CHAPTER 3

RIDGE REGRESSION

3.1 Introduction

The ridge regression procedure (Hoerl and Kennard, 1970; Connife and Stone, 1973; Jones, 1972; Smith and Golstein, 1975) is based on the matrix $(\mathbf{X}'\mathbf{X}+k\mathbf{I})$, $\mathbf{I}$ denoting the identity matrix and k being a positive scalar parameter. It is a procedure that can be used in "ill-condition" situations where correlations between the various predictors in the model cause the $\mathbf{X}'\mathbf{X}$ matrix to be close to singular. In particular, we can obtain a point estimate with a smaller mean square error.

Hoerl and Kennard (1970) suggested that in order to control inflation and general instability associated with the least squares estimates, one can use

$$\hat{\boldsymbol{\beta}}(k)=(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}\,;\quad k\geq 0 \tag{3.1.1}$$

Note that the LS estimator is a member of this family with $k=0$.

The ridge estimator, though biased, has lower mean square error than the BLUE (best linear unbiased estimator). Unfortunately, this mean-squared error is a function of the unknown parameters that we are trying to estimate. Let us denote the mean square error (MSE) of a biased estimator $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}$ as:

$$MSE(\hat{\boldsymbol{\beta}}^*)=E(\hat{\boldsymbol{\beta}}^*-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^*-\boldsymbol{\beta}) \tag{3.1.2}$$

Since the squared Euclidean distance between $\hat{\boldsymbol{\beta}}^*$ and $\boldsymbol{\beta}$ is

$$L^2=(\hat{\boldsymbol{\beta}}^*-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^*-\boldsymbol{\beta}), \tag{3.1.3}$$

the $MSE\left(\hat{\boldsymbol{\beta}}^{*}\right)$ can be interpreted as the mean squared Euclidean distance between the vectors $\hat{\boldsymbol{\beta}}^{*}$ and $\boldsymbol{\beta}$ (Koutsoyiannis, 1977). Thus, an estimator with low MSE will be close to the true parameter.

One property of the least squares estimator $\hat{\boldsymbol{\beta}}$ that is frequently noted in the ridge regression literature is (Judge et al., 1985)

$$E\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\right)=\boldsymbol{\beta}'\boldsymbol{\beta}+\sigma^{2}tr\left(\mathbf{X}'\mathbf{X}\right)^{-1}>\boldsymbol{\beta}'\boldsymbol{\beta}+\frac{\sigma^{2}}{\lambda_{k}}, \tag{3.1.4}$$

where λ_k is the minimum eigenvalue of $\mathbf{X}'\mathbf{X}$. Thus, if the data are collinear, and λ_k is small, this implies that the expected squared length of the least squares coefficient vector is greater than the squared length of the true coefficient vector. In addition, the smaller the λ_k, the greater the difference.

3.2 The Reparameterized model

Let us begin with the linear regression model as given in (2.2.1). We assume that the data are in standardized form and compute the correlation matrix, and the correlation coefficients between the dependent variable and the predictors, i.e. we compute $\mathbf{X}'\mathbf{Y}$. A parameterization that is popular in ridge regression is the one that is based on the singular value decomposition of $\mathbf{X}$. The matrix $\mathbf{X}$ can be written as

$$\mathbf{X}=\mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{P}', \tag{3.2.1}$$

where $\mathbf{Q}$ is a $(T\times p)$ matrix of the coordinates of the observations along the principal axes of $\mathbf{X}$ standardized in the sense that $\mathbf{Q}'\mathbf{Q}=\mathbf{I}$. The matrix $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p$, that is,

$$\Lambda=\begin{bmatrix}\lambda_1 & 0 & 0 & \ldots & 0\\ 0 & \lambda_2 & 0 & \ldots & 0\\ . & . & . & \ldots & .\\ . & . & . & \ldots & .\\ 0 & 0 & 0 & \ldots & \lambda_p\end{bmatrix}.$$

The $\mathbf{P}$ matrix is the $(p \times p)$ matrix of eigenvectors satisfying $\mathbf{X'X} = \mathbf{P\Lambda P'}$, and $\mathbf{P'P} = \mathbf{I}$. Then the regression model can be rewritten as follows:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{U} = \mathbf{XP'P\beta} + \mathbf{U} = \mathbf{X}^*\boldsymbol{\alpha} + \mathbf{U}, \tag{3.2.2}$$

which defines a parameter vector $\boldsymbol{\alpha} = \mathbf{P'\beta}$, and $\mathbf{X}^* = \mathbf{XP}$. The OLS estimate of $\boldsymbol{\alpha}$ is denoted by $\hat{\boldsymbol{\alpha}}$ and is given by

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y} = \left(\mathbf{P'X'XP}\right)^{-1}\mathbf{P'X'Y} = \left(\mathbf{P'X'XP}\right)^{-1}\mathbf{P'X'X}\hat{\boldsymbol{\beta}} \\ &= \left(\mathbf{P'X'XP}\right)^{-1}\mathbf{P'X'XPP'}\hat{\boldsymbol{\beta}} = \left(\mathbf{P'X'XP}\right)^{-1}\left(\mathbf{P'X'XP}\right)\mathbf{P'}\hat{\boldsymbol{\beta}} \\ &= \mathbf{P'}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{3.2.3}$$

As we showed in chapter 1 the variance of the OLS estimator is

$$V\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X'X}\right)^{-1} = \sigma^2\mathbf{P\Lambda}^{-1}\mathbf{P'},$$

while

$$\begin{aligned}V\left(\hat{\boldsymbol{\alpha}}\right) &= \mathbf{P'}V\left(\hat{\boldsymbol{\beta}}\right)\mathbf{P} = \sigma^2\mathbf{P'P\Lambda}^{-1}\mathbf{P'P} \\ &= \sigma^2\boldsymbol{\Lambda}^{-1}. \end{aligned} \tag{3.2.4}$$

The elements $\hat{\alpha}_i$ are called "uncorrelated components" because $V\left(\hat{\boldsymbol{\alpha}}\right) = \sigma^2\boldsymbol{\Lambda}^{-1}$ is diagonal. Since the ridge estimator of α is given by

$$\hat{\boldsymbol{\alpha}}(k) = \left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\mathbf{P'X'Y}, \tag{3.2.5}$$

we can easily obtain the relationship

$$\begin{aligned}\hat{\boldsymbol{\beta}}(k) &= \left(\mathbf{X'X} + k\mathbf{I}\right)^{-1}\mathbf{X'Y} = \left(\mathbf{P\Lambda P'} + k\mathbf{I}\right)^{-1}\mathbf{X'Y} = \left(\mathbf{P\Lambda P'} + \mathbf{kPP'}\right)^{-1}\mathbf{X'Y} \\ &= \mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\mathbf{P'X'Y} = \mathbf{P}\boldsymbol{\alpha}(k). \end{aligned} \tag{3.2.6}$$

We can also find from the above the relationship between the ridge and the ordinary estimate, which is given by:

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

$$= \mathbf{P}(\mathbf{\Lambda} + k\mathbf{I})^{-1}\mathbf{\Lambda}\mathbf{P}'\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{\Delta}\mathbf{P}'\hat{\boldsymbol{\beta}}, \tag{3.2.7}$$

where $\mathbf{\Delta} = diag(\delta_i)$; $\delta_i = \lambda_i(\lambda_i + k)^{-1}$, i = 1,…, p is a diagonal matrix of "shrinkage factors".

We must warn the user of ridge regression that the direct ridge estimators based on the model before standardization do not coincide with their unstandardized counterparts based on model (2.2.1) (Vinod, 1978).

3.3 Hoerl and Kennard's Reasoning

If $\mathbf{B}$ is an estimate of the vector $\boldsymbol{\beta}$, the residual sums of squares is given by

$$\begin{aligned}
\phi &= (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \\
&= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{B})\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{B})\right) \\
&= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \left(\hat{\boldsymbol{\beta}} - \mathbf{B}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{B}\right) \\
&= \phi_{\min} + \phi(\mathbf{B}),
\end{aligned} \tag{3.3.1}$$

since $2\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{B}\right) = 2\mathbf{Y}'\left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{B}\right) = 0$; $\phi_{\min}$ is the residual sums of squares of the OLS.

