CHAPTER 4

FURTHER RIDGE THEORY

4.1 Other Interpretations of Ridge Regression

In this section we will present three interpretations for the use of ridge regression. The first one is analogous to Hoerl and Kennard reasoning while the second one is based on a Bayesian approach. In addition, in recent literature one new characterization for ridge regression is presented based on an optimization problem.

4.1.1 Restricted Least Squares Interpretation

Ridge regression may be viewed as least squares subject to a spherical restriction on the parameters. Suppose that the regression problem under study is in correlation form and that we perform least squares subject to the spherical restriction

$$\boldsymbol{\beta}'\boldsymbol{\beta} \le c^2, \tag{4.1.1}$$

where c^2 is a specified value. A restricted least squares estimator can be estimated by minimizing $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$ subject to the constraint (4.1.1). Using the method of Lagrange multipliers, we can form

$$F = (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) + k(\boldsymbol{\beta}'\boldsymbol{\beta} - c^2), \tag{4.1.2}$$

Setting $\partial F/\partial\boldsymbol{\beta} = 0$ gives the equations

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}, \tag{4.1.3}$$

which is the ridge solution. (Vinod and Ullah, 1981).

4.1.2 Bayesian Interpretation

The Bayesian approach to ridge regression is based on the assumption that we have a regression situation where

$$\mathbf{Y} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2\right). \tag{4.1.4}$$

Consider the case where the individual regression coefficients in $\boldsymbol{\beta}' = \left(\beta_1,...,\beta_p\right)$ are exchangeable ("an assumption that may not be appropriate" as emphasized by Lindley and Smith, 1972) i.e. they are unaltered by a permutation of the suffixes (i = 1,2,…, p). Suppose further that

$$\beta_j \sim N\left(\xi, \sigma_\beta^2\right). \tag{4.1.5}$$

If we suppose vague prior knowledge for ξ, then the Bayes estimate is

$$\boldsymbol{\beta}^* = \left\{\mathbf{I}_p + k\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{I}_p - p^{-1}\mathbf{J}_p\right)\right\}^{-1}\hat{\boldsymbol{\beta}}, \tag{4.1.6}$$

where $k = \sigma^2/\sigma^2{}_\beta$ and $\mathbf{J}$ is a matrix of ones. If we assume ξ= 0, and thus imply that β_i's are small then the Bayes estimate is given by

$$\boldsymbol{\beta}^* = \left\{\mathbf{I}_p + k\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right\}^{-1}\hat{\boldsymbol{\beta}}. \tag{4.1.7}$$

When σ^2, the residual regression variance, and $\sigma^2{}_\beta$, the variance of the regression coefficients are both unknown we can estimate them and calculate k^* as follows: $k^* = s^2/s_\beta^2$.

In the estimates above k is a variance ratio and is estimated from the data while in Hoerl and Kennard's argument k is the constant where the regression estimates stabilize. Like ridge method the Bayesian method attempts to avoid some of the problems caused by non-orthogonality in the data but in addition it has the advantage "of dispensing with the rather arbitrary choice of k and allows data to estimate it" (Lindley and Smith, 1972).

4.1.3 An Optimization Problem

Consider linear estimators that can be written as

$$\mathbf{B} = \mathbf{J}\mathbf{R}_X\mathbf{B}_0,$$

where $\mathbf{J}$ is a $p \times p$ matrix, $\mathbf{B}_0$ is the ordinary LS estimator and $\mathbf{R}_X = \mathbf{X}'\mathbf{X}$ (the correlation matrix). Since $\mathbf{B}$ is a linear transform of $\mathbf{B}_0$, it is a biased estimator unless $\mathbf{J} = \mathbf{R}_X^{-1}$. We have $E(\mathbf{B}) = \mathbf{J}\mathbf{R}_X\boldsymbol{\beta}$. From (3.5.2) it can be shown that

$$MSE(\mathbf{B}) = D(\mathbf{B}) + \sigma^2 tr(\mathbf{J}\mathbf{R}_X\mathbf{J}'),$$

where $D(\mathbf{B})$ is the squared bias term of $\mathbf{B}$ and is equal to $\|(\mathbf{J}\mathbf{R}_X - \mathbf{I})\boldsymbol{\beta}\|^2$.

Ridge regression is a biased estimation method based on linear estimators. Qannari et al. (1997) present an optimization problem, which leads to the ridge estimator but from another viewpoint. They suggest keeping the total variance of the parameter estimates at an "acceptable level", whiling allowing the smallest possible bias.

Consider the inequality that holds for the Euclidean norm of a matrix $0 \le D(\mathbf{B}) \le \|\mathbf{J}\mathbf{R}_X - \mathbf{I}\|^2 \|\boldsymbol{\beta}\|^2$, it seems that

(i) $D(\mathbf{B})$ is zero when $\mathbf{J} = \mathbf{R}_X^{-1}$,

(ii) or approaches zero when $\|\mathbf{J}\mathbf{R}_X - \mathbf{I}\|^2$ approaches zero.

