J. R. Statist. Soc. B (1993)
55, *No.* 3, *pp.* 653–659

Continuum Regression and Ridge Regression

By ROLF SUNDBERG†

Stockholm University, Sweden

[Received December 1991. Revised May 1992]

SUMMARY

We demonstrate the close relationship between first-factor continuum regression and standard ridge regression. The difference is that continuum regression inserts a scalar compensation factor for that part of the shrinkage in ridge regression that has no connection with tendencies towards collinearity. We interpret this to mean that first-factor continuum regression is preferable in principle to ridge regression if we want protection against near collinearity but do not admit shrinkage as a general principle. Furthermore, our experience indicates that with first-factor continuum regression we can obtain predictors that are at least as mean-squared error efficient as with ridge regression but with less sensitivity to the choice of ridge constant. The scalar compensation factor is easily calculated by just an additional simple linear regression with the ridge regression predictor as regressor.

Keywords: CROSS-VALIDATION; NEAR COLLINEARITY; PARTIAL LEAST SQUARES; SHRINKAGE ESTIMATORS

1. INTRODUCTION

For multiple linear regression with non-orthogonal regressors there are several alternatives to the ordinary least squares (OLS) method which are advantageous when the regressors are near collinear. Examples are partial least squares (PLS), principal components regression (PCR) and ridge regression (RR). An important step towards the understanding of OLS, PLS and PCR and their interrelationships was the concept of continuum regression (CR), proposed by Stone and Brooks (1990). In CR we regard OLS, PLS and PCR as special cases corresponding to three different values of a 'parameter', with OLS at one extreme and PCR at the other. Fearn (1990) in the discussion of Stone and Brooks (1990) remarks on the similarity between first-factor CR and standard RR but does not explain fully. In this paper we extend Fearn's discussion and answer his demand 'the general question of when the two methods are similar would bear further investigation'.

We shall make the observation that RR may be regarded as bringing in two shrinkage effects: one to compensate for near collinearity in the regressors and the other to reduce the mean-squared error (MSE) by replacing variance by squared bias without any connection with collinearity or regression. First-factor CR with a parameter between OLS and PLS will be seen to correspond to the collinearity compensating aspect of RR. This motivates a modification of conventional RR so that it will correspond to the special case of CR. This modification will be preferable to conventional RR for statisticians who do not like shrinkage estimators as a principle, but it will also be seen in an example to have the advantage of less sensitivity to the choice of ridge constant.

†*Address for correspondence*: Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, S-106 91 Stockholm, Sweden.

 0035–9246/93/55653 $2.00

2. RELATIONSHIP BETWEEN CONTINUUM REGRESSION AND RIDGE REGRESSION

We want to explain or predict a response variable y by a linear function $a + b'x$ of a p-dimensional regressor vector x (where the prime denotes transpose). We have n observations (x_i, y_i) which enable us to construct estimates for the coefficients a and b. We assume that both y and x have been centred. This implies that all methods to be discussed in this paper will use $a = 0$, so the discussion can focus on the choice of estimate for b.

OLS tells us to use

$$b^{\mathrm{OLS}} = S^{-1}s,$$

where $S = X'X$ and $s = X'y$ are the usual sum of products matrix and vector respectively. When S is nearly singular b^{OLS} will have undesirable properties, being extremely sensitive to small changes in s.

In *standard RR* (Hoerl and Kennard, 1970) we replace S by a better conditioned matrix, $S + \delta I$ for a (typically small) positive coefficient δ, called the ridge constant, i.e.

$$b^{\mathrm{RR}}(\delta) = (S + \delta I)^{-1}s.$$

We shall not discuss methods for the choice of value of δ, but rather regard $b^{\mathrm{RR}}(\delta)$ as a class of estimators. In standard RR it is normally also assumed that the x-components are scaled so that S is of correlation form, but we do not require such an assumption here.

