Ridge Regression Revisited

Paul M.C. de Boer* Christian M. Hafner †

Econometric Institute Report EI 2005-29

In general ridge (GR) regression p ridge parameters have to be determined, whereas simple ridge regression requires the determination of only one parameter. In a recent textbook on linear regression, Jürgen Gross argues that this constitutes a major complication. However, as we show in this paper, the determination of these p parameters can fairly easily be done. Furthermore, we introduce a generalization of the GR estimator derived by Hemmele and by Teekens and de Boer. This estimator, which is more conservative, performs better than the Hoerl and Kennard estimator in terms of a weighted quadratic loss criterion.

Keywords: general ridge estimator, MSE performance

Running head: Ridge Regression Revisited

*Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands

†Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands

1 Introduction

Consider the normal linear regression model

$$y = X\beta + \varepsilon \tag{1}$$

where y is an $n \times 1$ vector of the variable to be explained, X is an $n \times p$ matrix of explanatory variables and ε is an $n \times 1$ vector of disturbances, distributed as $\varepsilon \sim N(0, \sigma^2 I_n)$. The $p \times 1$ parameter vector β is assumed unknown and to be estimated by the data, y and X. It is well known that the ordinary least squares (OLS) estimator for β is given by

$$b = (X'X)^{-1}X'y. \tag{2}$$

Under the model assumptions, the OLS estimator is the best linear unbiased estimator by the Gauss-Markov Theorem. However, comparing it with nonlinear or biased estimators, the OLS estimator may perform worse in particular situations. One of these is the case of near multicollinearity, where the matrix $X'X$ is nearly singular. In that situation, the variance of b, given by $\sigma^2(X'X)^{-1}$, can be very large. A biased estimator with less dispersion may in that case be more efficient in terms of the mean squared error criterion. This is the basic idea of ridge and shrinkage estimators, which were introduced by Hoerl and Kennard (1970) for the above regression model.

In a recent textbook on linear regression, Groß (2003) gives an excellent survey of alternatives to least squares estimation such as ridge estimators (pp. 115-150) and shrinkage estimators (pp. 150-162). He gives as possible justification that addition of the matrix κI_p (where κ is a scalar) to $X'X$ yields a more stable matrix $X'X + \kappa I_p$ and that the ridge estimator of β,

$$\hat{\beta} = (X'X + \kappa I_p)^{-1}X'y \tag{3}$$

should have a smaller dispersion that the OLS estimator.

To discuss the properties of the ridge estimator, one usually transforms the above linear regression model to a canonical form where the $X'X$ matrix is diagonal. Let P denote the (orthogonal) matrix whose columns are the eigenvectors of $X'X$ and let Λ be the (diagonal) matrix containing the eigenvalues.

Consider the spectral decomposition, $X'X = P\Lambda P'$, and define $\alpha = P'\beta$, $X^* = XP$, and $c = X^{*\prime}y$. Then model (1) can be written as

$$y = X^*\alpha + \varepsilon$$

and the OLS estimator of α as

$$\hat{\alpha} = (X^{*\prime}X^*)^{-1}X^{*\prime}y = (P'X'XP)^{-1}c = \Lambda^{-1}c$$

In scalar notation we have

$$\hat{\alpha}_i = \frac{c_i}{\lambda_i}, \quad i = 1, \ldots, p. \tag{4}$$

It easily follows from (3) that the principle of ridge regression is to add a constant κ to the denominator of (4), to obtain

$$\hat{\alpha}_i^R = \frac{c_i}{\lambda_i + \kappa}$$

Groß advances as criticism against this approach that all eigenvalues of $X'X$ are treated equally, while for the purpose of stabilization it would be reasonable to add rather large values to small eigenvalues but small values to large eigenvalues (Groß, 2003, p. 163). Thus, the general ridge (GR) estimator is defined to be

$$\hat{\alpha}_i^{GR} = \frac{c_i}{\lambda_i + \kappa_i} \tag{5}$$

Both types of estimators, ridge and shrinkage, are special cases of this general ridge estimator. Groß(2003) states that the disadvantage of this approach is that instead of a single ridge parameter, the determination of p ridge parameters, $\kappa_1, \ldots, \kappa_p$, is required.

