta}^R = [X'X + kI]^{-1}X'Y \ ,$$

it is possible to motivate ridge regression from a wide variety of viewpoints when one actually has a priori information of the special form

$$b = 0 \quad \text{and} \quad \Sigma = \sigma_\epsilon^2 I/k \ ,$$

which is to say that one has orthogonal priors with common variances centered at the origin. Conversely, the theoretical inadequacy of ridge regression is that little effort is made to assess the appropriateness of these implicit priors.

If the true prior probability distribution for β is Gaussian, then the prior distribution can always be centered, diagonalized, and scaled so that a ridge estimator is appropriate. That is, the model can be rewritten as in (2.2) so that the priors on $\gamma = A^{-1}\beta$ are orthonormal. The problem with ridge estimation in practice is that the model is linearly transformed to center, scale, and diagonalize partially the variables rather than the prior distributions of the parameters. Indeed, there is no discussion of the reasonableness of the implicit assumption that the parameters have zero means, zero covariances, and identical variances. It is not enough to assert that parameters are almost always less than three. The implicit ridge prior distributions cannot apply to all parameters, regardless of how the model is specified. Thinking back to the consumption function, which of the following parameters should be shrunk towards zero: the marginal propensity to consume out of income, the marginal propensity to save out of income, the difference between the marginal propensities to consume out of permanent and transitory income, or the difference between .9 and the marginal propensity to consume out of permanent income? The researcher's choice will make a difference and should be made consciously and explicitly. In the acetylene example; if the prior distributions implicit in a ridge approach are actually appropriate for the model in any one of the forms (3.1), (3.2), or (4.1), then ridge regression will be generally inappropriate for the remaining forms because the prior variances will not all be equal and some of the covariances will be nonzero.

If one blithely manipulates the data with no regard for the appropriateness of the implicit ridge prior distributions, the ridge estimates may literally be anything. Even if we restrict ourselves to diagonal transformations in (2.2) of the model (2.1), the implicit ridge estimates

$$\begin{aligned}\hat{\beta}^R = A\hat{\gamma}^R &= A[Z'Z + kI]^{-1}Z'Y \\ &= [X'X + kA^{-1}A^{-1}]^{-1}X'Y\end{aligned}$$

can be set equal to any arbitrary values $\bar{\beta}$ by selecting the n diagonal elements of A to satisfy the n equations

$$kA^{-1}A^{-1}\bar{\beta} = X'X(\hat{\beta}^{\mathrm{OLS}} - \bar{\beta}) \ .$$

It is of course not the objective of ridge users to obtain estimates that have preassigned values. This possibility is discussed here to dramatize the arbitrariness of the ridge estimates if data manipulation proceeds unchecked by an assessment of the reasonableness of the implicit ridge priors.

Similar difficulties arise in the selection of k if the ridge user does not take into account the fact that the implicit variances on the priors are being set at σ_ϵ^2/k. Instead, k is chosen by ad hoc procedures[2] whose loose theoretical underpinnings are indicated by the arbitrary restriction that k be less than one, which compels an assumption that the variance of the disturbance term is less than the variance on the priors. Inside this range, Marquardt and Snee choose a k that yields reasonable variance inflation factors (which we have seen to be a misleading objective) and parameter estimates that are relatively insensitive to small changes in k. Figure B gives a ridge trace for the acetylene data example. The intercepts are the least squares estimates ($k = 0$); the next unit corresponds to $k = .0001$, and each unit thereafter corresponds to a 50 percent increase in k over the preceding unit. This logarithmic scale seems the most appropriate for comparing changes in k. Marquardt and Snee indicate that is it not difficult to select k from the ridge trace, but we do not view the estimates as stable for the k's of .01 or .05 that they selected and, indeed, do not see much stability for any values of k less than one. In addition, there is the problem that this sensitivity analysis is not invariant to linear transformations of the model. That is, ridge traces of linear combinations of the parameters will not generally show the same regions of relative stability. More fundamentally, we do not understand why k should be chosen on the basis of local in-