In the method of *CR* (Stone and Brooks, 1990) a number of regressors are selected. The first to be chosen is given by the coefficient vector $c = c(\gamma)$ that maximizes the function

$$T = (c's)^2(c'Sc)^{\gamma - 1} \tag{2.1}$$

for a given 'CR parameter' $\gamma \geqslant 0$ ($= \alpha/(1-\alpha)$ in their alternative parameterization) and for given length $\|c\|$ of c, $c'c = 1$ say. If the construction is terminated here, the linear form $c'x_i$ is used as a regressor in a one-variable OLS. For $\gamma = 0$, the maximization of T is equivalent to the maximization of the sample correlation coefficient between y and $c'x$, with OLS as solution,

$$c(0) \propto S^{-1}s. \tag{2.2}$$

For $\gamma = 1$ the covariance between y and $c'x$ is maximized, and this yields the first latent factor of the PLS method. As $\gamma \to \infty$ we approach PCR, selecting as first regressor the form $c'x$ such that $c'Sc$ is maximized (under unit length of c), i.e. c is the eigenvector of S corresponding to the highest eigenvalue.

We shall later restrict our consideration to this first factor (first regressor) of CR, but further regressors are selected by analogous maximization of T under the constraints that the next c-vector to be chosen should be uncorrelated with all previously chosen c-vectors. The variable y is then regressed on the forms $c'x$ by OLS multiple regression. As a stopping rule for the number ω of CR regressors, and for the choice of γ, Stone and Brooks (1990) propose cross-validation, using a cross-validatory index $I_{\gamma,\omega} \leqslant 1$ ($= 1$ if the regression fit is perfect). In some illustrative examples $I_{\gamma,\omega}$ is shown plotted against γ for various values of the number ω. In several

of these diagrams we see a form of the $I_{\gamma,1}$ curve that is typical for RR, namely a rapid increase in $I_{\gamma,1}$ with γ for small γ, up to a peak, followed by a somewhat slower decrease in $I_{\gamma,1}$ with a further increase in γ.

Proposition. The vector $b^{CR}(\gamma)$ of regression coefficients from first-factor CR, $0 \leqslant \gamma < 1$, is proportional to a regression coefficients vector $b^{RR}(\delta)$ of standard RR, and vice versa; more precisely

$$b^{CR}(\gamma) = \left(1 + \frac{\gamma}{1-\gamma}\right) b^{RR}(\delta),$$

where the ridge constant $\delta \geqslant 0$ and the CR parameter $0 \leqslant \gamma < 1$ are monotonically related through

$$\delta(\gamma) = \tilde{e}(\gamma)\frac{\gamma}{1-\gamma},$$

where $\tilde{e}(\gamma)$ is a weighted average of the eigenvalues of S, increasing with γ,

$$\tilde{e}(\gamma) = \frac{b^{CR}(\gamma)'Sb^{CR}(\gamma)}{b^{CR}(\gamma)'b^{CR}(\gamma)} = \frac{b^{RR}(\delta)'Sb^{RR}(\delta)}{b^{RR}(\delta)'b^{RR}(\delta)}.$$

Proof. We start by demonstrating that the first stage of CR for $0 \leqslant \gamma < 1$ will yield $c(\gamma)$-vectors proportional to standard ridge estimators, and as a consequence we shall then see that the resulting CR regression will be proportional to an RR.

The function T to be maximized was given in formula (2.1). We take its logarithm and use the Lagrange multiplier method to cope with the unit length restriction $c'c = 1$, i.e. we differentiate

$$2\log(c's) - (1-\gamma)\log(c'Sc) - \lambda(c'c-1)$$

with respect to c and obtain the equation system

$$s/c's - (1-\gamma)Sc/c'Sc - \lambda c = 0.$$

Left multiplication by c' shows that $\lambda = \gamma$. Solving for c without bothering much about the scalars $c's$ and $c'Sc$ we obtain the relationship

$$c \propto (S+\delta I)^{-1}s \tag{2.3}$$

with $\delta = c'Sc\gamma/(1-\gamma)$, i.e. c is proportional to a standard ridge estimator with ridge constant δ. By writing

$$\delta = \frac{c'Sc}{c'c}\frac{\gamma}{1-\gamma} \tag{2.4}$$

we make its definition scale invariant in c.