The purpose of this note is twofold. First, we argue that the determination of p ridge parameters can fairly easily be done. Second, we derive the MSE properties of the GR estimator and show that estimators can be constructed that outperforms the OLS estimator in a weighted MSE sense.

We will only be concerned with GR estimators whose shrinkage intensity κ_i depends on the i-th component of the transformed data vector. Other estimators, such as the one of Strawderman (1978), depend on all components but, as noted by Lawless (1981), only provide small efficiency gains.

2 The explicit GR estimator

Minimizing the mean square error of the GR estimator with respect to κ_i, one obtains the optimal solution

$$\kappa_i = \frac{\sigma^2}{\alpha_i^2} \tag{6}$$

which is not feasible as α_i is unknown. However, replacing α_i^2 in the denominator of (6) by $(\hat{\alpha}_i^{GR})^2$ and plugging the resulting κ_i into equation (5), one obtains a quadratic equation for $\hat{\alpha}_i^{GR}$ that can be solved explicitly for a subset of the parameter space. Thus, the explicit general ridge estimator as derived independently by Hemmerle (1975) and Teekens and de Boer (1977), is given by

$$\alpha_i^* = \begin{cases} \gamma\hat{\alpha}_i & |\hat{\alpha}_i|/\sigma_i < 2 \\ \frac{1}{2}\hat{\alpha}_i(1+\sqrt{1-4\sigma_i^2/\hat{\alpha}_i^2}) & |\hat{\alpha}_i|/\sigma_i \geq 2 \end{cases} \tag{7}$$

where γ is a fixed parameter and $\hat{\alpha}_i$ is the OLS estimator (4). If σ^2 is unknown, then one can replace σ^2 in (7) by a sample estimator, $\hat{\sigma}^2$, but Teekens and de Boer (1977) show that the MSE performance of α_i^* is not affected substantially, so that in what follows we assume σ^2 to be known.

The MSE of α_i^* is defined as $MSE(\alpha_i^*) = \mathsf{E}[(\alpha_i^* - \alpha_i)^2]$. Note that $\hat{\alpha}_i \sim N(\alpha_i, \sigma_i^2)$, with $\sigma_i^2 = \sigma^2/\lambda_i$. The MSE of the OLS estimator is just given by σ_i^2. In the following we consider the relative efficiency of the ridge estimator with respect to OLS, defined by

$$r = \frac{MSE(\alpha_i^*)}{MSE(\hat{\alpha}_i)} = \frac{MSE(\alpha_i^*)}{\sigma_i^2}$$

We obtain $r = I_1 + I_2$ with

$$I_1 = \frac{1}{\sigma_i^3\sqrt{2\pi}} \int_{|x|<2\sigma_i} (\gamma x - \alpha_i)^2 e^{-\frac{1}{2}(\frac{x-\alpha_i}{\sigma_i})^2} dx \tag{8}$$

$$I_2 = \frac{1}{\sigma_i^3\sqrt{2\pi}} \int_{|x|>2\sigma_i} \left[\frac{x}{2}\left(1+\sqrt{1-4\frac{\sigma_i^2}{x^2}}\right)\right]^2 e^{-\frac{1}{2}(\frac{x-\alpha_i}{\sigma_i})^2} dx \tag{9}$$

Both integrals turn out to depend on α_i and σ_i only through the ratio, which we denote by $\theta = \alpha_i/\sigma_i$. Hence, also r only depends on θ. The second integral can be simplified to

$$I_2(\theta) = \frac{1}{\sqrt{2\pi}} \int_{|z|>2} \left[\frac{z}{2}(1+\sqrt{1-4/z^2}) - \theta\right]^2 e^{-\frac{1}{2}(z-\theta)^2} dz$$

which can be solved numerically. The first integral takes the form

$$I_1(\theta) = \frac{1}{\sqrt{2\pi}} \int_{|z|<2-\theta} \left\{\theta^2(\gamma-1)^2 + 2\theta\gamma(\gamma-1)z + \gamma^2 z^2\right\} e^{-\frac{1}{2}z^2} dz$$

Denoting the cdf and pdf of a standard normal distribution by $\Phi(\cdot)$ and $\phi(\cdot)$, respectively, note that we can write