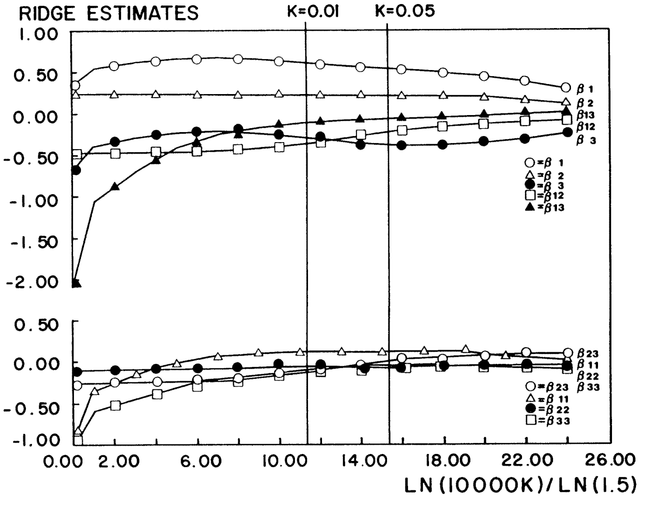
[2] Hoerl and Kennard list the following criteria: (a) stable parameter estimates, (b) reasonable absolute values for the parameter estimates, (c) correctly signed parameter estimates, and (d) a reasonable value for the residual sum of squares. This is a richer set of objectives, although attainment will often be difficult with only one control variable. In addition, the selection of k is not based on the theoretical interpretation of the parameter. If the researcher can cite reasonable values for the coefficients, then the researcher ought to incorporate this information directly in a Bayesian fashion, rather than tinker with k so that the ex post estimates will be close to the a priori values.

sensitivity. Even if the estimates were not much different with k at .01 or .02, this is not a convincing argument for using these estimates rather than the very different estimates that result from a k of .001 or 2.0.

Marquardt and Snee also write that "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" (p. 12). We have already pointed out that orthogonality does not imply strong data; therefore, the ridge trace is misleading if it favors the least squares point in this situation. And, since the data can always be orthogonalized, this statement by Marquardt and Snee illustrates our previous point that the selection of k depends on the parameters for which the ridge trace is drawn.

It seems to us that, without prior information, the only theoretically defensible use of ridge regression would be with a value of k that was so small as to give, for all practical purposes, the least squares estimates. One could then argue that one was simply assuming proper locally almost uniform priors. Marquardt and Snee do in fact argue that "the ridge estimate is equivalent to placing mild boundedness requirements on the coefficient vector" (p. 6). Were this so, ridge estimates would be only a formal curiosity. Of course it is not so, since ridge methods are specifically intended to obtain estimates that are significantly different from the presumedly unsatisfactory least squares estimates. This is clear in the acetylene data example from a comparison of the estimates in Table 2 of Marquardt and Snee (1975) and also from an examination of the values of k that were used, .01 and .05.

Using the least squares estimated value of σ_ϵ^2, the implicit common variances on the priors are, respectively, .0328 and .0066, which most would feel are fairly tight prior distributions given the size of the least squares estimates. In terms of confidence intervals, Marquardt and Snee acted as if for each parameter they were 95 percent confident that the population value of the parameter is no further than .36 or .16 from zero. In contrast, five of the nine least squares estimates are outside the larger interval and eight are outside the smaller. It seems that they have assumed more than "mild boundedness."


B. Acetylene Data Ridge Trace

In fact, Theil's suggested test indicates that Marquardt and Snee's implicit prior distributions are actually incompatible with the data. For $k = .01$, the chi-squared statistic is 24.7, and for $k = .05$ it is 93.5, as compared with a critical point of 16.9 for a test at the 5 percent level. With prior distributions of the ridge type, one would have to use a $k \leq .006$ to pass Theil's test and a much smaller k to be assuming only mild boundedness.

