4.5 Ridge Regression

Ridge Regression

Where there are many **correlated variables** in a linear regression model, their coefficients can become poorly determined and exhibit **high variance**. A wildly large positive coefficient on one variable can be cancelled by a similarly large negative coefficient on its correlated cousin. By imposing a **size constraint** on the coefficient, this phenomenon is prevented from occurring.

Model selection, as in the previous section, effectively sets some β_j equal to zero. Based on this idea, **ridge regression imposes a size constraint** on the coefficients, that is, **"shrinks"** them towards zero.
Benefits are similar to model selection: **Model complexity** is effectively **reduced**, leading to a **smaller variance**, and if done properly, to a **smaller mean squared prediction error**.

The task is to minimize

$$\mathrm{RSS}(\lambda) = \|\boldsymbol{y} - a_0\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n}(y_i - a_0 - x_i^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \quad (4.5.1)$$

with respect to a_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. Note that $x_i = (x_{i1}, \ldots, x_{ip})$, and that $\boldsymbol{X}$ here does not have a leading column of 1's. The parameter $\lambda > 0$ controls the degree of shrinkage. The **first term** in the above expression is **the least squares criterion**. The **additional term** $\lambda\sum_{j=1}^{p}\beta_j^2$ **penalizes large coefficient values** (note that there is no penalty for the intercept term).

How much large coefficient values are penalised is controlled by **the shrinkage parameter** λ: With $\lambda = 0$, the ridge estimator is just the ordinary least squares estimator; with $\lambda = \infty$, the fitted model is a constant model.

An equivalent way of writing the ridge estimate of $\boldsymbol{\beta}$ is to minimise the RSS subject to

$$\sum_{j=1}^{p}\beta_j^2 \leq \boldsymbol{s}. \quad (4.5.2)$$

This makes the size constraint on the parameters explicit. There is a one-to-one correspondence between the shrinkage parameter λ and $\boldsymbol{s}$.

Example: Polynomial Fitting

Recall the effects of fitting an overly complex polynomial. We considered a simulated data set of size

$n = 20$ from the model

$$y_i = x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, 0.25^2)$A polynomial model of order k is

$$y_i = a_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k + \epsilon_i.$$

The true order of the polynomial here is 1.

For our simulated data, consider a tenth-degree polynomial fit by ordinary least squares as well as ridge estimators for $\lambda = 0.05, 0.2, 0.5$.