Let $\phi_0 > 0$ be a fixed value for the error sum of squares. Then there exists a set of values of $\mathbf{B}_0$ that will satisfy the relationship $\phi = \phi_{\min} + \phi_0$. In this set we look for the estimate that has the minimum length. This can be stated as minimize $\mathbf{B}'\mathbf{B}$

$$\text{subject to } \left(\mathbf{B} - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{B} - \hat{\boldsymbol{\beta}}\right) = \phi_0 . \tag{3.3.2}$$

As a Lagrangian problem this is

$$\text{minimize } F = \mathbf{B}'\mathbf{B} + (1/k)\left[\left(\mathbf{B} - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{B} - \hat{\boldsymbol{\beta}}\right) - \phi_0\right]$$

where $(1/k)$ is the multiplier. Then

$$\frac{\partial F}{\partial \mathbf{B}} = 2\mathbf{B} + (1/k)\left[2(\mathbf{X}'\mathbf{X})\mathbf{B} - 2(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}\right] = 0\,.$$

Solving for $\mathbf{B}$ we obtain $\mathbf{B} = \hat{\boldsymbol{\beta}}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$; k is determined so that (3.3.2) is fulfilled. From (3.3.1) and the relationship $\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$ the residual sum of squares $\hat{\boldsymbol{\beta}}(k)$ is equal to

$$\begin{aligned}
\phi(k) &= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(k)\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(k)\right) \\
&= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(k)\right)\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(k)\right)\right) \\
&= \phi_{\min} + \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(k)\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(k)\right),
\end{aligned}$$

which after simple calculations becomes equal to $\phi_{\min} + k^2\hat{\boldsymbol{\beta}}(k)'(\mathbf{X}'\mathbf{X})^{-1}\hat{\boldsymbol{\beta}}(k)$ (Hoerl and Kennard, 1970).

3.4 Properties of the Ridge Estimator

As shown previously, Hoerl and Kennard's definition of the ridge estimate is

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

with $k \geq 0$ being the ridge parameter. Using the abbreviation $\mathbf{G}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}$ and $\mathbf{Z}_k = \mathbf{G}_k\mathbf{X}'\mathbf{X}$, we can rewrite the ridge estimate as

$$\hat{\boldsymbol{\beta}}(k) = \mathbf{G}_k\mathbf{X}'\mathbf{Y} = \mathbf{G}_k\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Z}_k\hat{\boldsymbol{\beta}}\,. \tag{3.4.1}$$

In what follows, we present some properties of the ridge estimator.

A) Let $\xi_i(\mathbf{G}_k)$ and $\xi_i(\mathbf{Z}_k)$ be the eigenvalues of $\mathbf{G}_k$ and $\mathbf{Z}_k$, respectively. Then

$$\xi_i(\mathbf{G}_k) = 1/(\lambda_i + k) \tag{3.4.2}$$

$$\xi_i(\mathbf{Z}_k) = \lambda_i/(\lambda_i + k). \tag{3.4.3}$$

B) The ratio of the largest characteristic root of the design matrix $(\mathbf{X}'\mathbf{X}+k\mathbf{I})$ to the smallest root is $(\lambda_1+k)/(\lambda_p+k)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ are the ordered roots of $\mathbf{X}'\mathbf{X}$, and is a decreasing function of k.

C) $\hat{\boldsymbol{\beta}}(k)$ for $k \neq 0$ is shorter than $\hat{\boldsymbol{\beta}}$, i.e.

$$\left(\hat{\boldsymbol{\beta}}(k)\right)'\left(\hat{\boldsymbol{\beta}}(k)\right) < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}. \qquad (3.4.4)$$

Recall (3.4.1) and since $\mathbf{Z}_k$ is symmetric positive definite the following holds (Hoerl and Kennard, 1970):

$$\left(\hat{\boldsymbol{\beta}}(k)\right)'\left(\hat{\boldsymbol{\beta}}(k)\right) \leq \xi^2_{\max}(\mathbf{Z}_k)\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}.$$

Since $\xi_{\max}(\mathbf{Z}_k) = \lambda_1/(\lambda_1+k)$ then (3.4.4) is verified (Hoerl and Kennard, 1970).

For $\hat{\boldsymbol{\beta}}(k)$ the residual sum of squares can be written as

$$\phi(k) = \left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}(k)\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}(k)\right) = \left(\mathbf{Y}'-\left(\hat{\boldsymbol{\beta}}(k)\right)'\mathbf{X}'\right)\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}(k)\right) =$$

$$\mathbf{Y}'\mathbf{Y}-\left(\hat{\boldsymbol{\beta}}(k)\right)'\mathbf{X}'\mathbf{Y}-\left(\mathbf{Y}'\mathbf{X}-\left(\hat{\boldsymbol{\beta}}(k)\right)'\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}(k).$$

From the definition of the ridge estimator we can replace the quantity $\mathbf{Y}'\mathbf{X}$ above with $\left(\hat{\boldsymbol{\beta}}(k)\right)'(\mathbf{X}'\mathbf{X}+k\mathbf{I})$ so the residual sum of squares becomes

$$\phi(k) = \mathbf{Y}'\mathbf{Y}-\left(\hat{\boldsymbol{\beta}}(k)\right)'\mathbf{X}'\mathbf{Y}-k\left(\hat{\boldsymbol{\beta}}(k)\right)'\left(\hat{\boldsymbol{\beta}}(k)\right).$$

This way the residual sum of squares can be described as the total sum of squares minus the "regression" sum of squares for $\hat{\boldsymbol{\beta}}(k)$ with a modification analogous to the squared length of $\hat{\boldsymbol{\beta}}(k)$.

Mean, bias and variance

The *mean* of the ridge estimator is given by

$$E\left(\hat{\boldsymbol{\beta}}(k)\right) = \mathbf{G}_k\mathbf{X}'E(\mathbf{Y})$$

$$= \mathbf{G}_k\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}_k\boldsymbol{\beta}. \qquad (3.4.5)$$

Note that when $k = 0$ then $\mathbf{Z}_k = \mathbf{I}$ and hence $E\left(\hat{\boldsymbol{\beta}}(k)\right) = \boldsymbol{\beta}$, but when $k \neq 0$, $\hat{\boldsymbol{\beta}}(k)$ provides a biased estimate of $\boldsymbol{\beta}$.

The *bias* of the estimator $\hat{\boldsymbol{\beta}}(k)$ is given by $Bias\left(\hat{\boldsymbol{\beta}}(k)\right) = -k\mathbf{G}_k\boldsymbol{\beta}$. Indeed, we know that the bias of an estimator $\mathbf{b}^*$ is defined as

$$Bias\left(\mathbf{b}^*\right) = E\left(\mathbf{b}^*\right) - \boldsymbol{\beta}.$$

Consequently, the bias of $\hat{\boldsymbol{\beta}}(k)$ is

$$\begin{aligned} Bias\left(\hat{\boldsymbol{\beta}}(k)\right) &= E\left(\hat{\boldsymbol{\beta}}(k)\right) - \boldsymbol{\beta} = \mathbf{Z}_k\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= \left[\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X} - \mathbf{I}\right]\boldsymbol{\beta} \\ &= \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\left[\mathbf{X}'\mathbf{X} - \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)\right]\boldsymbol{\beta} \\ &= -k\mathbf{G}_k\boldsymbol{\beta} \end{aligned} \tag{3.4.6}$$

or alternatively from the relationship between the ridge and the ordinary estimate (3.2.7)

$$\begin{aligned} Bias\left(\hat{\boldsymbol{\beta}}(k)\right) &= E\left(\hat{\boldsymbol{\beta}}(k)\right) - \boldsymbol{\beta} \\ &= \mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\boldsymbol{\Lambda}\mathbf{P}'\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= \left[\mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\boldsymbol{\Lambda}\mathbf{P}' - \mathbf{I}\right]\boldsymbol{\beta} \\ &= \left[\mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\boldsymbol{\Lambda}\mathbf{P}' - \mathbf{P}\mathbf{P}'\right]\boldsymbol{\beta} \\ &= \mathbf{P}\left[\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\boldsymbol{\Lambda} - \mathbf{I}\right]\mathbf{P}'\boldsymbol{\beta} \\ &= \mathbf{P}\left[\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\boldsymbol{\Lambda} - \left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)\right]\mathbf{P}'\boldsymbol{\beta} \\ &= \mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\left(\boldsymbol{\Lambda} - \boldsymbol{\Lambda} - k\mathbf{I}\right)\mathbf{P}'\boldsymbol{\beta} \\ &= -k\mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\mathbf{P}'\boldsymbol{\beta} \end{aligned} \tag{3.4.7}$$

Now it useful to give the squared bias (in its matrix version)

$$Bias\left(\hat{\boldsymbol{\beta}}(k)\right)Bias\left(\hat{\boldsymbol{\beta}}(k)\right)' = k^2\mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\mathbf{P}'\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{P}\left(\boldsymbol{\Lambda} + k\mathbf{I}\right)^{-1}\mathbf{P}' \tag{3.4.8}$$

The *variance-covariance matrix* for the ridge regression estimators is

$$\operatorname{cov}\left(\hat{\boldsymbol{\beta}}(k)\right) = \operatorname{cov}\left(\mathbf{Z}_k\hat{\boldsymbol{\beta}}\right) = \mathbf{Z}_k\operatorname{cov}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{Z}_k'$$

$$= \sigma^2 \mathbf{Z}_k (\mathbf{X}'\mathbf{X})^{-1} \mathbf{Z}_k{}' = \sigma^2 \mathbf{Z}_k \mathbf{G}_k . \qquad (3.4.9)$$