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}z^2} dz = \Phi(b) - \Phi(a) \tag{10}$$

$$\frac{1}{\sqrt{2\pi}} \int_a^b z e^{-\frac{1}{2}z^2} dz = \phi(a) - \phi(b) \tag{11}$$

$$\frac{1}{\sqrt{2\pi}} \int_a^b z^2 e^{-\frac{1}{2}z^2} dz = \phi(a) - \phi(b) + \Phi(b) - \Phi(a) \tag{12}$$

for any real numbers a, b, $a \leq b$. Thus, I_1 can be written as

$$I_1(\theta) = \theta^2(\gamma-1)^2 A(\theta) + 2\theta\gamma(\gamma-1)B(\theta) + \gamma^2 C(\theta)$$

where

$$A(\theta) = \Phi(2-\theta) - \Phi(-2-\theta) \tag{13}$$

$$B(\theta) = \phi(-2-\theta) - \phi(2-\theta) \tag{14}$$

$$C(\theta) = A(\theta) - (2+\theta)\phi(2+\theta) - (2-\theta)\phi(2-\theta) \tag{15}$$

Teekens and de Boer (1977) compare three alternative choices of γ, 0, 0.5 and 1, in terms of the implied relative efficiency of the ridge estimator, $r(\theta)$. Figure 1 reproduces their results.

One can compare the MSE performance of the above described ridge estimator with other types of ridge estimators where the shrinkage intensity is a function of the OLS estimator. Hoerl and Kennard (1970, HK hereafter) have proposed the following ridge estimator:

$$\alpha_i^* = \frac{\hat{\alpha}_i}{1 + \sigma_i^2/\hat{\alpha}_i^2}$$

Figure 2 shows the relative MSE of the HK estimator compared with the GR estimator with $\gamma = 1$.

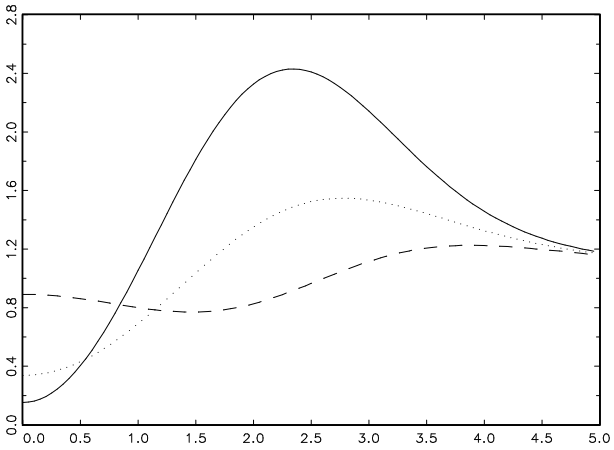

Figure 1: *Relative efficiency of the ridge estimator with* $\gamma = 0$ *(solid),* $\gamma = 0.5$ *(dotted) and* $\gamma = 1$ *(dashed).*

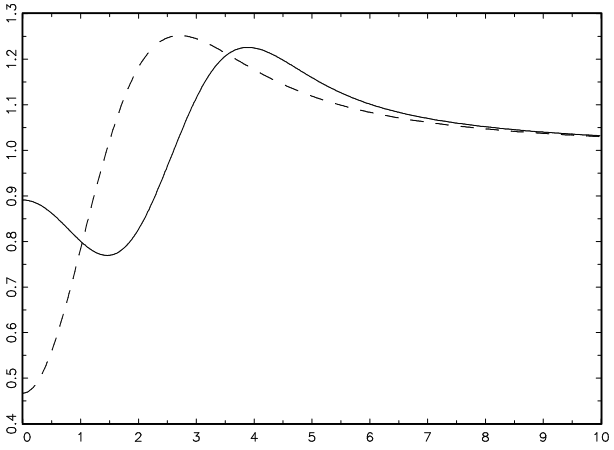

Figure 2: *Relative efficiency of the ridge estimator of Hoerl and Kennard (dashed) and* $\gamma = 1$ *(solid).*

Moreover, one can generalize the GR estimator given in (7) for general threshold value τ, $\tau \geq 2$, as

$$\alpha_i^*(\tau,\gamma) = \begin{cases} \gamma\hat{\alpha}_i & |\hat{\alpha}_i|/\sigma_i < \tau \\ \frac{1}{2}\hat{\alpha}_i(1+\sqrt{1-4\sigma_i^2/\hat{\alpha}_i^2}) & |\hat{\alpha}_i|/\sigma_i \geq \tau \end{cases} \tag{16}$$

where the estimator in (7) is given by $\alpha_i^*(2,\gamma)$. Figure 3 shows the relative MSE performance of $\alpha_i^*(\tau,\gamma)$ for $\tau = 2, 4, 6$.

