Proc. Aust. Soc. Anim. Prod. (1974) **10**; 379

ANALYSIS OF HIGHLY CORRELATED ANIMAL PRODUCTION VARIABLES

H. JEFFERY*, G.T. McKINNEY†, AND J.B. COOMBE†

Summary

Multiple regression analysis of data in which the independent variables are highly correlated frequently results in unstable estimates of the regression coefficients. The consequences of this instability can be that the analysis is of limited interpretative value and yields an equation which is a poor predictor when applied to observations other than those used in the analysis.

A technique, ridge regression, has recently been proposed to provide more stable estimates of multiple regression coefficients. This technique was compared with a stepwise procedure in the analysis of a number of sets of sheep metabolism data. The ridge regression procedure produced more consistent estimates of the coefficients and generally resulted in more accurate prediction equations. It is concluded that ridge regression analysis may be of considerable value in agricultural research.

I. INTRODUCTION

If a multiple regression equation is determined from a set of independent variables that are highly correlated the usual least squares estimates of the partial regression coefficients so obtained can be unstable (Snedecor and Cochran 1967). These estimates are unbiassed, but have a large variance.

Because of this large variance little biological meaning can be given to the estimates of the coefficients and the regression may be of limited interpretative value. Further, a predictive equation developed from one set of data will be a poor predictor when applied to other sets of similarly obtained data.

Recently a technique called "ridge regression" has been proposed in which, by the introduction of a small bias into the estimates of the regression coefficients, the variance of these coefficients can be greatly reduced (Hoerl and Kennard 1970a, 1970b). Ridge regression is aimed at obtaining regression coefficients with low mean square error, i.e. $(\text{bias})^2 + \text{variance}$, rather than maximising the coefficient of determination of the regression (R^2).

Generally the importance of different independent variables in a multiple regression equation cannot be judged by their so called "β coefficients". A simple linear transformation of a variable can alter its coefficient but will not change its importance. To directly compare variables it is necessary for them to be reduced to the same scale of measurement; this is usually achieved by "standardizing" the variables, i.e. dividing each observation by its standard deviation. The coefficients estimated in regressions calculated on standardized variables are termed standard partial regression coefficients (Snedecor and Cochran 1967), hereafter called a coefficients.

The theoretical considerations upon which ridge regression has been developed have been explained in detail by Hoerl and Kennard (1970a); a brief description follows. If the independent variable correlation matrix is called $\underline{R}$ and the dependent variable correlation vector is called $\underline{g}$, then the least squares estimate, $\underline{\hat{a}}$, of the vector of a values is given by

---

* On study leave from N.S.W. Department of Agriculture with the Division of Plant Industry, CSIRO, Canberra and Department of Environmental Biology, Australian National University, Canberra.

† Division of Plant Industry, CSIRO, Canberra.

$$\underline{\hat{\alpha}} = \underline{R}^{-1}\underline{g} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Ridge estimates of $\underline{\alpha}$, $\underline{\hat{\alpha}}^*$, are obtained from the following relation.

$$\underline{\hat{\alpha}}^* = (\underline{R} + K\underline{I})^{-1}\underline{g} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

where I is the identity matrix and K some non-negative constant. If K = 0 then equation (2) is identical to equation (1). Hoerl and Kennard (1970a) demonstrated that as K increases the bias of $\underline{\hat{\alpha}}^*$ increases but its variance decreases. Thus the aim of ridge regression is to select a K value which causes a large decrease in variance of $\hat{\alpha}^*$ without introducing too serious a bias.

This paper demonstrates the use of ridge regression analysis with data obtained from a sheep metabolism experiment. In particular, the results of the ridge regression analyses are compared with those obtained by stepwise regression using the "stepdown procedure" (Snedecor and Cochran 1967).

II. METHODS

(a) Data and model

The data were obtained from an experiment reported by Coombe, Christian and Holgate (1971) in which seven groups of adult Merino wethers were fed for 16 weeks on different diets of pelleted oat straw and urea, with or without mineral supplementation. Metabolic studies were conducted on four animals from each group. A detailed description of the methods used in the experiment can be found in the above publication.

Data used in the following analyses were obtained from the 16 sheep whose diet remained unchanged throughout the experiment. The variables used in the analysis were

Y = nitrogen balance (g/sheep/day)
X_1 = nitrogen intake (g/sheep/day)
X_2 = nitrogen digestibility (%)

These variables were selected as nitrogen balance is a difficult variable to measure and if it could be predicted with reasonable accuracy from the more simply measured variables (Xl and X2) then this could be useful in the evaluation of feeds. The independent variables chosen are measured in the determination of nitrogen balance and so also is urinary nitrogen loss. Use of the predictive equation would thus remove the need to measure urinary nitrogen loss for the estimation of nitrogen balance.

Data were fitted to the following statistical model with variables being deleted if they were not significant,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

The model thus allowed for both curvature and "interaction" by the inclusion of quadratic terms and the cross product term respectively.

(b) Analyses

Twenty-five sets, each of 10 observations, were obtained by randomly removing the data of six sheep from the total of 16 observations. Each of these sets were analyzed by both the stepdown multiple regression procedure (Snedecor and Cochran 1967), hereafter called the "t-test method", and ridge regression techniques.