Alternatively, we can write (3.4.9) using the matrices $\mathbf{P}$ and $\mathbf{\Lambda}$. Since

$$\hat{\boldsymbol{\beta}}(k) = \mathbf{P}(\mathbf{\Lambda} + k\mathbf{I})^{-1} \mathbf{\Lambda} \mathbf{P}' \hat{\boldsymbol{\beta}} \text{, then}$$

$$\operatorname{cov}\left(\hat{\boldsymbol{\beta}}(k)\right) = \mathbf{P}(\mathbf{\Lambda} + k\mathbf{I})^{-1} \mathbf{\Lambda} \mathbf{P}' \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{P} \mathbf{\Lambda} (\mathbf{\Lambda} + k\mathbf{I})^{-1} \mathbf{P}'$$

$$= \sigma^2 \mathbf{P}(\mathbf{\Lambda} + k\mathbf{I})^{-1} \mathbf{\Lambda} (\mathbf{\Lambda} + k\mathbf{I})^{-1} \mathbf{P}' \qquad (3.4.10)$$

$$= \sigma^2 \mathbf{P} \begin{bmatrix} \frac{\lambda_1}{(\lambda_1 + k)^2} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2}{(\lambda_2 + k)^2} & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & \frac{\lambda_p}{(\lambda_p + k)^2} \end{bmatrix} \mathbf{P}' .$$

3.5 Mean Squared Error Properties

We have already denoted in (3.1.2) the MSE of an estimator as the mean Euclidean distance between the estimator and the true value. MSE is also defined as the trace of the mean dispersion error matrix (Rao and Toutenburg 1999). The mean dispersion error matrix is

$$M(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$$

$$= E(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}) + E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}) + E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta})'$$

$$= V(\hat{\boldsymbol{\beta}}) + Bias(\hat{\boldsymbol{\beta}}) Bias(\hat{\boldsymbol{\beta}})' . \qquad (3.5.1)$$

Therefore,

$$MSE(\hat{\boldsymbol{\beta}}) = tr\{M(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})\} = tr[V(\hat{\boldsymbol{\beta}})] + [Bias(\hat{\boldsymbol{\beta}})]' [Bias(\hat{\boldsymbol{\beta}})]. \qquad (3.5.2)$$

For instance, recalling (3.1.2) the MSE of the OLS estimator is:

$$\text{MSE} = E(L^2) = tr(V(\hat{\boldsymbol{\beta}})) = \sigma^2 tr(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{3.5.3}$$

where λ_i is the ith eigenvalue of $\mathbf{X}'\mathbf{X}$. In the case of the ridge estimator we have from (3.4.8) and (3.4.10) the following:

$$MSE\left(\hat{\boldsymbol{\beta}}(k)\right) = tr\left\{ V\left(\hat{\boldsymbol{\beta}}(k)\right) + Bias\left(\hat{\boldsymbol{\beta}}(k)\right) Bias\left(\hat{\boldsymbol{\beta}}(k)\right)' \right\}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\beta_i^2}{(\lambda_i + k)^2}$$

or

$$MSE\left(\hat{\boldsymbol{\beta}}(k)\right) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \boldsymbol{\beta}' (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2} \boldsymbol{\beta}$$

$$= \gamma_1(k) + \gamma_2(k). \tag{3.5.4}$$

Hoerl and Kennard (1970) proved that $\gamma_1(k)$ is a monotonic decreasing function of k, while $\gamma_2(k)$ is monotonic increasing. In addition, $\gamma_2(k)$ can be considered the square of a bias introduced when $\hat{\boldsymbol{\beta}}(k)$ is used instead of $\hat{\boldsymbol{\beta}}$ while $\gamma_1(k)$ can be shown to be the sum of the variances of the parameter estimates. The sum of the variances of all $\hat{\beta}_i(k)$'s is the sum of the diagonal elements of (3.4.10). Note that since $\mathbf{X}'\mathbf{X} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ then $\gamma_2(k)$ can be written as

$$\gamma_2(k) = k^2 \sum_{1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \tag{3.5.5}$$

where $\boldsymbol{\alpha} = \mathbf{P}'\boldsymbol{\beta}$.

3.6 Existence Theorems

The main justification for ridge regression by Hoerl and Kennard is their theorem that there always exists a $k > 0$ such that $E\left[L^2(k)\right] < E\left[L^2(0)\right] = \sigma^2 \sum_{1}^{p} (1/\lambda_i)$, where

$L^2(k)$ is the Euclidean distance between the ridge estimator and $\boldsymbol{\beta}$ while $L^2(0)$ is the Euclidean distance between the OLS and $\boldsymbol{\beta}$. To see this from (3.5.3) (3.5.4) and (3.5.5)

$$\frac{dE\left[L^2(k)\right]}{dk} = \frac{d\gamma_1(k)}{dk} + \frac{d\gamma_2(k)}{dk}$$

$$= -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} \qquad (3.6.1)$$

As mentioned in the previous paragraph, $\gamma_1(k)$ and $\gamma_2(k)$ are monotonically decreasing and increasing and thus their first derivatives are always non-positive and non-negative, respectively. So the result can be proved if we can show that there always exists a $k > 0$ such that $\frac{dE\left[L^2(k)\right]}{dk} < 0$. And this holds when

$$k < \sigma^2 / \alpha_{\max}^2 \qquad (3.6.2)$$

where $\alpha_{\max}^2$ is the squared value of the larger α_i. In most applications, interesting values of *k* usually lie in the range (0, 1). For standardized variables, this is always the case.

The difficulty in the above result is that *k* depends on σ^2 and $\boldsymbol{\beta}$, neither of which is known. Thus although k exists, we do not know whether or not we have attained a value for k which provides a lower MSE than that of LS in a specific practical problem (Draper and Smith, 1981).

In Hoerl and Kennard's existence theorem the mean square error of $\hat{\boldsymbol{\beta}}(k)$ has been compared with $\sigma^2 tr(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \sum_{i=1}^{p} (1/\lambda_i)$. Banerjee and Carr (1971) suggested comparing it with

$$\sigma^2 tr(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} = \sigma^2 \sum_{1}^{p} 1/(\lambda_i + k), \qquad (3.6.3)$$

and not to the larger quantity $\sigma^2 \sum_{i=1}^{p} (1/\lambda_i)$. In order to explain their suggestion, Banerjee and Carr (1971) introduced (see appendix B) the augmented model:

$$\begin{bmatrix} \mathbf{Y}_X \\ \cdots \\ \mathbf{Y}_A \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \cdots \\ k^{1/2}\mathbf{I}_p \end{bmatrix} \begin{bmatrix} \\ \boldsymbol{\beta} \\ \\ \end{bmatrix} + \mathbf{U}, \tag{3.6.4}$$

where $\mathbf{Y}_x$ is the original $\mathbf{Y}$, $\mathbf{Y}_A$ is a $(p \times 1)$ observation vector corresponding to the augmented part, $\mathbf{I}_p$ is a $(p \times p)$ identity matrix, and $\mathbf{U}$ is $(n+p)\times 1$ error vector. In addition, we have $E(\mathbf{Y}_X) = \mathbf{X}\boldsymbol{\beta}$ and $E(\mathbf{Y}_A) = \sqrt{k}\boldsymbol{\beta}$. The least squares estimate of $\boldsymbol{\beta}$ in the augmented model is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{Y} + \sqrt{k}\mathbf{Y}_A) \\ &= \hat{\boldsymbol{\beta}}(k) + \sqrt{k}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{Y}_A . \end{aligned}$$

For the augmented model, the authors have proved a corresponding "existence theorem". There always exists a $k > 0$ such that $E\left[L^2(k)\right] < \sigma^2 \sum_{i=1}^{p} 1/(\lambda_i + k)$. For the proof we refer the interested reader to Banerjee and Carr (1971). It is interesting to note that the same condition for k was obtained in the augmented model, namely $k < \sigma^2/\alpha_{\max}^2$, where $\alpha_{\max}^2$ is the largest component of $\boldsymbol{\alpha}$.

Conniffe and Stone (1973) comment that only if the appropriate value of k is assumed known is the proof of Hoerl and Kennard's existence theorem valid. What is important is whether the estimator with *estimated* k has better mean square error properties than least squares estimators. They also note that mean square error is not the only criterion that determines the quality of a particular estimator. Other criteria, such as that of having a tractable distribution, are also important.