Therefore the authors, as explained earlier, suggest minimizing the bias, i.e. $\min_J \|\mathbf{J}\mathbf{R}_X - \mathbf{I}\|^2$, under the constraint that the total variance is fixed, i.e. $tr(\mathbf{J}\mathbf{R}_X\mathbf{J}') = c$, where c is a fixed positive scalar. Solving the Lagrangian problem we obtain

$$\mathbf{J} = (\mathbf{R}_X + k\mathbf{I})^{-1},$$

which is the ridge estimator (Qannari et al., 1997).

4.2 Application of Ridge Regression in Special Cases

In chapter 3 we only consider the use of ridge regression in the multivariate linear regression model. However, many authors have used ridge regression in different cases,

for example, in logistic regression. We will discuss some cases which we consider rather useful.

4.2.1 Rank deficient model

Let us consider the case where our model is *rank deficient*. Brown (1978) examines the ridge estimator in the context of a linear model, which may be rank deficient (**X** is an $(T \times p)$ given matrix of rank m $(\leq p)$). In such a case the ridge estimator, $(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$ is not defined at $k = 0$, so Brown (1978) suggests the following definition. Let

$$\hat{\boldsymbol{\beta}}(k) = \begin{cases} (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} & for \quad k > 0 \\ \mathbf{X}^{+}\mathbf{Y} & for \quad k = 0 \end{cases}$$

where $\mathbf{X}^{+}$ denotes the Moore-Penrose pseudoinverse (Appendix A).

4.2.2 Straight line regression with a small number of observations

Carmer and Hsieh (1978) try to apply biased techniques to *straight line regression* with a small number of observations.

Having **Y** and **X** in standardized form leads to a LS estimate equal to the simple correlation $\hat{r}$, between **X** and **Y**; the regression sum of squares is equal to $\hat{r}^2$ and the residual mean square is $\hat{\sigma}^2 = (1 - \hat{r}^2)/(T - 2)$, where *T* is the number of observations.

The biased estimate of the standardized regression coefficient is $\tilde{\beta} = \tilde{r} = \hat{r}/(1 + k)$. Farebrother (in Carmer and Hsieh, 1978) proposed for an estimate of *k* the following:

$$k_1 = \frac{\hat{\sigma}^2}{\hat{r}^2} = \frac{(1 - \hat{r}^2)}{(T - 2)\ \hat{r}^2}$$

The results of the simulation study of the authors showed that none of the biased procedures are recommended for use in straight line regression problems with a small number of observations. According to the authors "all the procedures rather severely

reduced the estimate of the slope, relative to least squares, and none of the procedures produced dramatic improvements in the mean square error".

4.2.3 Models with lagged effects

In *models with lagged effects* we have

$$\mathbf{Y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}_0 \mathbf{X}_t + \boldsymbol{\beta}_1 \mathbf{X}_{t-1} + \ldots + \boldsymbol{\beta}_\ell \mathbf{X}_{t-\ell} + \mathbf{U}_t; \quad t = 1,2,\ldots,n \tag{4.2.1}$$

where $\mathbf{Y}_t$ is a dependent variable, $\mathbf{X}_t$ represents the matrix of regressors and $\mathbf{U}_t$ the random error. As we can notice from (4.2.1) the regressors involve time series which are often autocorrelated. So using ridge regression particularly for large values of ℓ is a way to tackle this problem.

However, a problem of lagged effects model is to select an appropriate number of lagged terms, i.e. the right ℓ. Erickson (1981) deals with the topic of variable selection utilizing ridge regression. In order to select variables he minimizes a prediction error, or at least an estimate of the prediction error based on ridge regression- Ridge Regression Prediction criterion (*RP*). *RP* depends on which observations and regressors are used and on the value of *k*- the ridge constant. Using ridge regression on some data the author shows that in order to find the "right" estimates for the number of lagged terms one should first calculate a *k* that minimizes the *RP* criterion for each value of ℓ and then find the overall minimum of ℓ's.

4.2.4 Subset selection

The ridge regression has also been used as a *subset selection technique* by Hoerl et al. (1986). They propose a ridge selection method that examines a full ridge solution and then deletes terms that are not significant. The deletion of the terms is based on a modified *t-test*, $t = \hat{\boldsymbol{\beta}}(k) / S_{Ri}$, where S_{Ri}^2 is the i_{th} diagonal element of $\sigma^2 (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}$. This means that we are actually testing the hypothesis $H_0 : E(\hat{\boldsymbol{\beta}}(k)) = 0$.

