The Canadian Journal of Statistics
Vol. 8, No. 2, 1980, Pages 217–225
La Revue Canadienne de Statistique

An example of ridge regression difficulties

Gary SMITH*

Pomona College, Claremont, California

Key words and phrases: Ridge regression, multicollinearity, principal components.
AMS 1980 *subject classifications:* Primary 62J07; secondary 62J05, 62H12.

ABSTRACT

A simple consumption function is used to illustrate two fundamental difficulties with ridge regression and similarly motivated procedures. The first is the ambiguity of multicollinearity measures for judging the data's "ill-conditioning". The second is the sensitivity of the estimates to the arbitrary normalization of the model. Neither of these poses a problem for least squares or Bayesian estimates. The logical restructuring of ridge procedures to avoid these difficulties leads to a more explicitly Bayesian approach.

1. INTRODUCTION

Economic data are seldom rich enough to decisively answer the questions that economists pose. There are frequently only a few relevant observations and in these "nature experiments" the independent variables often display little variation and/or high covariation. This seems to be a permanent part of economics, and over the years researchers have proposed and tried a wide variety of methods for coping with inadequately informative data.

Ridge regression is a recent remedy that has attracted a great deal of attention (e.g., Hoerl and Kennard 1970, Marquardt 1970, Mayer and Wilke 1973, and Theobald 1974). Ridge procedures seem motivated by the 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 parameter estimates toward zero. Although flexibility is provided by the abstention from exact exclusion restrictions, Smith and Campbell (1980) have argued that 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 arbitrary labeling of nonorthogonal data as weak, and a loose representation of *a priori* beliefs and reliance at times on *ad hoc* pseudo information.

Any standard regression model

$$\underset{t\times 1}{\mathbf{y}} = \underset{t\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{t\times 1}{\boldsymbol{\varepsilon}} \tag{1.1}$$

can always be rewritten as

$$\mathbf{y} = (\mathbf{XA})(\mathbf{A}^{-1}\boldsymbol{\beta}) + \boldsymbol{\varepsilon} = \mathbf{Zy} + \boldsymbol{\varepsilon} \tag{1.2}$$

using any nonsingular matrix **A**.

* Research completed while at the University of Houston, Central Campus.

The application of a simple ridge estimator to (1.2),

$$\hat{\gamma} = (\mathbf{Z}'\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}'\mathbf{y} \tag{1.3}$$

is equivalent to use of a generalized ridge estimator for (1.1)

$$\hat{\boldsymbol{\beta}} = \mathbf{A}\hat{\gamma} = (\mathbf{X}'\mathbf{X} + k(\mathbf{A}\mathbf{A}')^{-1})^{-1}\mathbf{X}'\mathbf{y}. \tag{1.4}$$

From a Bayesian (Lindley and Smith 1972) or classical mixed estimation (Theil and Goldberger 1961) perspective, these estimators are appropriate for prior information of the specific form

$$\gamma \sim \mathcal{N}[\mathbf{0}, (\sigma_\epsilon^2/k)\mathbf{I}] \Leftrightarrow \boldsymbol{\beta} \sim \mathcal{N}[\mathbf{0}, (\sigma_\epsilon^2/k)\mathbf{A}\mathbf{A}']. \tag{1.5}$$

That is, the prior distributions for the γ_i are independent with common variances and centered at the origin.

Thus, there is a simple direct theoretical interpretation of ridge regression. With a Gaussian prior distribution, a model (1.1) can always be transformed to (1.2) so that ridge regression is appropriate. In practice the primary inadequacy of ridge regression is that the appropriateness of the implicit priors (1.5) is never addressed. There is inevitably a transformation of the initial model, but it is based upon the sample characteristics of $\mathbf{X}'\mathbf{X}$ rather than *a priori* notions about $\boldsymbol{\beta}$. Smith and Campbell (1980), however, argued that the nature of $\mathbf{X}'\mathbf{X}$ cannot reasonably be used to decide whether or not to use a ridge procedure, nor to determine the transformation to which it should be applied. If ridge estimates are calculated with $\mathbf{A}$ chosen by some *ad hoc* method, then they will be largely arbitrary and never optimal.

The present paper aims to clarify and reinforce these arguments with a detailed illustrative analysis of a simple and familiar model.

2. THE EXAMPLE

The variables are:

C = real per capita consumption

Y = real per capita disposable income

Y_P = real per capita permanent income

Y_T = real per capita transitory income, $Y - Y_P$.