3.7 Generalized Ridge Estimator

In Vinod and Ullah (1981) one can find the definition of a generalized ridge estimator (GRE) of $\boldsymbol{\alpha}$ (as given in (3.2.3)). It is obtained by augmenting the i_{th} diagonal element of $\boldsymbol{\Lambda}$ by a positive constant k_i, and using the singular value decomposition of $\mathbf{X}$. Specifically, GRE of $\boldsymbol{\alpha}$ is given by:

$$\boldsymbol{\alpha}_K = \left(\boldsymbol{\Lambda} + \mathbf{K}\right)^{-1} \boldsymbol{\Lambda}^{1/2} \mathbf{Q}'\mathbf{Y}, \tag{3.7.1}$$

where $\mathbf{K} = diag\left(k_i\right)$ is a diagonal matrix. The GRE of $\boldsymbol{\beta}$ in (2.2.1) can be written as

$$\begin{aligned}
\mathbf{b}_K &= \mathbf{P}\boldsymbol{\alpha}_K = \mathbf{P}\left(\boldsymbol{\Lambda} + \mathbf{K}\right)^{-1} \boldsymbol{\Lambda}^{1/2} \mathbf{Q}'\mathbf{Y} \\
&= \mathbf{P}\left(\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P} + \mathbf{K}\right)^{-1} \boldsymbol{\Lambda}^{1/2} \mathbf{Q}'\mathbf{Y} \\
&= \mathbf{P}\left(\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P} + \mathbf{P}'\mathbf{P}\mathbf{K}\mathbf{P}'\mathbf{P}\right)^{-1} \boldsymbol{\Lambda}^{1/2} \mathbf{Q}'\mathbf{Y} \\
&= \left(\mathbf{X}'\mathbf{X} + \mathbf{P}\mathbf{K}\mathbf{P}'\right)^{-1} \mathbf{X}'\mathbf{Y}.
\end{aligned}$$

Alternatively it can be written as

$$\mathbf{b}_K = \left(\mathbf{X}'\mathbf{X} + \mathbf{P}\mathbf{K}\mathbf{P}'\right)^{-1} \mathbf{X}'\mathbf{Y} = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}'\hat{\boldsymbol{\beta}}, \tag{3.7.2}$$

where $\boldsymbol{\Delta} = diag\left(\delta_i\right)$, the diagonal matrix of $\delta_i = \lambda_i\left(\lambda_i + k_i\right)^{-1}$.

Guilkey and Murphy (1975) considered a modification of the GRE which they called "Direct Ridge Estimator" (DRE). They suggest that only the diagonal elements of $\boldsymbol{\Lambda}$ corresponding to relatively small eigenvalues (λ_i is defined as small if $\lambda_i < 10^{-c}\lambda_{\max}$ where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ and c arbitrary constant) of $\mathbf{X}'\mathbf{X}$ should be augmented by a k_i value. This DRE will result in an estimate of $\boldsymbol{\beta}$, that is less biased than $\mathbf{b}_K$, and in cases with severe multicollinearity DRE will have a smaller MSE than the GRE.

3.8 The Ridge Trace Plot

Hoerl and Kennard (1970) claimed that a method to select the "right" value of k is the ridge trace. The ridge trace is a two-dimensional plot of $\hat{\beta}_i(k)$ against k, where $\hat{\beta}_i(k)$ is the ridge estimate of β_i obtained using the fixed value k; it usually includes a plot of

$RSS\left(\hat{\boldsymbol{\beta}}_k\right)$ against k. Typically, k runs through a short interval, beginning at k =0. As k increases, the estimates become smaller in absolute value, tending to zero as k tends to infinity. Hoerl and Kennard propose to choose the value where the "system" stabilizes.

Below we present the ridge trace for the Longley data (Appendix, Part 2); the lines present the ridge coefficients for values of k = 0 to k =0.1

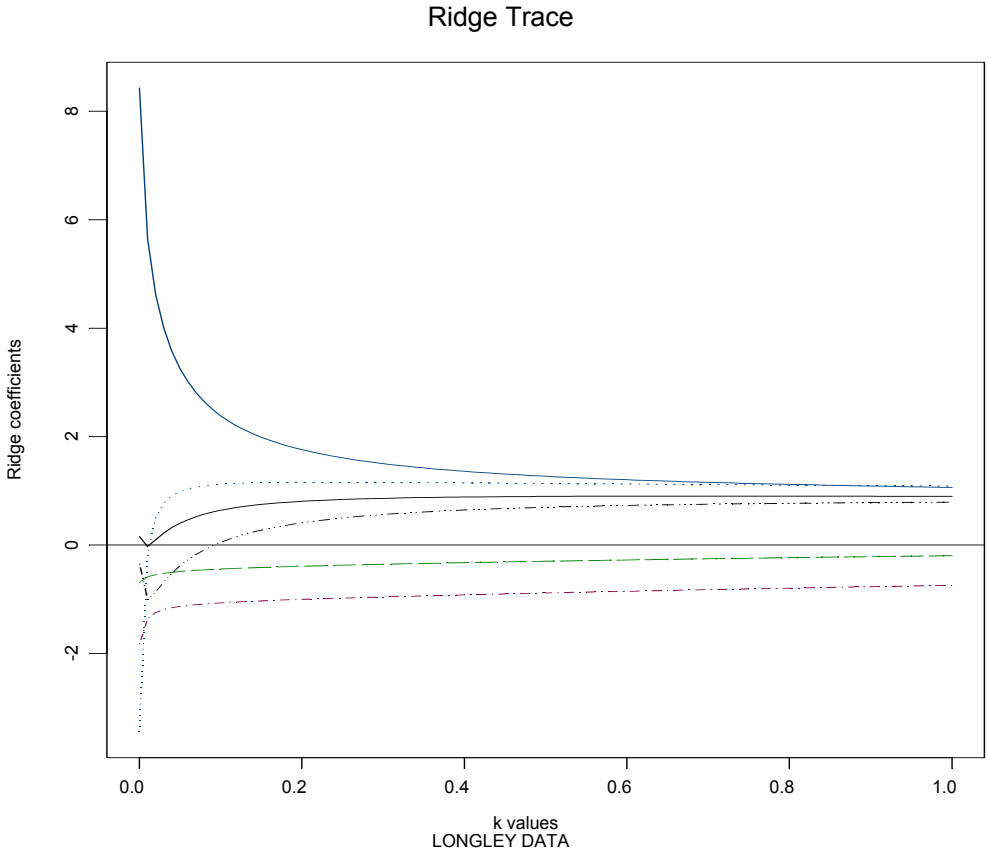

Figure 3.1 Ridge Trace

Hoerl and Kennard claimed that the *ridge trace* is a diagnostic tool that can help the analyst to estimate the value of k. However, since this procedure is based on the user's personal judgment it may be considered unreliable. Judge et. al. (1985) seem to doubt ridge trace as this "visual inspection" will lead to estimates of unknown properties. In addition, the ridge trace leads to a k which is a random variable and therefore the bias

introduced complicates the confidence intervals. They accept however, that one can learn from the data using the ridge trace.

3.8.1 An alternative Scaling for the Ridge Trace

Vinod (1976) has choosen another scaling on the horizontal axis for the ridge trace called "multicollinearity allowance", *m*, defined by

$$m = p - \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)} = p - \sum_{i=1}^{p} \delta_i \,, \tag{3.8.1}$$

where $\delta_i = \lambda_i / \lambda_i + k$, *i*=1,2,…, *p*. Note that, when $k = k_1 = ... = k_p = 0$, $m = 0$ and when $k = \infty$ then $m = p$. Some of the advantages of the *m* scale are (Vinod and Ullah, 1981):

- Finite range: In general, the *k* can have an infinite range $0 \le k \le \infty$. For the *m* scale the range is $0 \le m \le p$, which is finite.
- Generality: The *k* scale ridge trace cannot be plotted for generalized ridge regressions when the k_i's are distinct. In contrast, it is simple to plot an *m*-scale ridge trace for GRE.
- More reliable stable region: When choosing *k* from the stable region of the ridge trace, one can note that *k* may appear to be more stable for larger *k* even for completely orthogonal data; *m* scale does not have this property.

3.8.2 Quantification of the concept of a Stable region

As discussed above the ridge trace may appear to be more stable for larger *k* even for completely orthogonal data; this is not the case for the *m* scale which will not give greater stability at larger *m.* It is this property of the *m* scale that suggested a numerical measure called Index of Stability of Relative Magnitudes (ISRM), defined for $m < p$

$$ISRM = \sum_i \left[\left(p\delta_i^{\,2} / \bar{S}\lambda_i\right) - 1\right]^2, \tag{3.8.2}$$

where $\bar{S} = \dfrac{dm}{dk} = \sum \dfrac{\lambda_i}{(\lambda_i + k)^2}$. For completely orthogonal systems ISRM is equal to zero. It is possible to compute ISRM for each m (< p) and choose m where *ISRM* is the smallest. Vinod notes an important advantage of *ISRM*, that is not stochastic. The $\hat{\beta}_i(k)$ plotted in a ridge trace are stochastic and therefore k is a random variable.