Before we continue, let us temporarily follow Fearn (1990) and make expressions (2.3) and (2.4) more explicit by introducing the canonical orthogonal transformation to the eigenvectors of S. In this representation S is diagonal, with eigenvalues $e_1 \leqslant e_2 \leqslant \ldots \leqslant e_p$, say. Formula (2.3) then reads

$$c_i \propto \frac{s_i}{e_i+\delta} = \frac{b_i^{OLS}}{1+\delta/e_i} \tag{2.5}$$

(with a proportionality factor independent of i), and

$$\delta = \frac{\gamma}{1-\gamma} \frac{\sum c_i^2 e_i}{\sum c_i^2}.$$

For $\gamma = 0$ we retain the b^{OLS} of expression (2.2),

$$c_i(0) = s_i/e_i.$$

From equation (2.5) we see that, as γ increases from 0 towards 1, the weights vector $c = c(\gamma)$ will be successively redistributed towards the higher eigenvalues, and consequently $\tilde{e} = \Sigma c_i^2 e_i / \Sigma c_i^2$ is a strictly increasing function of γ. The one-to-one relationship between γ and δ follows. In the limit as $\gamma \to 1$ we obtain the PLS first latent factor, $c_i \propto s_i$ (or $c \propto s$ in the original variables).

So far we have demonstrated that the first-stage CR regressor coefficients vector $c(\gamma)$ is proportional to an RR estimator $b^{RR}(\delta)$. It remains to derive the corresponding CR estimator $b^{CR}(\gamma)$. According to the principles of CR this is done by simple linear regression of y on $c'x$, i.e. the vector of CR regression coefficients for x is

$$b^{CR}(\gamma) = c(c's)/c'Sc. \tag{2.6}$$

However, equation (2.6) is scale invariant in c, so we may choose $c = b^{RR}(\delta)$ and conclude first that

$$c's/c'Sc = c'(S+\delta I)c/c'Sc = 1 + \gamma/(1-\gamma),$$

by use of equation (2.4), and next as a consequence the desired formula

$$b^{CR}(\gamma) = \{1 + \gamma/(1-\gamma)\} b^{RR}(\delta). \qquad \square$$

For the relationship between $b^{CR}(\gamma)$ and b^{OLS} we find

$$b^{CR}(\gamma) = \left(1 + \frac{\gamma}{1-\gamma}\right)\left(I + \frac{\gamma}{1-\gamma}\tilde{e}(\gamma)S^{-1}\right)^{-1} b^{OLS}.$$

In the canonical transformation this reads

$$b_i^{CR}(\gamma) = \frac{b_i^{OLS}}{1+\gamma(\tilde{e}/e_i - 1)}. \tag{2.7}$$

For γ close to 0 we have

$$\tilde{e} \approx \frac{\sum (s_i/e_i)^2 e_i}{\sum (s_i/e_i)^2},$$

but as the example will show (see Fig. 2 later) $\tilde{e}$ may increase significantly for only a slight increase in γ near $\gamma = 0$.

3. EXAMPLE

We illustrate the behaviour of standard RR and first-factor CR by applying the methods to the cement heat evolution data set used by Stone and Brooks (1990) to illustrate CR, and previously used by Hald (1952) to illustrate multiple regression and by Draper and Smith (1981), chapter 6, to demonstrate RR as a tool for handling near collinearity (the condition number of the standardized S is 1379). The response

variable y is the heat evolved from $n = 13$ cement samples of different compositions as given by $p = 4$ explanatory variables $x_1, \ldots, x_4$ (with a sum $x_1 + \ldots + x_4$ of little variation between cement samples; the cause of the near collinearity).