In the t-test method the variable with the lowest absolute t-value was deleted from the analysis if this value was less than 1.5. A new regression was then calculated and the above procedure repeated until the absolute t-values of all variab was greater than 1.5. With ridge regression analyses the importance of different variables was assessed when their values were stabilized and then the least important variables deleted. Details of this decision algorithm will be published, in full, elsewhere (Jeffery and McKinney, unpublished).

The predictive power of the regressions established within each set of 10 observations was judged by comparing the sum of squared deviations between the predicted and actual values for the six remaining observations (D2).

III. RESULTS

The correlations that were found between the different variables for the total of 16 observations are presented in Table 1. High correlations can be seen to exist between the respective linear and quadratic terms and between the interaction term and the linear and quadratic forms of X_1.

TABLE 1

Simple correlations between variables for the full set of data

Variable	X_1	X_2	X_1^2	X_2^2	X_1X_2	Y
X_1	1.000					
X_2	.277	1.000				
X_1^2	.900	.300	1.000			
X_2^2	.279	.999	.302	1.000		
X_1X_2	.983	.446	.980	.448	1.000	
Y	.535	.016	.540	.019	.506	1.000

Neither technique consistently chose the same predictive variables. Final regressions from the ridge method contained at most two variables and each coefficient was stable when K = 0 (least squares fit). On three occasions the t-test method yielded predictive equations in four variables and these regressions were highly unstable (D2 values were 13.6, 19.1 and 41.1). On 10 occasions all t-statistics in the regression exceeded 2.00. For these regressions D^2 values ranged from 2.4 to 165.2 and in 6 cases exceeded 10.

The distribution of D^2 in Table 2 clearly shows the relatively high probability of aberrant predictions that resulted from equations determined by the t-test method. The maximum D^2 obtained with ridge regression was 13.1 whereas 10 regressions selected by the t-test method had D2 values in excess of 13.1, the maximum value being 165.2.

TABLE 2

Comparison of the frequency distribution of the sum of squared deviations (D^2) between predicted and actual values for the two methods of analysis

D^2 / Analysis	less than 10	10 to 20	20 to 30	greater than 30
t-test method	10	10	2	3
ridge method	21	4	0	0

On six occasions equal D^2 values were obtained by the different methods, i.e. identical regressions were calculated. In two cases out of the 25 the D^2 from the t-test method was lower than that from the ridge analysismethod.

From Table 3 it can be seen that the estimates of the coefficients were more consistent when obtained by the ridge method. For example, the ridge estimates of the coefficient of Xl varied between .117 and .251 whereas the t-test method estimates ranged from -9.14 to 14.91. Within each variable all estimates of the

coefficients obtained by ridge analysis had the same sign whereas the t-test method estimates were never consistently positive or negative.

TABLE 3
Comparison of the coefficients of variation of the significant coefficients obtained from the two methods

Variable \ Method	t-test	ridge
X_1	1201	24
X_2	682	15
X_1^2	296	17
X_2^2	559	37
X_1X_2	299	23

IV. DISCUSSION

In the example considered, ridge regression provided estimates of regression coefficients that were more consistent and provided better predictive equations than did the stepdown procedure. The estimates obtained using ridge technqiues were also within a meaningful range; this did not often apply to the t-test estimates. For example , β_1 = 14.91 can have no biological meaning as it suggests that for every g increase in nitrogen intake the nitrogen balance is increased by 14.91 g. In any case, the high variability (Table 3) of the t-test estimates could not permit anymore than a crude interpretation of effects.

The low correlations between Y and both X2 and $X2^2$ may suggest that these variables should not be considered in a predictive equation. However, the correlations between Y and X2 varied from -.66 to .39 in the individual sets of data. Thus it was by no means clear that X_2 or X_2^2 would not be important predictive variables. In any case, the existence of a low correlation between a dependent and independent variable does not imply that the independent variable will be of no predictive value when incorporated in a regression with other variables.

The most significant aspect of the analyses was the distribution of D^2 (Table 2). Clearly if the intent of regression analysis is to obtain a predictive equation then a technique which (a) lowers the average D^2, and (b) is unlikely to produce very high D^2 values; is to be preferred. Where the penalty for poor prediction is greatthen point (b) increases in importance.

In conclusion, the major aim of this paper is to inform research workers of a relatively new analytic technique. We feel that it is useful where reliable, meaningful prediction equations are required from highly correlated data. Ridge analysis is by no means a panacea, but it is a most useful additional analytic-tool.

V. ACKNOWLEDGEMENT

One of us (H.J.) was in receipt of a C.E.S.G. Postgraduate study award.

VI. REFERENCES

COOMBE, J.B., CHRISTIAN, K.R., and HOLGATE, M.D. (1971). J. agric. Sci., Camb. 77: 1

HOERL, A.E., and KENNARD, R.W. (1970a). Technometrics 12:55.

HOERL, A.E., and KENNARD, R.W. (1970b). Technometrics 12: 69.

SNEDECOR, G.W., and COCHRAN, W.G. (1967). "Statistical Methods" 6th Ed. (Iowa State University Press: Ames).