3.9 Selecting value of k

A very important statistical challenge in ridge regression research is to determine the optimal value for k. In this section our aim is to bring together the methods that have been proposed in the literature and employed in practice for the selection of k. It will be assumed again that $\mathbf{X}$ and $\mathbf{Y}$ are standardized so that $\mathbf{X'X}$ is in correlation form and $\mathbf{X'Y}$ is the vector of correlations of the dependent variable with each explanatory variable.

First we present two methods that are partly based on the following optimization problem: Ridge estimators should minimize the residual sum of squares subject to the constraint that the length of the coefficient vector is something less than the least squares length.

1) Hoerl (1962) proposed reducing the length of the coefficient vector without increasing the residual sum of squares. We take

$\dfrac{d^2\left(\phi(k)^{1/2}\right)}{dC^2} = \phi(k)^{-1/2}\left\{\dfrac{-(kC)^2}{\phi(k)} + \dfrac{C^2}{\left[\hat{\boldsymbol{\beta}}'(k)\mathbf{G}_k\hat{\boldsymbol{\beta}}(k)\right]} - k\right\}$, where $\phi(k)$ is the residual sum of squares and $C^2 = \left(\hat{\boldsymbol{\beta}}(k)\right)'\hat{\boldsymbol{\beta}}(k)$ is the squared length of the vector. We choose the value k that yields the maximum value for the above derivative (Gibbons, 1981).

2) McDonald and Galarneau (1975). The choice of k is made in such a way that the squared length of the corresponding ridge estimator equals an estimated squared length of $\boldsymbol{\beta}$.

$$Q = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - s^2 \sum_{j=1}^{p} \lambda_j^{-1} \tag{3.9.1}$$

where $s^2 = \dfrac{\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{\left(T - p - 1\right)}$. Choose k such that $\left(\hat{\boldsymbol{\beta}}(k)\right)'\hat{\boldsymbol{\beta}}(k) = Q$, if $Q > 0$; choose $k = 0$ otherwise.

The next nine estimators are based on the MSE property of ridge estimators:

3) Consider the general ridge estimator (Hoerl, Kennard and Baldwin, 1975) as given in (3.7.2), i.e.

$$\mathbf{b}_K = \left(\mathbf{X}'\mathbf{X} + \mathbf{PKP}'\right)^{-1}\mathbf{X}'\mathbf{Y},$$

where $\mathbf{K} = diag\left(k_1, k_2, ...k_p\right)$. The MSE function is minimized at $k_i = \sigma^2 / \alpha_i^2$ where $\boldsymbol{\alpha} = \mathbf{P}'\boldsymbol{\beta}$. This optimal choice for k_i was also presented by Hoerl and Kennard (1970) and Goldstein and Smith (1974).

Hoerl et al. (1975) propose the use of the harmonic mean of these k_i to obtain a single value, namely k_h is given by

$$k_h = p\sigma^2 / \boldsymbol{\beta}'\boldsymbol{\beta}. \tag{3.9.2}$$

And using the estimates of σ^2 and $\boldsymbol{\beta}$ for the calculation of (3.9.2) we obtain

$$k_{HKB} = ps^2 / \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}. \tag{3.9.3}$$

4) Hoerl, Kennard, Baldwin, Thisted rule (see Lin and Kmenta, 1982)

$$k_{HKBM} = (p-2)s^2 / \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}. \tag{3.9.4}$$

This estimator was suggested because the Hoerl, Kennard and Baldwin (HKB) estimator seems to overshrink towards zero.

5) Dwividi and Srivastava (1978) select *k* in a way similar to the one of Hoerl, Kennard and Baldwin by using

$$k = \frac{s^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}} \tag{3.9.5}$$

6) The optimal value of k_i for which the *MSE* of the almost unbiased generalized ridge regression estimator (AUGRR) proposed by Singh, Chaubey and Dwivedi (1986) is minimum, is given by:

$$k_i^* = \frac{\left\{\sigma^2 + \left(\sigma^4 + \sigma^2\lambda_i\alpha_i^2\right)^{1/2}\right\}}{\alpha_i^2}$$

$$= \frac{\sigma^2}{\alpha_i^2}\left[1 + \left\{1 + \lambda_i\left(\frac{\alpha_i^2}{\sigma^2}\right)\right\}^{1/2}\right]. \tag{3.9.6}$$

In the case of the almost unbiased ordinary ridge regression estimator (AUORR) estimator where $k = k_1 = k_2 = ... = k_p$, we can obtain *k* by considering the harmonic mean of k_i^* in (3.9.6). It is given by

$$k^h = p\sigma^2 \Big/ \sum_{i=1}^{p}\left(\alpha_i^2 / \left\{1 + \left(1 + \lambda_i\left(\alpha_i^2/\sigma^2\right)\right)^{1/2}\right\}\right). \tag{3.9.7}$$

Since (3.9.8) depends on the unknown $\boldsymbol{\alpha}$ and σ^2, we replace them by their OLS estimates. Therefore the parameter in (3.9.7) becomes

$$k_{HMO} = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p}\left(\hat{\alpha}_i^2 / \left\{1 + \left(1 + \lambda_i\left(\hat{\alpha}_i^2/\hat{\sigma}^2\right)\right)^{1/2}\right\}\right), \tag{3.9.8}$$

where $\hat{\sigma}^2 = \dfrac{\left(\mathbf{Y} - \mathbf{X}^*\hat{\boldsymbol{\alpha}}\right)'\left(\mathbf{Y} - \mathbf{X}^*\hat{\boldsymbol{\alpha}}\right)}{\left(T - p\right)}$ and $\hat{\boldsymbol{\alpha}}$ as given in (3.2.3).

7) If we assume that all k_i are equal to *k,* then the MSE function is minimized when $\sum \lambda_i\left(k\alpha_i^2 - \sigma^2\right)/\left(\lambda_i + k\right)^3 = 0$ (Dempster, Schatzoff, and Wermuth, 1977). The algorithm evaluates

$$\left|\sum \lambda_i\left(k\hat{\alpha}_i{}^2(k) - s^2\right)/\left(\lambda_i + k\right)^3\right|,$$

for values of *k* and selects that value of *k* that gives the observed minimum.

8) Recall the MSE of $\boldsymbol{\beta}$ in (3.1.2) and the following criterion which differs by a constant,

$$E\left(\hat{\boldsymbol{\beta}}^{*}-\boldsymbol{\beta}\right)'\boldsymbol{\Lambda}\left(\hat{\boldsymbol{\beta}}^{*}-\boldsymbol{\beta}\right). \tag{3.9.9}$$

Both are minimized when $k_i=\sigma^2/\alpha_i$. Hoerl and Kennard (1970) suggested an iterative estimation procedure for the selection of k_i by using the OLS estimators for σ^2 and α_i as the initial values in $k_i=\sigma^2/\alpha_i$. Hemmerle (1975) proposed $s^2/\lambda_i\hat{\alpha}_i^2$ as the initial values for the iteration. Hemmerle and Brantle (1978) consider the minimization of the estimators of (3.1.2) and (3.9.9) using optimization methods as an alternative to estimating the k_i's. Specifically, the value that minimizes both (3.1.2) and (3.9.9) is

$$\frac{\lambda_i}{\lambda_i+k_i}=\begin{Bmatrix}1-s^2/\lambda_i\hat{\alpha}_i^2, & s^2/\lambda_i\hat{\alpha}_i^2<1\\ 0, & s^2/\lambda_i\hat{\alpha}_i^2\geq 1\end{Bmatrix}.$$

The authors also consider including a priori information about β by constraining the estimates of the parameters in the linear model. They obtain the ridge estimators using quadratic programming methods.

9) Consider the MSE of the ridge estimator as function of k, λ, $\boldsymbol{\alpha}$ and σ, i.e.

$$MSE\left(\hat{\beta}(k)\right)=\gamma\left(k,\lambda,\boldsymbol{\alpha},\sigma\right)=\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{\left(\lambda_i+k\right)^2}+k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{\left(\lambda_i+k\right)^2}$$

To get the best k Nordberg (1982) suggest to use "the empirical MSE-function" $\gamma\left(k,\lambda,\hat{\boldsymbol{\alpha}},\hat{\sigma}\right)$ where $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}$ are "good" estimators of $\boldsymbol{\alpha}$ and σ. The procedure is the following:

(i) Choose a preliminary $k=k_0\geq 0$.

(ii) Set $\hat{\boldsymbol{\alpha}}=\mathbf{P}'\hat{\boldsymbol{\beta}}(k_0)$.

(iii) Set $\hat{\sigma}^2=\dfrac{1}{T-p}\left|\mathbf{Y}-X\hat{\boldsymbol{\beta}}(0)\right|^2$,

(iv) Compute the k-value which minimizes the function $\gamma=\left(k,\lambda,\hat{\boldsymbol{\alpha}},\hat{\sigma}\right)$.