For simplicity, we will assume that the sample means have been subtracted from each variable. Such "centering" seems to be standard ridge procedure although ridge estimates are, unfortunately, sensitive to this arbitrary decision. Brown (1977) gives a computational procedure for obtaining centered estimates without actually centering the data. However, this evades the basic question of the desirability of centered estimates; i.e., whether the intercept should be shrunk and if so towards what point.

The assumed sample data are

$$\frac{1}{n-1}\begin{bmatrix}\mathbf{C}'\\ \mathbf{Y}'\\ \mathbf{Y}_P'\\ \mathbf{Y}_T'\end{bmatrix}[\mathbf{C}\ \mathbf{Y}\ \mathbf{Y}_P\ \mathbf{Y}_T] = \begin{bmatrix}7.3 & 8.3 & 8 & 0.3\\ 8.3 & 10 & 9 & 1\\ 8 & 9 & 9 & 0\\ 0.3 & 1 & 0 & 1\end{bmatrix}.$$

A plausible model is that consumption is a linear function of both current income

and also longer run permanent income,

$$\mathbf{C} = \alpha\mathbf{Y} + \beta\mathbf{Y}_P + \boldsymbol{\varepsilon}. \tag{2.1}$$

With the data given above, the correlation between $\mathbf{Y}$ and $\mathbf{Y}_P$ is 0.95. Does this indicate a collinearity problem and a need for ridge regression or some other procedure advertised to pacify ill-conditioned data? Invariably, ridge users gauge the strength of the data by the intercorrelations among the explanatory variables as measured by correlation coefficients or eigenvalues. But these are incomplete measures of the precision of the estimates (see Smith 1980).

The variance of the least squares estimate of γ_i is

$$\mathcal{V}\hat{a}r(\hat{\gamma}_i) = \frac{\sigma_\varepsilon^2}{(n-1)\sigma_i^2}\frac{1}{1-R_i^2},$$

where σ_i^2 is the sample variance of the variable associated with γ_i, and R_i^2 is the squared multiple correlation coefficient between this variable and the remaining explanatory variables. A high squared correlation R_i^2 can help explain why an estimate is imprecise, but it cannot be used alone to measure the scale of the imprecision.

If for instance $\sigma_\varepsilon^2 = 0$, then all of the parameters would be estimated without error even if the explanatory variables were almost perfectly correlated. In practice, σ_ε^2 will rarely be zero but the strength of the data cannot be assessed without reference to it, the number of observations, and the variance of the associated explanatory variables.

The precision of a particular estimate also has to be measured against the scale of the parameter and its intended use. A given standard deviation of say 0.3 might mean that the point estimate is much too vague or may be quite satisfactory if the parameter is very large or of little importance. The only scalar measure of overall data strength that comes to mind is the estimated mean squared forecasting error of y for those values of $\mathbf{X}$ for which the model will be used. This, of course, depends upon considerably more than the sample intercorrelations among the explanatory variables.

For the assumed data (2.1), $\hat{\sigma}_\varepsilon^2 = 0.1$ and the least squares estimates are

$$\hat{C} = \underset{(0.32)}{0.30}\ Y + \underset{(0.33)}{0.59}\ Y_P.$$

The estimated standard deviations are displayed in parentheses. For most applications of the model, these standard deviations would probably be considered large. One reason for the imprecision is that Y and Y_P are highly correlated. If Y and Y_P were uncorrelatetd, and the sample data otherwise unchanged, the estimated standard deviations would only be 0.10 and 0.11. However, it cannot be inferred simply from the correlation between Y and Y_P that the estimates are imprecise. The standard deviations would also be one-third their present size if (for a given 0.95 correlation between Y and Y_P) either the number of observations or the variances of Y and Y_P were increased tenfold or σ_ε^2 reduced by a factor of ten.

The inadequacy of measuring imprecision solely by R_i^2 (or some equivalent) is clearly shown by simple innocent transformations of (2.1) based upon $Y = Y_P + Y_T$,

$$C = (\alpha + \beta)Y - \beta Y_T + \varepsilon, \tag{2.2}$$

$$C = \alpha Y_T + (\alpha + \beta)Y_P + \varepsilon. \tag{2.3}$$

These are entirely equivalent to (2.1). But in form (2.2), the correlation among the explanatory variables is 0.32, and in (2.3) it is 0. The least squares estimates are no more precise in (2.2) or (2.3) than in (2.1); indeed they coincide.