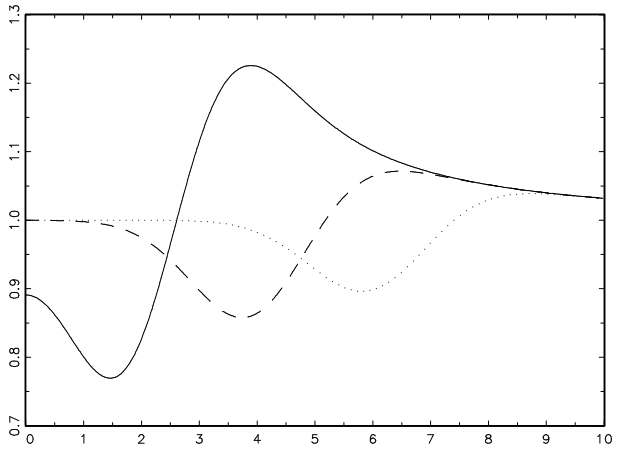

Figure 3: *Relative efficiency of the ridge estimator with* $\gamma = 1$ *and threshold at* $\tau = 2$ *(solid),* $\tau = 4$ *(dashed), and* $\tau = 6$ *(dotted).*

As there is no dominance among the estimators, we have to look for global performance criteria. For example, we can take the integrated relative difference of MSE, defined as

$$L = \int r(\theta)w(\theta)d\theta \tag{17}$$

where $w(\theta)$ is a weight function that integrates to one. The value of L could be interpreted as the average efficiency gain of using a GR estimator relative to OLS, where the average is weighted according to $w(\theta)$. Obviously, the choice of weight function is crucial for determining an optimal estimator. In the following section we motivate possible choices.

3 A Bayesian interpretation

In a Bayesian framework the parameter vector β is considered as random. Suppose one has prior information about the distribution of β. In the normal linear regression model, this prior distribution is usually Gaussian with mean $\bar{\beta}$ and covariance matrix $\bar{\Sigma}_\beta$. The Bayesian estimator for a quadratic loss criterion is then the mean of the posterior distribution and is given by

$$\bar{\bar{\beta}} = (\bar{\Sigma}_\beta^{-1} + X'X/\sigma^2)^{-1}[\bar{\Sigma}_\beta^{-1}\bar{\beta} + (X'X/\sigma^2)b] \tag{18}$$

where b is the OLS estimator in (2), see e.g. Judge et al. (1985, pp. 286). If the prior mean is set to zero, $\bar{\beta} = 0$, then $\bar{\bar{\beta}}$ simplifies to

$$\bar{\bar{\beta}} = (\sigma^2\bar{\Sigma}_\beta^{-1} + X'X)^{-1}X'y. \tag{19}$$

On the other hand, the generalized ridge estimator (5), rewritten in terms of the original variables, is given by

$$\hat{\beta}^{GR} = (PKP' + X'X)^{-1}X'y, \tag{20}$$

where $K = \mathsf{diag}(\kappa_1, \ldots, \kappa_p)$, and the ridge estimator (3) by

$$\hat{\beta}^{R} = (\kappa I_p + X'X)^{-1}X'y. \tag{21}$$

Comparing these estimators, it is clear that the Bayesian estimator (19) is identical to the GR estimator (20) for the prior covariance matrix $\bar{\Sigma}_\beta = \sigma^2(PKP')^{-1}$, and identical to the ridge estimator (21) if $\bar{\Sigma}_\beta = (\sigma^2/\kappa)I_p$. Thus, in the generalized ridge case, the prior distribution that corresponds to a particular choice of K is $N(0, \sigma^2(PKP')^{-1})$, where the covariance matrix is in general not diagonal. The prior distribution that corresponds to a particular choice of κ in the simple ridge case is $N(0, (\sigma^2/\kappa)I_p)$, with diagonal covariance matrix. In both cases, however, any marginal distribution of the prior will be normal with mean zero and some variance v_i^2, say. In the simple ridge case, we have $v_i^2 = \sigma^2/\kappa$, and in the general case, v_i^2 is a function of σ^2, $\kappa_1, \ldots, \kappa_p$, and of the eigenvectors of $X'X$.