5. RIDGE REGRESSION, PRINCIPAL COMPONENTS, AND MARQUARDT'S GENERALIZED INVERSE

Marquardt (1970) shows that when the data are transformed into their principal components, there is a clear relationship between ridge regression, principal components analysis, and Marquardt's generalized inverse technique.

If the columns of A in (2.2) are the orthonormal eigenvectors of $X'X$, then the columns of Z are the principal components of $X'X$, and $Z'Z$ is a diagonal matrix with the eigenvalues (λ_i) of $X'X$ as its diagonal elements. The implicit ridge priors on β can be similarly transformed as follows:

$$\beta = 0 + U \ , \quad E(UU') = (\sigma_\epsilon^2/k)I$$

$$\gamma = A'\beta = 0 + A'U \ , \quad E(A'UU'A) = (\sigma_\epsilon^2/k)I \ .$$

Thus, a ridge analysis could be made by using either the original data or the principal components. In this latter form, however, the orthogonality of both the data and the priors provides estimates that are simple weighted averages of the likelihood estimate and the prior mean

$$\hat{\gamma}_i^R = \left(\frac{\lambda_i}{\lambda_i + k}\right)\hat{\gamma}_i + \left(\frac{k}{\lambda_i + k}\right)0 \ .$$

Those estimates with the largest variances ($\sigma_\epsilon^2/\lambda_i$) are shrunk the most, and the larger is k, the closer all these estimates are to zero.

There is an obvious similarity between this and the usual principal components analysis. In the latter, the estimates are typically

$$\begin{aligned} \hat{\gamma}_i^P &= \hat{\gamma}_i \ , \quad \lambda_i \geq h \ , \\ &= 0 \ , \quad \lambda_i < h \ , \end{aligned}$$

where h is some selected cutoff point for retaining components. In this form of principal components analysis, either zero or the least squares point is selected, depending on the relative variance of the least squares estimate. In ridge regression, estimates are selected that lie between zero and the least squares estimate based also on the relative variance of the estimate.

In Marquardt's generalized inverse technique, allowance is made for intermediate eigenvalues that are

Estimates of the Coefficients of the Principal Components

Parameter	Least Squares Estimate (t statistic)	λ_i	Ridge Estimates $k = .01$	Ridge Estimates $k = .05$	Generalized Inverse $r = 3.8$
γ_1	.352 (39.88)	4.205	.351	.348	.352
γ_2	−.005 (−.39)	2.163	−.005	−.005	−.005
γ_3	−.600 (−35.38)	1.138	−.595	−.575	−.600
γ_4	.238 (13.43)	1.041	.236	.227	.190
γ_5	.009 (.33)	.3845	.009	.008	.0
γ_6	.217 (2.67)	.0495	.181	.108	.0
γ_7	−.383 (−2.47)	.0136	−.221	−.082	.0
γ_8	.521 (2.06)	.0051	.176	.048	.0
γ_9	−2.401 (−1.31)	.0001	−.023	−.005	.0
SSR	.0023		.0039	.0067	.0110

neither obviously large nor small, and the associated estimates are put partway between zero and the least squares estimate:

$$\begin{aligned} \hat{\gamma}_i^G &= \hat{\gamma}_i \ , & \lambda_i &\geq h_1 \\ &= r\hat{\gamma}_i \ , & h_2 < \lambda_i &< h_1 \ . \\ &= 0 \ , & \lambda_i &\leq h_2 \end{aligned}$$

All three of these procedures consequently fit into the class of estimators

$$\tilde{\gamma}_i = \alpha_i \hat{\gamma}_i + (1 - \alpha_i) C_i \ , \qquad (5.1)$$

which use simple weighted averages of the least squares estimate $\tilde{\gamma}_i$ and some point C_i, where the weights $0 \leq \alpha_i \leq 1$ depend on the variances of the least square estimates.

These techniques are consequently all subject to the same criticisms of ridge regression that we have made here. They are sensitive to unimportant linear transformations of the model. The strength of the data is inadequately measured, and insufficient attention is paid to the appropriateness of the implicit priors.