Denote by $K^*(k_0)$ the *k*-value obtained by the above procedure with k_0 as a "start value". By setting $k_0 = 0$ and by iterating the above procedure by $k_{j+1} = K^*(k_j)$, j = 0,1,2,… until it "stabilizes", i.e. until $k_{j+1} \approx k_j$ we can obtain good *k*-values.

10) As already shown the GRE is given by $\boldsymbol{\alpha}_K = (\boldsymbol{\Lambda} + \mathbf{K})^{-1} \boldsymbol{\Lambda}^{1/2} \mathbf{Q}'\mathbf{Y}$. Minimizing the MSE of the GRE term-by-term i.e., minimizing the diagonal elements of the mean squared error matrix

$$\sigma^2 \sum_{i=1}^{p} \lambda_i / (\lambda_i + k_i)^2 + k_i^2 \sum_{1}^{p} \alpha_i^2 / (\lambda_i + k_i)^2 \tag{3.9.9}$$

with respect to k_i yields the optimum value

$$k_{i(opt)} = \frac{\sigma^2}{\alpha_i^2} \quad (i = 1,2,..., p). \tag{3.9.10}$$

Hoerl and Kennard suggested to start with $\frac{S^2}{\hat{\beta}_i^2} = \hat{k}_i$ where $\hat{\beta}_i$ is the *ith* element of the least squares estimator and S^2 is an unbiased estimator of σ^2.

Replacing k_i in $\mathbf{K}$ by $\hat{k}_i$ to form $\hat{\mathbf{K}}$ and substituting it in place of $\mathbf{K}$ in (3.9.9) leads to an adaptive estimator of $\boldsymbol{\alpha}$ (Dwivedi et al., 1980): $\boldsymbol{\alpha}_{ad} = (\boldsymbol{\Lambda} + \hat{\mathbf{K}})^{-1} \mathbf{X}^{*'} \mathbf{Y}$.

11) Obenchain (see Gibbons, 1981) considers a family of two-parameter estimators $\boldsymbol{\beta}^*(k,q) = \left[(\mathbf{X}'\mathbf{X})^{-q+1} + k\mathbf{I}\right]^{-1} (\mathbf{X}'\mathbf{X})^{-q} \mathbf{X}'\mathbf{Y}$. For q = 0 we obtain the ridge estimator. In order to obtain the minimum mean squared error we choose q so as to maximize

$$C(q) = \frac{\sum_{i=1}^{p} |r_i| \lambda_i^{1-q/2}}{\left[\sum_{i=1}^{p} r_i^2 \sum_{i=1}^{p} \lambda_i^{(1-q)}\right]^{1/2}},$$

where $r = \boldsymbol{\Lambda}^{-1/2} \mathbf{P}'\mathbf{X}'\mathbf{Y}$. The parameter k is then chosen so as to minimize $\tilde{L} = n \ln(2\pi e \tilde{\sigma}) + \xi'\xi - (r'\xi)/\tilde{\sigma}$, where $\xi_i = sign(r_i)[\delta_i/(1-\delta_i)]^{1/2}$, $\delta_i = \lambda / (\lambda_i + k\lambda_i^q)$

and $\tilde{\sigma} = 2\left\{ \left[(r'\xi)^2 + 4n\right]^{1/2} + (r'\xi)\right\}^{-1}$. Goldstein and Smith (1974) have considered an equivalent two-parameter estimator where the parameter $m=1-q$ is an integer.

Next we consider Bayesian approaches to the selection of k.

12) Lindley and Smith (1972) showed that if $\mathbf{Y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$ and the prior for $\boldsymbol{\beta}$ is $N(0, \sigma_\beta^2\mathbf{I})$, then $\hat{\boldsymbol{\beta}}(k)$ is the Bayes estimator where $k = \dfrac{\sigma^2}{\sigma_\beta^2}$. Since σ^2, the residual regression variance, and σ_β^2, the variance of the regression coefficients are usually both unknown we should estimate them and calculate k as follows:

$$k_{LS} = \frac{s^2}{s_\beta^2}. \tag{3.9.11}$$

13) Lawless and Wang (1976) also adopt a Bayesian approach and estimate the variance ratio by

$$k_{LW} = \frac{ps^2}{\sum \lambda_i \hat{\alpha}_i^2} \tag{3.9.12}$$

14) Dempster, Schatzoff, and Wermuth (1977) in a large simulation study suggested an estimator RIDGM, which is motivated by the Bayesian interpretation and is similar to the McDonald-Galarneau estimator. Given $\boldsymbol{\alpha} \sim N(0, \omega^2\mathbf{I})$ it follows that

$$\sum_{i=1}^{p} \hat{\alpha}_i^2 / \sigma^2 \left[(1/k) + (1/\lambda_i)\right] \sim \chi_p^2 \tag{3.9.13}$$

where $k = \sigma^2/\omega^2$. Replacing σ^2 by s^2 and using the fact that $E(\chi_p^2) = p$, i.e.

$$\sum_{i=1}^{p} \hat{\alpha}_i^2 / s^2 \left[(1/k) + (1/\lambda_i)\right] = p. \tag{3.9.14}$$

The authors propose to use that value of k that satisfies (3.9.14).

Finally, two more approaches are given below:

15) Consider the relation of the ridge estimator with the LS estimator. Based on (3.2.5) and (3.7.2) respectively one can easily obtain:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}(k) &= (\boldsymbol{\Lambda} + k\mathbf{I})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{Y} \\ &= (\boldsymbol{\Lambda} + k\mathbf{I})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= (\boldsymbol{\Lambda} + k\mathbf{I})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P}\hat{\boldsymbol{\alpha}} \\ &= (\boldsymbol{\Lambda} + k\mathbf{I})^{-1}\mathbf{P}'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{P}\hat{\boldsymbol{\alpha}} = (\mathbf{I} + k\boldsymbol{\Lambda}^{-1})^{-1}\hat{\boldsymbol{\alpha}} \qquad (3.9.15)\end{aligned}$$

Hocking et al. (1976) introduce a class of biased estimators of the coefficients in the linear regression model defined by

$$\tilde{\boldsymbol{\alpha}} = \mathbf{B}\hat{\boldsymbol{\alpha}}. \qquad (3.9.16)$$

B is a diagonal matrix with diagonal components given by $b_i = \sum_{i=1}^{p}\gamma_i$, where γ_i are to be determined. Comparing the ridge estimator as given in (3.9.15) with the general estimator given in (3.9.16) yields the relation

$$b_i = (1 + k\lambda_i^{-1})^{-1}$$

or equivalently,

$$k = \lambda_i(1 - b_i)/b_i. \qquad (3.9.17)$$

If we specify b_i using, for instance, the shrinkage estimator then *k* can be obtained as a mean of the values in (3.9.17) or as a least squares determination. Specifically,

$$k = p^{-1}\sum_{i=1}^{p}\lambda_i(1 - b_i)/b_i, \qquad (3.9.18)$$

or

$$k = \sum_{i=1}^{p}\lambda_i b_i/(1 - b_i)\Big/\sum_{i=1}^{p}b_i^{\,2}/(1 - b_i)^2. \qquad (3.9.19)$$

The authors suggest a special case of (3.9.19) namely,

$$k = \sum_{i=1}^{p} \alpha_i^2 \, \lambda_i^{\,2} / \sigma^2 \Big/ \sum_{i=1}^{p} \alpha_i^{\,4} \lambda_i^2 / \sigma^4 \; .$$

16) Golub et al. (1979) consider the generalized-cross validation (GCV) method for choosing the value of *k* in ridge regression. Specifically, *k* is the value that minimizes $V(k/n)$ where

$$V(k/n) = \frac{1}{n} \left\| (I - \mathbf{A}(k/n)\mathbf{Y}) \right\|^2 \Big/ \left[\frac{1}{n} tr(\mathbf{I} - \mathbf{A}(k/n)) \right]^2$$

and

$$\mathbf{A}(k/n) = \mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'$$

The authors point out that since this method does not require the estimation of σ^2 it can be used when *n*-*p* is small or when $p \geq n$.