One change is that different parameters are explicitly estimated. In (2.2) and (2.3), $\alpha + \beta$ is directly estimated and, because of the positive correlation between Y and Y_P, $\alpha + \beta$ is indeed more accurately estimated than either α or β. But an arbitrary choice of which parameters to estimate directly and which indirectly should not influence a decision to abandon least squares for ridge regression.

The other change in (2.2) and (2.3) is that Y_T appears explicitly. The imprecision in the estimates of the associated parameters is then attributed to the low variance of Y_T rather than as in (2.1) to the high correlation between Y and Y_P. Looking at α and comparing (2.3) with (2.1), for instance, $1/(1 - R_i^2)$ has been reduced by a factor of 10, but so has the variance of the variable associated with α. The estimate of α and its standard deviation are unchanged. In (2.1) there is a collinearity problem; in (2.3) there is insufficient variation in Y_T. These are equivalent descriptions. A high correlation between Y and Y_P means that there is little variation in $Y_T = Y - Y_P$. In general, high covariation can always be restated as low variation. And again a decision to use ridge regression should not be based upon an arbitrary habit of speech.

The second fundamental difficulty is deciding which ridge procedure to use. The application of a simple ridge estimator to (2.1), (2.2), or (2.3) will, as indicated by (1.3), (1.4), and (1.5), impose different implicit priors and result in different estimates. For example, ridge estimates of (2.1) implicitly assume equal prior variances on α and β. With (2.2), however, the implicit prior variance on α is twice that for β; with (2.3), α's variance is half that of β. In (2.1), the correlation between the implicit priors on α and β is zero; in (2.2) and (2.3) it is -0.7.

The differences in the implicit priors invoked by ridge estimates of (2.1), (2.2) and (2.3) relate to the covariance structure. All assume prior means of zero. But there seem to be no compelling reasons for zero means. Indeed, the explanatory variables were presumably chosen precisely because they were expected to have some effect on the dependent variable. In addition, the researcher might just as well have chosen other arbitrary arrangements of the model for which a ridge procedure would have mechanically used nonzero means for α and β. For instance, he might have acquired the habit of specifying saving rather than spending functions, say

$$S = (1 - \alpha)Y - \beta Y_P + \varepsilon, \tag{2.4}$$

$$S = (1 - \alpha - \beta)Y + \beta Y_T + \varepsilon, \tag{2.5}$$

and

$$S = (1 - \alpha)Y_T + (1 - \alpha - \beta)Y_P + \varepsilon, \tag{2.6}$$

where $S = Y - C$. Least squares will correctly recognize that these are fully equivalent to (2.1), (2.2), or (2.3). Ridge estimates of (2.4), (2.5), or (2.6) will, however, implicitly assume a prior mean for α of one, a zero mean for β, and again restructure the covariances.

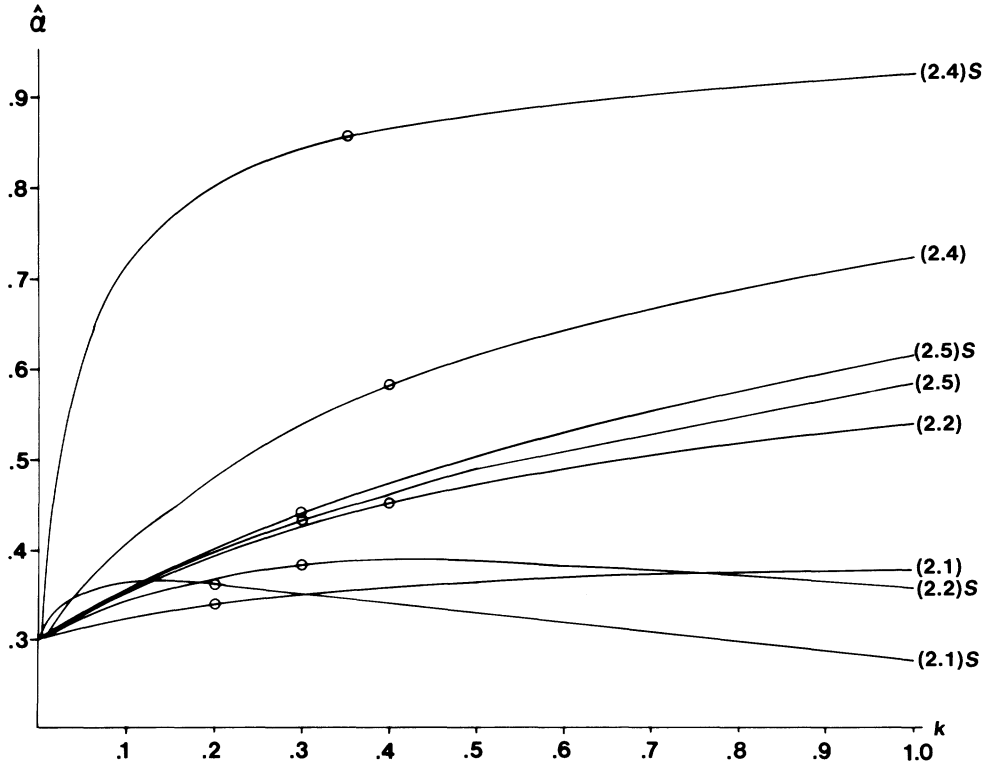
There is no substantive difference to representations (2.1) through (2.6). Any estimation procedure which produces different estimates for such transformed data is fundamentally flawed. The only theoretically legitimate way to choose between ridge estimates in (2.1) through (2.6) is to decide which implicit prior information is