Following Draper and Smith (1981) and Stone and Brooks (1990) we apply the two methods to variance-standardized explanatory variables. The so-called ridge trace is given by Draper and Smith (their Figs 6.4 and 6.5), showing a rapid change in estimated coefficients as δ is increased from 0 to about 0.002. They do not use a cross-validatory index, but a rule-of-thumb formula suggesting the value $\delta = 0.0131$ to be a reasonable choice. Fig. 1 of Stone and Brooks (1990, 1992) gives the cross-validatory index I for the one-factor CR as a function of $\alpha = \gamma/(1+\gamma)$, with further specifications in their Table 1 for $\alpha = 0$, $\alpha = 0.006$ and $\alpha = 0.5$ ($\gamma = 0$, $\gamma \approx 0.006$ and $\gamma = 1$). Our Fig. 1 shows the cross-validatory index ($I_{\gamma,1}$ in the notation of Stone and Brooks (1990)) as a function of the RR parameter (ridge constant) δ for each of the two methods.

For both methods we observe that the index increases rapidly as δ is moved away from 0. For RR the optimum value is attained at a slightly smaller δ than for one-factor CR. The latter method has in fact a slightly higher optimal I-value than RR. Of more significance is the behaviour when δ is increased above its optimal value. For RR the index curve is seen to go down much faster than for one-factor CR. The values for $\delta = \infty$ are $I = 0$ and $I = 0.959$ respectively.

Fig. 2 shows how the weighted eigenvalue average $\tilde{e}$, that relates δ with γ, increases with γ. The four eigenvalues of S are 2.236, 1.576, 0.187 and 0.002. We note a first steep increase followed by a flattening towards the limit value 2.232 for $\gamma = 1$ ($\delta = \infty$, one-factor PLS). As a function of δ the curve looks quite similar to that in Fig. 2, since Fig. 2 also indicates that in round figures δ is about twice γ (for small γ; otherwise twice $\gamma/(1-\gamma)$).

Remark. Generally a reasonable upper bound for the cross-validatory index value is given by the 'adjusted R^2-value'. In our example this quantity is 0.974 for the full model. Hence, we should not expect an I-value much higher than the maximum seen in Fig. 1 for any method of estimation.

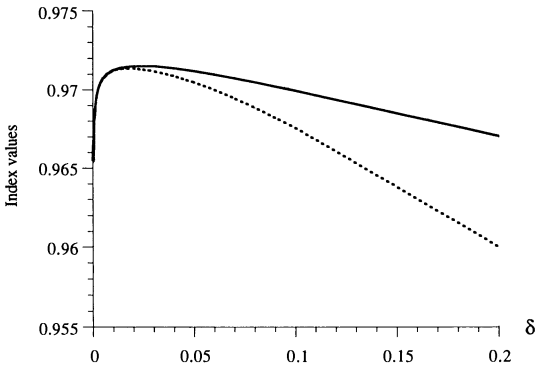

Fig. 1. RR (·····) and CR (———) cross-validatory indices in the example: in round figures δ is twice γ (see Fig. 2) and twice $\alpha = \gamma/(1+\gamma)$ here; more precisely the end point $\delta = 0.2$ corresponds to $\gamma = 0.086$ and $\alpha = 0.079$

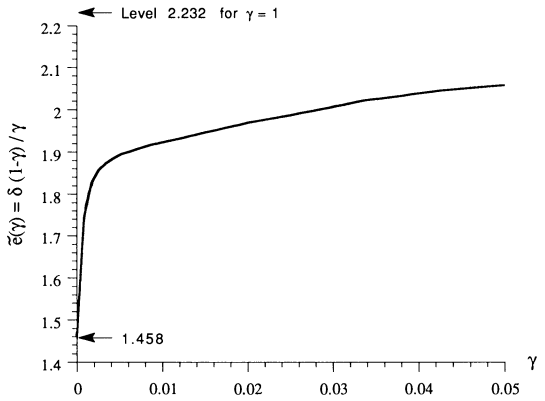

Fig. 2. Weighted eigenvalue average $\tilde{e}$ as a function of γ, equal to the factor of proportionality between δ and $\gamma/(1-\gamma)$