In this principal components framework, arbitrary linear combinations of parameters are shrunk toward zero. We will return to the choice of target in a moment. First, we should note that the degree of shrinkage depends only on the relative variance of the least squares estimate. This is an inadequate description of the strength of the data relative to one's priors since it ignores the absolute size of the variances (which depends on σ_ϵ^2), the distance between the least squares estimate and the prior mean, and the true strength of one's prior beliefs. These inadequacies are again apparent in the acetylene data example. The table displays the least squares estimates of the coefficients of the principal components, the associated eigenvalues, the ridge estimates, and the generalized inverse estimates.

Many of the coefficients with low eigenvalues are significantly different from zero because of the size of the estimates and the small value of σ_ϵ^2. As a consequence, four of the six restrictions imposed by the generalized inverse procedure would be rejected by individual classical hypothesis tests at the 5 percent level, and the five exact restrictions would even be rejected by a joint test.

The inadequacy of retaining components on the basis of eigenvalues was recognized several years ago by Hotelling (1957), who pointed out that components that are of little use in explaining variation in the explanatory variables may still be very powerful in explaining the dependent variable. This had led Massy (1965) and others to advocate the deletion of components whose coefficients are statistically insignificant. Although this avoids the imposition of weakly held restrictions that are rejected by the data, it still mechanically overrules the likelihood point whenever an arbitrarily selected point (zero) is inside the confidence interval. Thus, this approach permits one to impose ad hoc constraints where the data are very informative (when zero is inside a tight band) and to refrain when constraints are badly needed (when zero is outside a large band).

More generally, it is distressing how little effort is expended on the selection of a shrinkage target. Zero should have no special claim on our attention. The arbitrariness is indicated by the fact that the shrinkage of some parameters toward zero necessarily expands other linear transformations away from zero. In the earlier comsumption function, as the marginal propensity to consume moves toward zero, the marginal propensity to save moves away from zero. Instead of trying to select parameters to shrink towards zero (or choosing them arbitrarily), why not select reasonable values to shrink the parameters toward?

These points can be illustrated in more detail with the specific loss function adopted by Marquardt and Snee (1975) and by Hoerl and Kennard (1970). The MSE (or expected squared distance to β) of the estimator $\tilde{\beta}$ is

$$L = E(\beta - \tilde{\beta})'(\beta - \tilde{\beta}) = \sum_i \text{MSE}(\tilde{\beta}_i) \ .$$

The equal weighting of these particular parameters is arbitrary. If the data are not orthonormal, this loss function does not ensure smaller mean squared prediction error.

Again, we can work with principal components as these preserve average MSE

$$\begin{aligned} E(\beta - \tilde{\beta})'(\beta - \tilde{\beta}) &= E(\beta - \tilde{\beta})'AA'(\beta - \tilde{\beta}) \\ &= E(A'\beta - A'\tilde{\beta})'(A'\beta - A'\tilde{\beta}) \\ &= E(\gamma - \tilde{\gamma})'(\gamma - \tilde{\gamma}) \\ &= \sum_i \text{MSE}(\tilde{\gamma}_i) \ . \end{aligned}$$

Now, for the procedures considered here (5.1), if the α_i were nonstochastic,[3] the MSE could be broken into two parts,

$$\begin{aligned} E(\gamma_i - \tilde{\gamma}_i)^2 &= \alpha_i^2\,\mathrm{MSE}(\hat{\gamma}_i) + (1 - \alpha_i)^2(\gamma_i - C_i)^2 \\ &= \mathrm{var}(\tilde{\gamma}_i) + [\mathrm{bias}(\tilde{\gamma}_i)]^2\ . \end{aligned}$$

Thus, as compared with the least squares estimate $(\hat{\gamma}_i)$, shrinking unambiguously reduces the variance and increases the bias. If the least squares variance is not zero and $(\gamma_i - C_i)^2$ is bounded, then there will always be some weights α_i that reduce the MSE. If, however, γ_i is unknown, then one will also not know whether or not the MSE has been reduced.