more apt. But there are, of course, an infinite variety of other possible transformations, and a choice between all possiblilities is really just an awkward way of eliciting prior information for applying a Bayesian or mixed estimation procedure. This is quite different from standard ridge methods, and there is no reason to use the additional ridge label.

To illustrate the sensitivity of ridge estimates to transformations of a model, we calculated ridge estimates of representations (2.1), (2.2), (2.4), and (2.5). Representations (2.3) and (2.6) were neglected because practitioners consider ridge estimates unnecessary for orthogonal data. For comparative purposes, it can be assumed that the ridge estimates of (2.3) and (2.6) would be the least squares estimates. Tongue firmly in cheek, Leamer (1976) has suggested that ridge clients unlucky enough to select orthogonal data can use "valley regression", supplementing the sample data with a matrix containing zeroes on the diagonal and k's off-diagonal. He shows that, as with ridge regression, there is some k for which average mean squared error is less than with least squares. Since there do not seem to presently be many valley enthusiasts, we have not examined any valley estimates in our models.

Many ridge users standardize their data by dividing each variable by its standard deviation. This arbitrary and innocent transformation affects the ridge results. If with unstandardized data it is implicitly assumed that the prior variances on β_i are equal, then with standardized data it is implicity assumed that the prior variances on $\sigma_i\beta_i$ are equal; i.e., that the variance of β_j relative to β_k is equal to the variance of X_k relative to X_j. Different implicit priors, of course, yield different estimates. We have,

FIGURE 1: Implicit ridge estimates of α.



consequently, also calculated ridge estimates for standardized versions of (2.1), (2.2), (2.4), and (2.5).

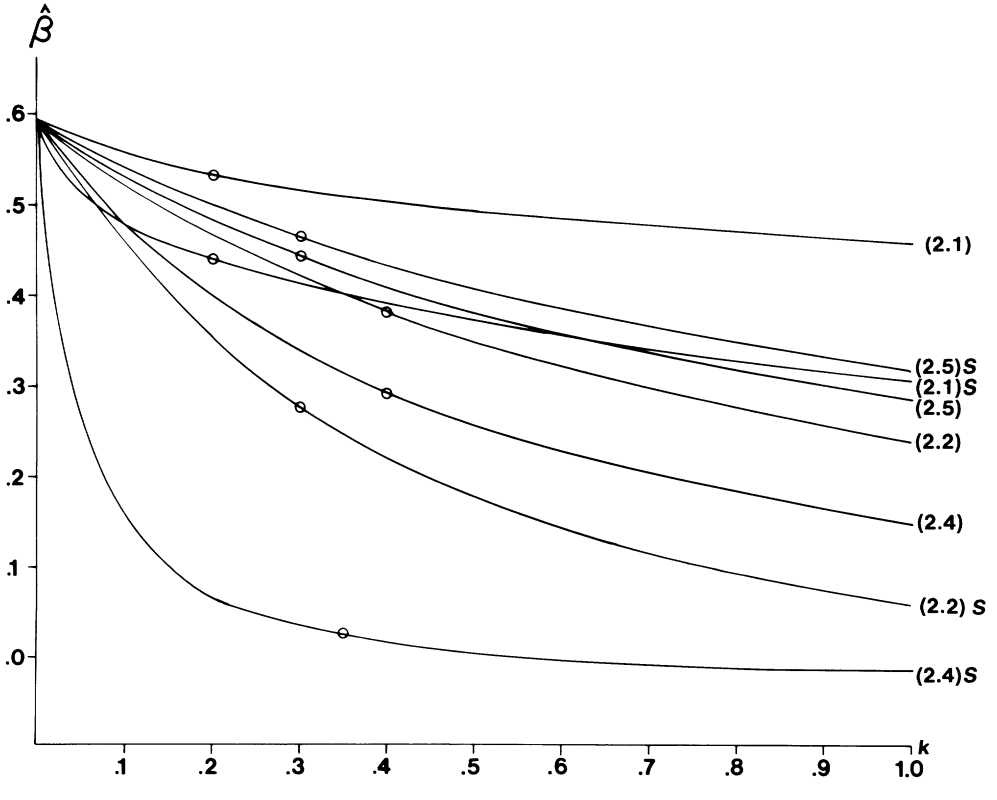
The eight ridge traces were sent to a ridge enthusiast, who was promised anonymity and who selected values of k for which the traces stabilized. This practitioner was not told that the traces all related to the same model, and indeed did not even ask the nature of the data being analyzed.