This discussion now motivates the choice of the prior for the i-th component, $N(0, v_i^2)$ for the weight function $w(\theta)$ in the loss function (17). In fact, a

Bayesian estimator with a quadratic loss criterion minimizes the expectation of the mean squared error, where the expectation is taken with respect to the prior distribution. The minimizer is the mean of the posterior distribution, see e.g. Judge et al. (1985, pp. 139). Thus, considering $w(\theta)$ in (17) as the prior for the i-th component has a natural Bayesian interpretation.

Table 1 reports the values of L in (17) when the weight function $N(0, v_i^2)$ is used. The variance v_i^2 reflects the precision of the prior information about the parameter. If this information is extremely imprecise ($v_i = \infty$), this could be interpreted as a noninformative prior and we use a uniform density over the interval (0,1). We see that the HK estimator is for each choice of v_i dominated by some (τ, γ)-estimator. If prior information is precise, then small values of γ are preferable. If it is diffuse, then $\gamma = 1$ is best. Note that for some estimators with $\gamma = 1$, the value of L is smaller than one for every choice of v_i, which implies that there is an efficiency gain no matter how precise or imprecise prior information. However, the larger the value of τ, the closer the estimator will be to OLS and the efficiency gains are rather small. Under a noninformative prior ($v_i = \infty$), the best choice for γ and τ is 1 and 8, respectively, in which case $L = 0.9817$ and the average efficiency gain is 1.83%. One could still refine the optimization of L with respect to τ, but it seems that $\tau \approx 8$ is a conservative but good choice if prior information is diffuse.

4 Conclusions

The estimator (16) outperforms the Hoerl and Kennard (1970) estimator in terms of the weighted quadratic loss criterion. However, the choice of (τ, γ) has to made according to the prior information. If this information is diffuse, then the estimator with $(8, 1)$ is best, although efficiency gains will be small.

References

Groß, J. (2003), *Linear regression*, Lecture Notes in Statistics, Springer Verlag.

Hemmerle, W.J. (1975), An explicit solution for generalized ridge regression, *Technometrics* **17**, 309–314.

v_i	0.1	0.5	1	2	4	∞
HK	0.2792	0.5507	0.7036	0.9032	1.0137	1.0477
(τ,γ)=(2,0)	0.1389	0.3866	0.8410	1.3455	1.4118	1.3013
(2,1/2)	0.3170	0.4270	0.6280	0.9327	1.0802	1.1014
(2,1)	0.8659	0.8662	0.8425	**0.8962**	0.9895	1.0350
(4,0)	**0.0110**	**0.2522**	0.9911	2.8815	3.7309	3.1267
(4,1/2)	0.2532	0.3151	**0.5170**	1.1301	1.5672	1.5270
(4,1)	0.9998	0.9994	0.9951	0.9723	**0.9701**	0.9944
(6,1)	1.0000	0.9999	0.9999	0.9970	0.9846	0.9853
(8,1)	1.0000	1.0000	0.9999	0.9998	0.9940	**0.9817**
(10,1)	1.0000	1.0000	1.0000	0.9999	0.9981	0.9906

Table 1: *Relative efficiency (17) of alternative GR estimators with respect to OLS. The weight function $w(\theta)$ is $N(0, v_i^2)$, where $v_i = \infty$ is a uniform density over the interval (0,10). HK is the estimator of Hoerl and Kennard (1970). (τ,γ) is the estimator in (16). Bold entries are the minima for a given v_i.*

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

Lawless, J.F. (1981), Mean squared error properties of generalized ridge estimators, *Journal of the American Statistical Association* **76**, 462–466.

Strawderman, W.E. (1978), Minimax adaptive generalized ridge regression estimators, *Journal of the American Statistical Association* **73**, 623–627.

Teekens, R. and P. de Boer (1977), The exact MSE-efficiency of the general ridge estimator relative to OLS, *Econometric Institute*, Erasmus University Rotterdam, available at `http://www.few.eur.nl/few/people/pmdeboer`.