4.2.5 Logistic regression

Consider the *logistic regression model*:

$$\pi = \frac{1}{\left(1+e^{-\mathbf{X}\boldsymbol{\beta}}\right)} \tag{4.2.2}$$

where $\boldsymbol{\beta} = \left(\beta_0, \beta_1, ..., \beta_p\right)'$ and π is the probability that the event $\mathbf{Y}$ occurs, $\pi = P\left(\mathbf{Y}=1\right)$. The unknown parameter vector $\boldsymbol{\beta}$ can be estimated by $\hat{\boldsymbol{\beta}}$, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$. Schaefer et al (in Lee and Silvapulle, 1988) have derived the ridge estimator for the logistic regression model as

$$\hat{\boldsymbol{\beta}}(k) = \left(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X} + k\mathbf{I}\right)^{-1}\left(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}\right)\hat{\boldsymbol{\beta}},$$

where $\hat{V} = V\left(\hat{\boldsymbol{\beta}}\right)$. They have also shown that if the degree of multicollinearity is high then $MSE\left\{\hat{\boldsymbol{\beta}}(k)\right\} < MSE\left\{\hat{\boldsymbol{\beta}}\right\}$ for many observations and small value of *k*.

Lee and Silvapulle (1988) propose a method for the determination of *k* using Bayesian methods. They obtained the following two choices of *k*:

$$\hat{k}_a = (\pi+1)\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\right)^{-1}, \tag{4.2.3}$$

$$k_b = \left[tr\left(\operatorname{cov}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\right]\left[\hat{\boldsymbol{\beta}}'\left(\operatorname{cov}\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1}\hat{\boldsymbol{\beta}}\right]^{-1}. \tag{4.2.4}$$

After a Monte Carlo study for the examination of the performance of the above estimators the authors concluded that $\hat{k}_a$ is considered the "best" choice for *k*.

4.2.6 Autocorrelated disturbances

Firinguetti (1989) studies the effect of collinearity and autocorrelated disturbances in the performance of several ridge regression estimators. The use of ridge regression in generalized linear models has been considered by other authors too. Yet it had only been discussed in cases where the error variance-covariance matrix $\left(\sigma^2\boldsymbol{\Omega}\right)$ was known. Firinguetti suggests that even when one has to estimate *k* and $\boldsymbol{\Omega}$, conditions can be found where the ordinary ridge regression estimator dominates the generalized least squares (GLS) estimator.

Consider the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$ as described in (2.2.1), where $\mathbf{U}$ is a vector of T disturbances such that

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad |\rho| < 1, \quad t = 1,2,...,T \tag{4.2.5}$$

and

$$\varepsilon_t \sim N(0, \sigma^2), \quad E(\varepsilon_t \varepsilon_{t'}) = 0 \quad \text{for each } t, t' \neq t. \tag{4.2.6}$$

The GLS estimator

$$\mathbf{b} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y}, \tag{4.2.7}$$

where

$$\boldsymbol{\Omega} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & ... & \rho^{T-1} \\ \rho & 1 & ... & \rho^{T-2} \\ . & . & . & . \\ \rho^{T-1} & \rho^{T-2} & ... & 1 \end{bmatrix} \tag{4.2.8}$$

is the minimum variance unbiased estimator. Since in practice ρ is usually unknown it is estimated by $\hat{\rho} = \dfrac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2}$ where $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, the OLS residuals. Then the GLS estimator of $\boldsymbol{\beta}$ becomes

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y}. \tag{4.2.9}$$

In case when collinearity is present in a GLRM, the author suggested considering a generalized version of some well-known ridge estimators. For example, the generalized Hoerl, Kennard and Baldwin RR (GHKB) estimator is:

$$\begin{aligned} \hat{\mathbf{b}}(k_1) &= \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X} + k_1\mathbf{I}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y} \\ &= \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X} + k_1\mathbf{I}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\hat{\mathbf{b}} \end{aligned}$$

with $k_1 = \dfrac{ps^2}{\hat{\mathbf{b}}'\hat{\mathbf{b}}}$ and $s^2 = \dfrac{\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}\right)'\hat{\boldsymbol{\Omega}}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}\right)}{(n-p)}$.

One can also define the generalized Lawless and Wang RR (GLWR) estimators as

$$\hat{\mathbf{b}}(k_2) = \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X} + k_2\mathbf{I}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y}$$

$$= \left(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X} + k_2\mathbf{I}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\hat{\mathbf{b}}$$

with $k_2 = \dfrac{ps^2}{\hat{\mathbf{b}}'\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\hat{\mathbf{b}}}$.

Comparing the different estimators using MSE and absolute bias the author suggests that in the presence of multicollinearity and autocorrelation the generalized ridge regression estimators can perform better than the other methods.

4.3 A Recent Advance in Ridge Regression

It is not unusual to have collinearity and influential cases simultaneously in a data set. Walker and Birch (1988) discuss about the effect that collinearity can have on the influence of any given case and propose some influence measures in case we use ridge regression. Part C in the Appendix provides a brief overview of influence analysis.