We then calculated the implicit ridge estimates of α and β for different values of k, and these are displayed in Figures 1 and 2. The number labels on these transformed traces correspond to the equation numbers in the text. The appearance of an "S" denotes the use of standardized data. The circled points on these traces are the implicit point estimates using the values of k selected by the ridge enthusiast from the ridge traces for the explicitly estimated parameters.

There is enormous variety in the implicit ridge traces; arbitrary rearrangements of the variables produce strikingly different parameter estimates. Even the simple scaling of the variables by their standard deviations has a considerable effect. Again, the question has to be asked, how can one logically choose among these varied estimates, or the many other possibilities? It hardly seems satisfactory to select a representation and the associated parameter estimates arbitrarily. The logical alternative of using *a priori* information leads away from ridge traces to a straightforward and explicitly Bayesian approach.

Since ridge traces seem a bit mystical, even to some ridge users, a number of more mechanical rules for choosing k have been proposed. For example, Hoerl, Kennard,

FIGURE 2: Implicit ridge estimates of β.



and Baldwin (1975) argue that "a reasonable choice for an automatic selection of k" in (1.3) is an estimate of $p\sigma_\varepsilon^2/\gamma'\gamma$. For this purpose, they use the least squares estimates of σ_ε^2 and $\gamma'\gamma$. (The usefulness of their reported simulations is lessened by their use of parameter values which conform to the implicit ridge priors (1.5).) Tables 1 and 2 contain the implicit estimates of α and β using the Hoerl, Kennard, and Baldwin rule. The estimates are again sensitive to the explicit representation of the model, though not quite so much as when the ridge trace is used to select k. Since many ridge advocates (e.g., Hoerl and Kennard 1970) emphasize the excessive length of the least squares estimates of γ, it might be supposed that they would prefer larger estimates of $p\sigma_\varepsilon^2/\gamma'\gamma$ than the least squares values. If larger values of k are in fact used, this would increase the dispersion of the estimates in Tables 1 and 2.

Ridge procedures are not unique in their sensitivity to renormalizations of the data. It is a widespread practice to delete variables whose coefficients are statistically insignificant. The outcome of this procedure is sensitive to which coefficients are subjected to such tests. If it is the explicitly estimated parameters which are tested, then the final estimates will depend heavily upon the initial normalization. The final columns in Tables 1 and 2 illustrate this for the model here.

The transformation to and deletion of principal components is similarly sensitive.

TABLE 1: Estimates of α.

Equation	k from trace	$k = 2\sigma_\varepsilon^2/\gamma'\gamma$	Principal components: Low λ	Principal components: $t < 2$	Delete variable with $t < 2$
(2.1)	0.34	0.36	0.45	0.45	0
(2.1)S	0.36	0.35	0.43	0.43	0
(2.2)	0.45	0.38	0.90	0.90	0.83
(2.2)S	0.38	0.31	1.46	0.30	0.83
(2.3)	0.30	0.24	0	0	0
(2.3)S	0.30	0.29	0.30	0.30	0
(2.4)	0.58	0.51	0.93	0.38	0.83
(2.4)S	0.86	0.51	0.93	0.37	0.83
(2.5)	0.44	0.50	0.81	0.81	0.30
(2.5)S	0.43	0.48	0.38	0.38	0.30
(2.6)	0.30	0.50	1.00	0.30	0.30
(2.6)S	0.30	0.48	0.30	0.49	0.30

TABLE 2: Estimates of β.