Table 3.1 summarizes the different criteria.

aa	Criterion	Function to minimize-maximize	Reference
1		Choose k that yields the observed maximum of $\frac{d^2\left(\phi(k)^{1/2}\right)}{dC^2} = \phi(k)^{-1/2}\left\{\frac{-(kC)^2}{\phi(k)} + \frac{C^2}{\left[\hat{\boldsymbol{\beta}}'(k)\mathbf{G}_k\hat{\boldsymbol{\beta}}(k)\right]} - k\right\}$	Hoerl, 1962 (in Gibbons, 1981)
2		Choose k such that $\left(\hat{\boldsymbol{\beta}}(k)\right)'\hat{\boldsymbol{\beta}}(k) = Q = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - s^2\sum_{j=1}^{p}\lambda_j^{-1}$, if $Q > 0$; choose k = 0 otherwise	McDonald and Galarneau (1975)
3	$k_{HKB} = ps^2/\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$		Hoerl, Kennard and Baldwin (1975)
4	$k_{HKBM} = (p-2)s^2/\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$		Hoerl, Kennard, Baldwin, Thisted (in Lin and Kmenta, 1982)
5	$k_{DS} = \frac{s^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$		Dwividi and Srivastava (1978)
6	$k_{HMO} = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p}\left(\hat{\alpha}_i^2/\left\{1+\left(1+\lambda_i\left(\hat{\alpha}_i^2/\hat{\sigma}^2\right)\right)^{1/2}\right\}\right)$		Singh, Chaubey and Dwivedi (1986)

(continued from previous page)

7		Choose k such that $\left\|\sum \lambda_i \left(k\hat{\alpha}_i^{\,2}(k) - s^2\right) \big/ (\lambda_i + k)^3\right\|$ is minimum	Dempster, Schatzoff, and Wermuth (1977)
8	$\dfrac{\lambda_i}{\lambda_i + k_i} = \begin{cases} 1 - s^2 \big/ \lambda_i \hat{\alpha}_i^2, & s^2 \big/ \lambda_i \hat{\alpha}_i^2 < 1 \\ 0, & s^2 \big/ \lambda_i \hat{\alpha}_i^2 \geq 1 \end{cases}$		Hemmerle and Brantle (1978)
9		Choose k by minimising $MSE\left(\hat{\boldsymbol{\beta}}(k)\right) = \gamma(k, \lambda, \boldsymbol{\alpha}, \sigma)$ $= \sigma^2 \sum_{i=1}^{p} \dfrac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \dfrac{\alpha_i^2}{(\lambda_i + k)^2}$ using an algorithm which gives values to k then to $\boldsymbol{\alpha}$ and σ.	Nordberg (1982)
10	For the general ridge regression estimator $\dfrac{S^2}{\hat{\beta}_i^2} = \hat{k}_i$, where S^2 is an unbiased estimator of σ^2		Hoerl and Kennard (in Dwivedi et al., 1980)
11		For the two-parameter estimator $\boldsymbol{\beta}^*(k, q) = \left[(\mathbf{X}'\mathbf{X})^{-q+1} + k\mathbf{I}\right]^{-1} (\mathbf{X}'\mathbf{X})^{-q} \mathbf{X}'\mathbf{Y}$, choose q so as to maximize $C(q) = \left[\sum \lvert r_i \rvert \lambda_i^{((1-q)/2)}\right] \Big/ \left[\left(\sum r_i^2\right)\left(\sum \lambda_i^{(1-q)}\right)\right]^{1/2}$,	Obenchain (in Gibbons, 1981)

(continued from previous page)

		choose k so as to minimize $\tilde{L} = n\ln(2\pi e\tilde{\sigma}) + \xi'\xi - (r'\xi)/\tilde{\sigma}$	
12	$k = s^2/s_\beta^2$		Lindley and Smith (1972)
13	$k = ps^2/\sum \lambda_i \hat{\alpha}_i^2$		Lawless and Wang (1976)
14		k is obtained by solving $\sum_{i=1}^{p} \hat{\alpha}_i^2/s^2[(1/k)+(1/\lambda_i)] = p$	Dempster, Schatzoff, and Wermuth (1977)
15	$k = \sum_{i=1}^{p} \alpha_i^2 \lambda_i^2/\sigma^2 \Big/ \sum_{i=1}^{p} \alpha_i^4 \lambda_i^2/\sigma^4$		Hocking,Speed and Lynn (1976)
16		k is the value that minimizes $V(k/n)$, $V(k/n) = \frac{1}{n}\|(I - \mathbf{A}(k/n)\mathbf{Y})\|^2 \Big/ \left[\frac{1}{n} tr(\mathbf{I} - \mathbf{A}(k/n))\right]^2$	Golub, Heath and Wahba (1979)

Table 3.1: Selection Criteria

3.10 Illustration to Real Data

In order to illustrate the use of ridge regression we applied the method to a real data set.

3.10.1 Bodyfat data

The data are the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. The dependent variable (**Y**) is the PERCENT BODY FAT (from Siri's equation). The data were obtained from StatLib (Dataset Archive) and were submitted by Dr. A. Garth Fisher. More details about the data can be found in http://lib.stat. cmu.edu /datasets/ bodyfat.

The independent variables (matrix *X*) are

- AGE(years)
- WEIGHT(lbs)
- HEIGHT(inches)
- NECK CIRCUMFERENCE(cm)
- CHEST CIRCUMFERENCE(cm)
- THIGH CIRCUMFERENCE(cm)
- FOREARM CIRCUMFERENCE(cm)

Note: This data set included another 7 explanatory variables, which were left out for reasons of convenience and efficient data presentation.

Accurate measurement of body fat is inconvenient or costly so it is desirable to have easy methods of estimating body fat that are not inconvenient or costly. Eventually, we wish to produce a regression equation which will predict percentage body fat in terms of anatomical measurements.

3.10.2 Data analysis

The regression model is: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$. We wish to examine the inclusion of correlated variables to our model in order to illustrate collinearity diagnostics and the ridge regression solution. As one can see from the next scatterplot matrix (containing all the scatterplots of one variable against another) some of the explanatory variables are highly correlated.

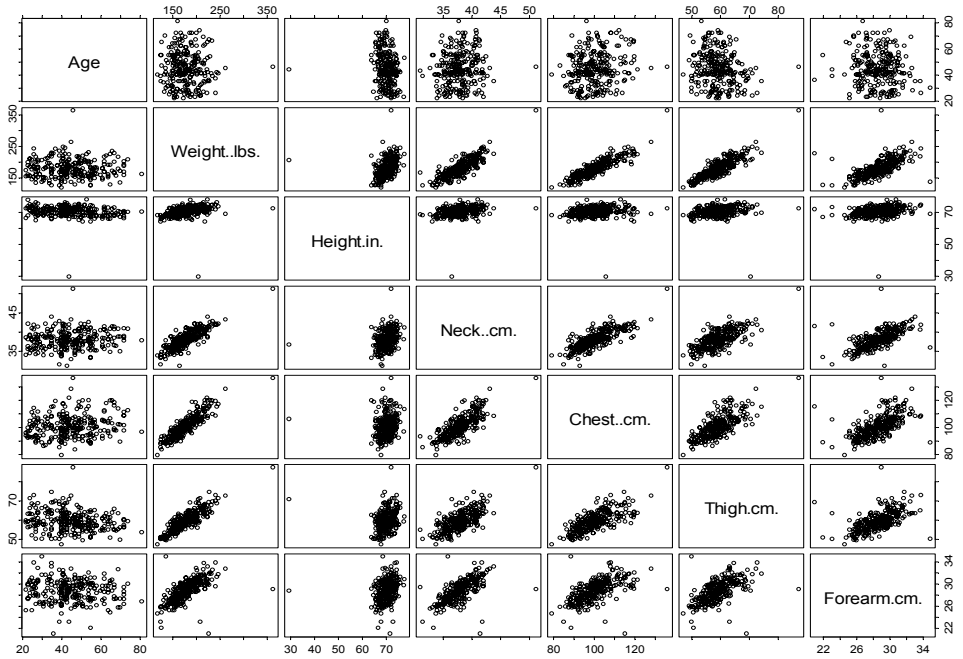

Figure 3.2: Correlation matrix

Checking the correlation coefficients as well we find several large pairwise correlations. For example, the correlation between chest circumference and weight is 0.894 which is rather large. In addition, we check whether $r_{ij} \geq R^2$ where $R^2 = 0.5894$. This holds in 7 cases (denoted with bold, italics in Table 3.2) so we can say that multicollinearity is present.

	Age	Weight lbs	Height in	Neck cm	Chestcm	Thigh cm	Forearm cm
Age	1.000						
Sig. (2-tailed)	.						
Weightlbs	-0.013	1.000					
Sig. (2-tailed)	.840	.					
Heightin	-0.172	0.308	1.000				
Sig. (2-tailed)	.006	.000	.				
Neckcm	0.114	***0.831***	0.254	1.000			
Sig. (2-tailed)	.072	.000	.000	.			
Chestcm	0.176	***0.894***	0.135	***0.785***	1.000		
Sig. (2-tailed)	.005	.000	.032	.000	.		
Thighcm	-0.200	***0.869***	0.148	***0.696***	***0.730***	1.000	
Sig. (2-tailed)	.001	.000	.018	.000	.000	.	
Forearmcm	-0.085	0.630	0.229	***0.624***	0.580	0.567	1.000
Sig. (2-tailed)	.178	.000	.000	.000	.000	.000	.