4.3.1 Influence in Ridge Regression

Using a different notation for convenience the ridge estimator of $\boldsymbol{\beta}$ is now denoted as

$$\mathbf{b}^* = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}, \tag{4.3.1}$$

The ridge residuals are defined as $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}^*$. In order to measure the influence of a single case a version of *DFFITS* (difference in fit standardized) for RR can be used, namely

$$DFFITS^*(i) = \frac{x_i\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)}{SE\left(x_i\mathbf{b}^*\right)},$$

where $\mathbf{b}^*(i)$ is the ridge estimator of $\boldsymbol{\beta}$ without the i_{th} case, $SE\left(x_i\mathbf{b}^*\right)$ is an estimator of the standard error (SE) of the fitted value without the i_{th} case and x_i is the i^{th} row of matrix $\mathbf{X}$.

The authors also define two versions of Cook's distance D_i

$$D_i^* = \frac{\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)' \mathbf{X}'\mathbf{X}\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)}{ps^2} \text{ or } D_i^* = \frac{\left(\hat{\mathbf{Y}}^* - \hat{\mathbf{Y}}^*(i)\right)'\left(\hat{\mathbf{Y}}^* - \hat{\mathbf{Y}}^*(i)\right)}{ps^2}$$

and

$$D_i^{**} = \frac{\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)'\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\left(\mathbf{b}^* - \mathbf{b}^*(i)\right)}{ps^2}.$$

For choosing the value of k the authors suggest the value of k that minimizes the following quantity

$$C_k = \left(SSR_k / s^2\right) - T + 2tr\left(\mathbf{H}^*\right), \tag{4.3.2}$$

where SSR_k is the sum of squares of residuals from RR and $\mathbf{H}^* = \mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'$.

As one can conclude from the definitions of *DFFITS* and Cook's distance, the influence of each case is a function of the ridge parameter k. It is interesting to note that while the influence of some cases decreases the influence of some others increases. Thus, the authors advise to determine the value of k and then compute the influence measures for that k. If it is necessary to delete certain cases, the process described should be repeated.

4.3.2 Local change of small perturbations

Shi and Wang (1999) presented another approach in order to measure the influence of observations on the ridge estimator. Instead of examining the influence of case deletion they perform local influence analysis. In local influence analysis we try to estimate the local change of small perturbations on the variance or on the explanatory variables.

The functions used to estimate these changes are the generalized influence function (GIF) and the generalized Cook statistic (GC)

- Perturbing the variance

The variance of *the errors* becomes $\sigma^2\mathbf{W}^{-1}$ where $\mathbf{W} = diag(\boldsymbol{\omega})$ with diagonal elements of $\boldsymbol{\omega} = \left(\omega_1, ..., \omega_n\right)'$. The perturbed version of the ridge estimator is

$$\mathbf{b}^*(\omega) = \left(\mathbf{X}'\mathbf{W}\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}. \tag{4.3.3}$$

The generalized influence function of $\mathbf{b}^*$ under the perturbation is given by

$GIF(\mathbf{b}^*, \mathbf{l}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'D(\mathbf{e}^*)\,\mathbf{l}$, where $D(\mathbf{e}^*) = diag(\mathbf{e}_i{}^*)$ and $\mathbf{l}$ is a unit-length vector.

Again two versions of the generalized Cook statistic of $\mathbf{b}^*$ can be defined

$$GC_1(\mathbf{b}^*, \mathbf{l}) = \mathbf{l}'D(\mathbf{e}^*)\mathbf{H}D(\mathbf{e}^*)\mathbf{l}/\,ps^2\,, \tag{4.3.4}$$

and

$$GC_2(\mathbf{b}^*, \mathbf{l}) = \mathbf{l}'D(\mathbf{e}^*)\mathbf{H}^{*2}D(\mathbf{e}^*)\mathbf{l}/\,ps^2\,, \tag{4.3.5}$$

where $\mathbf{H}^* = \mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'$ and $\mathbf{H}$ is the hat matrix of LS regression.

- Perturbing the explanatory variables

Similar influence measures can be defined when we have perturbation of the explanatory variables.

Finally, recall the quantity (4.3.2) and consider the perturbation of the variance. Let $C_k(\boldsymbol{\omega})$, $SSR_k(\boldsymbol{\omega})$ and $\mathbf{H}^*(\boldsymbol{\omega})$ denote the perturbed versions of C_k, SSR_k and $\mathbf{H}^*$, respectively. Then

$$C_k(\boldsymbol{\omega}) = SSR_k(\boldsymbol{\omega})/s^2 - T + 2tr(\mathbf{H}^*(\boldsymbol{\omega})). \tag{4.3.6}$$