4. DISCUSSION

It is apparent that CR with one regressor differs from RR by only a scalar factor that moderates the RR shrinkage effect. Consider the special case e_i = constant ($=\tilde{e}$), i.e. complete orthonormality in the x-space. The shrinkage of the RR estimator relative to the OLS estimator is then given by the scalar factor $\{1+\gamma/(1-\gamma)\}^{-1} = 1-\gamma$, which is exactly compensated in the CR estimator, so $b^{\mathrm{CR}}(\gamma) = b^{\mathrm{OLS}}$ for all γ. In passing, note that this case is not possible unless the x-variables are variance standardized; this could be taken as an argument for standardization in combination with CR. Another special case is when one of the eigenvectors of S contains all the correlation with y, i.e. $s_i = 0$ for all except one index. Also in this case $b^{\mathrm{CR}}(\gamma) = b^{\mathrm{OLS}}$, whereas $b^{\mathrm{RR}}(\delta)$ is shrunk.

Multiplying an unbiased estimator by a scalar less than 1 will reduce the variance of the estimator but instead introduce a bias. If the factor is sufficiently close to 1, the MSE is reduced. It is a controversial statistical principle to refrain from unbiasedness just to gain such a shrinkage effect in the MSE. In some situations, however, a selective shrinkage undoubtedly can have quite a favourable effect on the estimator, as typified by the ridge estimator in near-collinear regression. The interpretation of the method of CR with one regressor that we have found is that it ignores the general shrinkage of the RR method but retains the shrinkage effect that protects against near collinearity. Thus, for those of us who would not shrink the OLS estimator in simple linear regression, CR with one regressor (and $\gamma < 1$) is the sensible procedure to use rather than conventional RR. Moreover, as the cross-validation index curves of the example show, standard RR need not be better in terms of MSE.

We have not seen this modified RR method discussed elsewhere. In particular, the modification is not the same as in the so-called almost unbiased ridge estimator of Singh *et al.* (1986).

From a more pragmatic point of view we should ask how the two (classes of) estimators typically will behave in practice. We have seen one example earlier. The same general picture appeared for Fearn's (1983) data used in example 3 of Stone and Brooks (1990). The optimal cross-validation index values were the same size for both methods, but first-factor CR was much less sensitive than RR to over-

estimation of the ridge parameter (γ or δ). In both examples the optimal ridge parameter values were quite small, but this is typical for cases where RR is advocated. Small values of δ (or γ) correspond to modification factors $1/(1-\gamma)$ close to 1, but, although this does not necessarily imply that the CR and RR optimal choices must be close in γ-values and in estimated coefficients b, we believe this to be typical. In other cases, pictures of other proportions may be obtained. For a different data set, with index I maximized not far from $\delta = 1$ for RR, first-factor CR gave a slightly higher but very flat index maximum between PLS ($\delta = \infty$, $\gamma = 1$) and PCR ($\gamma = \infty$).

Our conclusion from the theoretical and empirical investigations is that (first-factor) CR is preferable to standard RR. More specifically, RR should not be used as such, but rather used to yield (one-dimensional) regressors for OLS fitting.

REFERENCES

Draper, N. R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn. New York: Wiley.

Fearn, T. (1983) A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Appl. Statist.*, **32**, 73–79.

——(1990) Discussion of Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (by M. Stone and R. J. Brooks). *J. R. Statist. Soc.* B, **52**, 260–261.

Hald, A. (1952) *Statistical Theory with Engineering Applications*. New York: Wiley.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Singh, B., Chaubey, Y. P. and Dwivedi, T. D. (1986) An almost unbiased ridge estimator. *Sankhya* B, **48**, 342–346.

Stone, M. and Brooks, R. J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Statist. Soc.* B, **52**, 237–269; corrigendum, **54** (1992), 906–907.