Notice also that the variance reduction is entirely independent of the shrinking target, C_i. That is, shrinking toward the origin is of no advantage for variance reduction. Where the choice of target does show up is in the squared bias, and here a more accurate target is unambiguously beneficial. Thus, the origin can only be justified as a shrinking target if it is favored over other potential targets on a priori grounds. But if one has such a priori beliefs, then they should be directly incorporated. When the least squares estimates are imprecise, auxiliary information is quite useful. Pseudoinformation is of dubious value.

[3] This type of analysis is considerably more complex when the α_i's depend on the least squares estimates and are therefore stochastic and is intractable when the α_i's are selected from visual inspections of ridge traces.

REFERENCES

Chipman, J.S. (1964), "On Least Squares With Insufficient Observations," *Journal of the American Statistical Assocation*, 59, 1078–1111.

Dickey, J.M. (1974), "Bayesian Alternatives to the *F*-Test and Least Squares Estimates in the Normal Linear Model," in *Studies in Bayesian Econometrics and Statistics*, eds. A Zellner and S. Fienberg, Amsterdam: North-Holland Publishing Co.

Farrar, D., and R. Glauber (1967), "Multicollinearity in Regression Analysis: The Problem Revisited," *Review of Economics and Statistics*, 49, 92–107.

Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

——— (1970), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, 12, 69–82.

Hotelling, H. (1957), "Relation of the Newer Multivariate Statistical Methods to Factor Analysis," *British Journal of Statistical Psychology*, 10, 69–79.

Leamer, E.E. (1973), "Multicollinearity: A Bayesian Interpretation," *Review of Economics and Statistics*, 55, 371–380.

Marquardt, D.W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation," *Technometrics*, 12, 591–612.

Marquardt, D.W., and Snee, R.D. (1975), "Ridge Regression in Practice," *The American Statistician*, 29, 3–20.

Massy, W. (1965), "Principal Components in Exploratory Statistical Research," *Journal of the American Statistical Association*, 60, 234–256.

Mayer, L.S., and Wilke, T.A. (1973), "On Biased Estimation in Linear Models," *Technometrics*, 15, 497–508.

Smith, Gary (1974), "Multicollinearity and Forecasting," Cowles Foundation Discussion Paper No. 383.

Smith, Gary, and Brainard, William (1976), "The Value of a Priori Information in Estimating a Financial Model," *Journal of Finance*, 31, 1299–1322.

Theil, H. (1971), *Principles of Econometrics*, New York: John Wiley & Sons.

Theil, H., and Goldberger, A.S. (1961), "On Pure and Mixed Statistical Estimation in Economics," *International Economic Review*, 2, 65–78.

Theobald, C.M. (1974), "Generalizations of Mean Square Error Applied to Ridge Regression," *Journal of the Royal Statistical Society*, Ser. B, 36, 103–106.

Comment

RONALD A. THISTED*

1. INTRODUCTION

In their critique Smith and Campbell mount a spirited attack on some basic ridge regression techniques, and in this assault I find myself caught on middle ground. I am able neither to accept their arguments without reservation nor to defend wholeheartedly current ridge practice. The issues Smith and Campbell raise concern the foundations of ridge regression, which have never been adequately examined. It is true that some ridge practices rest on shallow footings that can easily be undermined, as I demonstrate later. But certain structures seem to me to be solid despite Smith and Campbell's objections, and their analysis does not justify the conclusion that ridge methods have no utility.

In some respects I agree with Smith and Campbell. I am quite sympathetic with their call for explicit

* Ronald A. Thisted is the Leonard Jimmie Savage Assistant Professor in the Department of Statistics and the College, University of Chicago, 5734 University Ave., Chicago, IL 60637. Support for this research was provided by the National Science Foundation under Grant No. MCS76-81435. This article contains comments presented at the *JASA* Theory and Methods Invited Paper session of the 139th annual meeting of the American Statistical Association in Washington, D.C., on August 16, 1979.

© Journal of the American Statistical Association
March 1980, Volume 75, Number 369
Theory and Methods Section