Equation	k from trace	$k = 2\sigma_\varepsilon^2/\gamma'\gamma$	Principal components: Low λ	Principal components: $t < 2$	Delete variable with $t < 2$
(2.1)	0.54	0.50	0.43	0.43	0.89
(2.1)S	0.44	0.51	0.45	0.45	0.89
(2.2)	0.38	0.48	−0.09	−0.09	0
(2.2)S	0.28	0.55	−1.11	0.59	0
(2.3)	0.59	0.62	0.89	0.89	0.89
(2.3)S	0.59	0.57	0.59	0.59	0.89
(2.4)	0.29	0.37	−0.07	0.66	0
(2.4)S	0.03	0.37	−0.07	0.66	0
(2.5)	0.44	0.37	0.02	0.02	0.70
(2.5)S	0.47	0.43	0.47	0.47	0.70
(2.6)	0.59	0.39	−0.11	0.70	0.70
(2.6)S	0.59	0.44	0.59	0.34	0.70

Kendall (1965) has recommended the deletion of components with relatively low eigenvalues. This entails the setting equal to zero of certain parameters (linear combinations of the initial parameters) whose estimates have relatively large variances. If any one of these estimates is itself large, then zero may actually be outside a conventional confidence interval. Massy (1965) has consequently recommended the alternative procedure of deleting components whose coefficients are statistically insignificant. In either approach the initial normalization determines the parameters subjected to such scrutiny and thus affects the final estimates. Tables 1 and 2 display the implicit estimates of α and β for normalizations (2.1)–(2.6). The normalization and component deletion rules are important. Indeed, these principal components estimates show considerably greater variation than do the ridge estimates.

3. SUMMARY

A simple familiar model was used to illustrate two fundamental difficulties with ridge regression and similarly motivated procedures. The first is the ambiguity of multicollinearity measures for judging the data's "ill-conditioning". The second is the sensitivity of the estimates to the arbitrary normalization of the model. Neither of these poses a problem for least squares or Bayesian estimates. The logical restructuring of ridge procedures to avoid these difficulties leads to a more explicitly Bayesian approach.

RÉSUMÉ

Une simple fonction de consommation est employée pour illustrer deux difficultés fondamentales qui surgissent dans la régression "ridge" et d'autres procédures de même source. La première est l'ambiguïté des mesures de multicolinéarité employées pour évaluer la "mauvaise condition" des données. La seconde est la sensibilité des estimations à la normalisation arbitraire du modèle. Ces difficultés ne se présentent pas dans le cas des estimateurs bayesiens ou des estimateurs des moindres carrés. La restructuration logique des procédures "ridge" mène à une approche dont le caractère bayesien est plus explicite.

REFERENCES

Brown, P.J. (1977). Centering and scaling in ridge regression. *Technometrics*, 19, 35–36.

Hoerl, Arthur E., and Kennard, Robert W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

Hoerl, Arthur E., and Kennard, Robert W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12, 69–82.

Hoerl, Arthur E.; Kennard, Robert W., and Baldwin, Kent F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4, 105–123.

Kendall, M.G. (1965). *A Course in Multivariate Statistical Anaalysis.* Third Edition. Griffin, London.

Leamer, Edward (1976). Valley regression: Biased estimation for orthogonal problems. Unpublished manuscript.

Lindley, D.V., and Smith. A.F.M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B*, 34, 1–18.

Marquardt, Donald W. (1970). Generalized inverses, ridge regression, biased linear estimation. *Technometrics*, 12, 591–612.

Massy, William F. (1965). Principal components in exploratory statistical research. *J. Amer. Statist. Assoc.*, 60, 234–256.

Mayer, Lawrence S., and Wilke, Thomas A. (1973). On biased estimation in linear models. *Technometrics*, 15, 497–508.

Smith, Gary (1980). Can multicollinearity be meaningfully measured? Unpublished manuscript.
Smith, Gary, and Campbell, Frank (1980). A critique of some ridge regression methods. (With discussion and rejoinder.) *J. Amer. Statist. Assoc.*, 75, 74–103.
Theil, H., and Goldberger, A.S. (1961). On pure and mixed statistical estimation in economics. *Internat. Econom. Rev.*, 2, 65–78.
Theobald, C.M. (1974). Generalizations of mean square error applied to ridge regression. *J. Roy. Statist. Soc. Ser. B*, 36, 103–106.

Received 15 January 1980
Revised 7 May 1980

Department of Economics
Pomona College
Claremont, California 91711, U.S.A.