Table 3.2: Correlation coefficients of the predictors

Then the least squares estimates obtained by fitting the regression model are calculated and given below:

Parameter estimates					
	Parameter Estimate	Std. Error	Standardized Estimate	t value	p-value
Intercept	-30.808	14.769		-2.086	0.038
Age	0.175	0.033	0.264	5.287	0.000
Weightlbs	0.011	0.046	0.036	0.223	0.824
Heightin	-0.293	0.114	-0.128	-2.572	0.011
Neckcm	-0.744	0.270	-0.216	-2.751	0.006
Chestcm	0.555	0.107	0.559	5.168	0.000
Thighcm	0.537	0.155	0.337	3.457	0.001
Forearmcm	0.041	0.229	0.010	0.179	0.858

Multiple R-squared: 0.5894

Table 3.3: The values of the regression coefficients and the *p*-values

Only two predictor coefficient estimates (weightlbs and forearmcm) have large *p*-values i.e. they are not significant.

To decide for multicollinearity we calculate some diagnostics (see next table):

The predictors	VIF	R_i^2	Leamer's c_i
Age	1.482	0.325	0.822
Weightlbs	15.292	0.935	0.256
Heightin	1.474	0.322	0.824
Neckcm	3.668	0.727	0.522
Chestcm	6.960	0.856	0.379
Thighcm	5.644	0.823	0.421
Forearmcm	1.817	0.445	0.742

Table 3.4: The multicollinearity diagnostics

		Variance Proportions							
Dimension	Eigenvalue	(Constant)	AGE	WEIGHT	HEIGHT	NECK	CHEST	THIGH	FOREARM
1	7.906	.00	.00	.00	.00	.00	.00	.00	.00
2	.070	.00	.61	.00	.00	.00	.00	.00	.00
3	.017	.01	.04	.06	.02	.00	.00	.00	.00
4	.003	.00	.00	.04	.31	.00	.00	.01	.48
5	.002	.03	.00	.04	.09	.00	.01	.27	.37
6	.001	.01	.34	.01	.06	.06	.38	.27	.10
7	.001	.01	.01	.00	.03	.87	.18	.00	.05
8	.000	.95	.01	.84	.49	.07	.43	.44	.00

Table 3.5: Eigenvalues and variance proportions

A variable X_i is harmfully multicollinear only if its multiple correlation with other members of the independent variable set, R_i^2, is greater than the dependent variable's multiple correlation with the entire set, R^2 (Greene, 1993). This is the case in four cases as we can see from Table 3.4. We can reach the same conclusion using Leamer's diagnostic which is small for the same cases. The *VIF* for weightlbs is 15.292 which is quite large.

Calculating the condition number we find $K = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2} = \left(\frac{7.90631}{0.0003}\right)^{1/2} = 26{,}385.36$ which is rather large, while small eigenvalues (i.e. 0.002) indicate near linear dependencies. Another diagnostic is the determinant of the correlation matrix $|\mathbf{X}'\mathbf{X}|$ = 0.0038. The closer $|\mathbf{X}'\mathbf{X}|$ is to 0, the greater the severity of multicollinearity. Finally the sum of λ_i^{-1} = 6,273.9487. Recall that in an orthogonal system it would be 7.

Since our data suffer from multicollinearity we will try to implement ridge regression. To this aim we calculate *k* by four methods.

a) Hoerl and Kennard: $k_{HK} = s^2 / \max(\hat{\alpha}_i^2) = 0.008$, where s^2 is the estimate of the variance and $\hat{\alpha}$ the least square estimate (see (2.2.4)).

b) Hoerl Kennard and Baldwin: $k_{HKB} = \frac{ps^2}{\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}} = 0.021$

c) Lawless and Wang: $k_{LW} = \dfrac{ps^2}{\sum \lambda_i \hat{\alpha}_i^2} = 0.020$

d) Vinod's *ISRM*: $k_{ISRM} = 0.44$

Ridge Regression Results			
	Ridge Estimate	Std. Error	Standardized Ridge Estimate
Intercept	-22.978	6.321	
Age	0.120	0.019	0.181
Weightlbs	0.046	0.006	0.160
Heightin	-0.285	0.068	-0.125
Neckcm	0.037	0.099	0.011
Chestcm	0.270	0.026	0.272
Thighcm	0.281	0.043	0.176
Forearmcm	0.116	0.128	0.028

Table 3.6: Ridge estimates

Observing the Ridge Trace we note that when $k_{ISRM} = 0.44$ the coefficients stabilize. So in Table 3.6 we give the ridge estimates using the k obtained by minimizing the Index of Stability of Relative Magnitudes (*ISRM*).

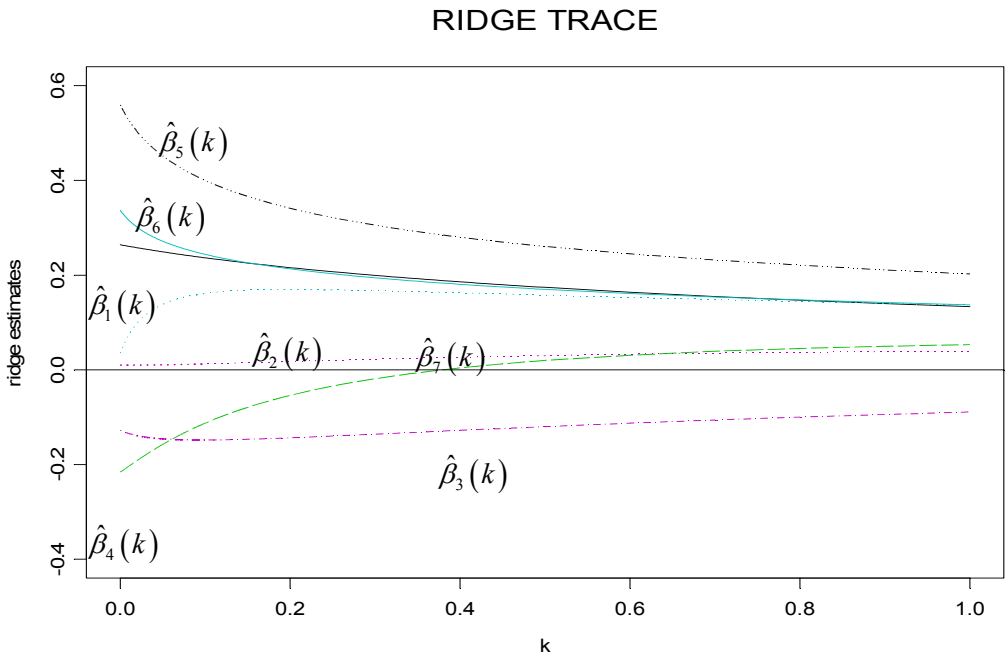

Figure 3.3: The Ridge Trace

3.11 A Critical View of Ridge Regression

There are a lot of controversies in the literature about the success of ridge regression. Some authors are in favour of using ridge regression while others greatly criticize it stating that it is not "always" better than least squares.

According to Draper and Van Nostrand (1979) ridge regression is a technique that enables one to assume prior information of a specific kind. So ridge regression is an appropriate multicollinearity remedy in case we consider a Bayesian formulation or in case of a restricted least squares formulation. In any other circumstances, they claim that ridge regression should not be used. They doubt the value of ridge regression, since they find that it is favored over least squares only when the ridge estimators are close to the least squares values. Marquardt and Snee (1975) express the same opinion with Draper and Van Nostrand stating that ridge regression is reasonable under a Bayesian interpretation. They also comment that since in correlation form regression coefficients rarely exceed three, one can consider bounded priors.

Recalling the fact that the ridge estimator is just the OLS estimator biased by $(\lambda_i/(\lambda_i + k))$, Pagel (1981) points out that since this fraction declines with λ_i, ridge applies the greatest shrinkage, and thus reduces the variability most, for the coefficients associated with small eigenvalues. However, it is not always right to treat the coefficients of small eigenvalues as less "important" than those of large eigenvalues. Small eigenvalues may derive from the fact that the data are inadequate for the estimation of the model parameters; or from a misspecification of the model. Ridge regression ignores such problems and tries to obtain the regression estimates.

Gunst and Mason (1977), in their evaluation on five estimators of the regression coefficients (least squares (LS), principal components (PC), ridge regression (RR), latent root (LR) and a shrunken estimator (SH)), concluded that the PC and LR estimators appeared to offer the best opportunity for large decrease in MSE over LS for the multicollinear data, while ridge regression and SH performed well for the near-orthogonal data.

Many simulations have been performed to compare ridge regression estimates to least squares estimates, in a mean square error sense. Pagel (1981) notes that based on

Monte Carlo studies, ridge regression reliably reduces the mean squared error of the estimated coefficients under conditions of multicollinearity and low signal to noise ratios. However, these simulations must be viewed with caution. Draper and Van Nostrand (1979) claim that these simulations involve restrictions on the parameter values (the situations where ridge regression is the appropriate technique theoretically). Opponents of ridge regression also cite *inconsistent findings* among studies and criticize the modeling of fixed length of the regression coefficient vectors in many simulations.