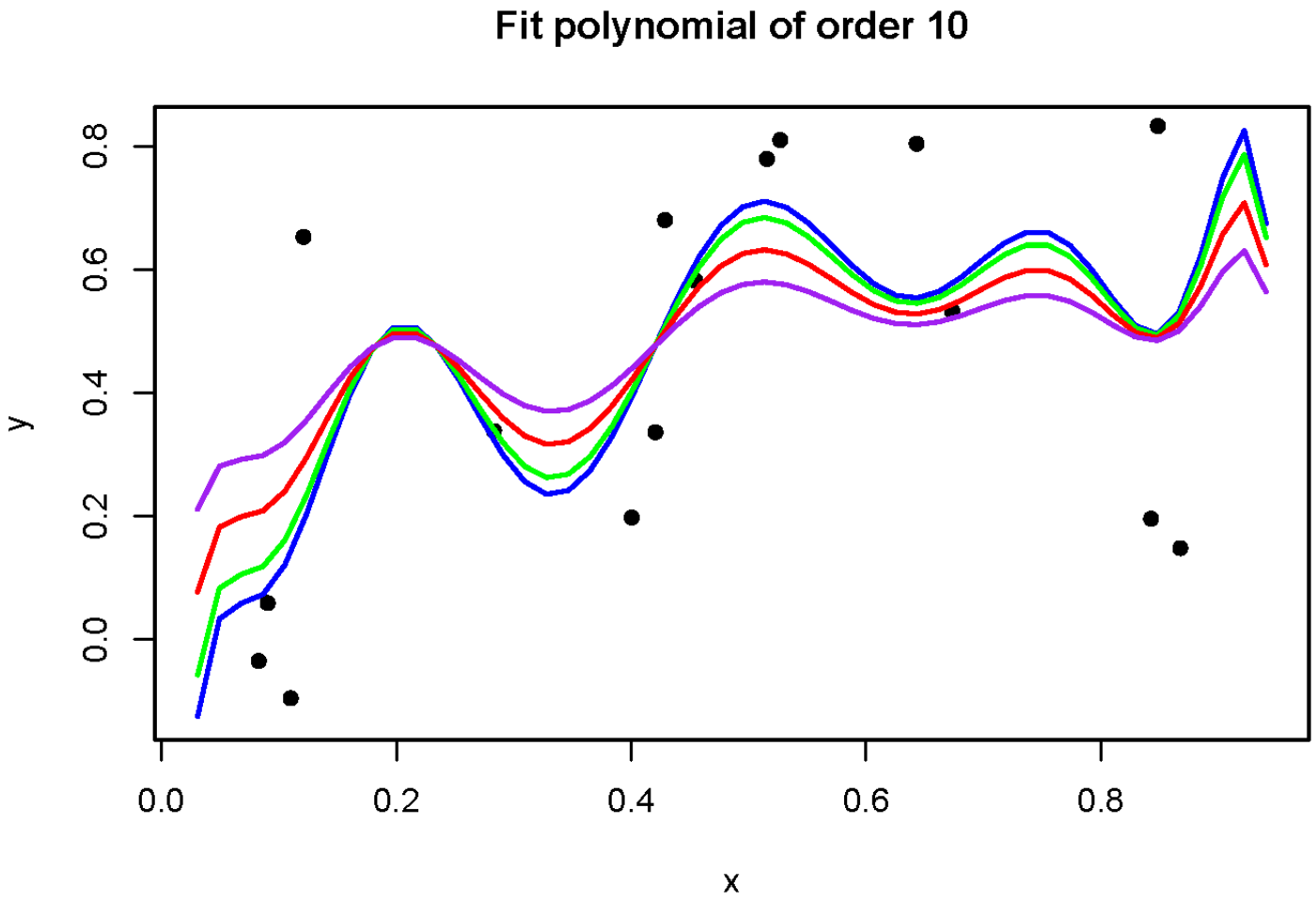


Figure 4.5.1: Fitted polynomials using least squares (blue) and ridge regression with $\lambda = 0.05$ (green), 0.2 (red) and 0.5 (purple).

We can see that **prediction variance is reduced by taking $\lambda > 0$ although for large λ** we **seem to incur substantial prediction bias**.

There are better curve fitting methods than high order polynomials, and this is not a serious

application of ridge regression -- it is only meant to illustrate how ridge regression reduces prediction variance.

Centring and scaling

Let

$$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n,$$

the scaled predictors are defined as

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}.$$

The **two models**

$$y_i = a_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \quad \text{and} \quad y_i = a_0 + \beta_1 z_{i1} + ... + \beta_p z_{ip} + \epsilon_i$$

are equivalent, in the following sense: Shifts of $\boldsymbol{X}_j$ (by $\bar{x}_j$) result in a shift in the intercept β_0, but leave the other coefficients unchanged; division of $\boldsymbol{X}_j$ by S_j effectively multiplies β_j by S_j. Moreover, least squares estimates in one scale linearly rescale to least squares estimates in another. This means that the choice of scale is irrelevant for inference and prediction purposes in linear models without shrinkage.

The ridge solutions are NOT equivalent under scaling of the inputs, and so one normally standardises the inputs before solving (4.5.1). It can be shown that the solution to the criterion (1) can be separated into two parts, after reparametrization using centered inputs: each x_{ij} gets replaced by $z_{ij} = x_{ij} - \bar{x}_j$. We estimate a_0 by $\bar{y}$. The remaining coefficients get estimated by a ridge regression without an intercept, using the centred z_{ij}.

The input matrix Z has p (rather than $p+1$) columns. The ridge regression criterion is than

$$\mathrm{RSS}(\lambda) = \|\boldsymbol{y}^{(c)} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

with solution

$$\hat{\beta} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}^{(c)},$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix and $\boldsymbol{y}^{(c)} = \boldsymbol{y} - \bar{\boldsymbol{y}}$.

Note that the solution adds a positive constant to the diagonal of $\boldsymbol{Z}^\top \boldsymbol{Z}$ before inversion. This makes

the problem nonsingular, even if $\boldsymbol{Z}^\top \boldsymbol{Z}$ is not of full rank.

This means that the least squares estimate exists if the columns of $\boldsymbol{Z}$ (and equivalently $\boldsymbol{X}$) are not linearly independent, or even if $p > N$. In fact, this was the primary motivation for ridge regression when it was first introduced in Hoerl & Kennard (1970).

Example: Hospital manpower Data

To import the hospital **`manpower`** data:

```
require(bestglm)
data(manpower)
```

A data set with $N = 17$ was collected from 17 US Naval hospitals at various sites around the world. The variables are:

$$
\begin{aligned}
Y &= \text{Monthly man hours } (\texttt{Hours}) \\
X_1 &= \text{Average daily patient load } (\texttt{Load}) \\
X_2 &= \text{Monthly X-ray exposures } (\texttt{Xray}) \\
X_3 &= \text{Monthly occupied bed days } (\texttt{BedDays}) \\
X_4 &= \text{Eligible population in the area divided by 1000 } (\texttt{AreaPop}) \\
X_5 &= \text{Average length of patient's stay in days } (\texttt{Stay})
\end{aligned}
$$

The factors appear multiplicative instead of additive, and the response Y is highly right-skewed. For these reasons Y is log-transformed.

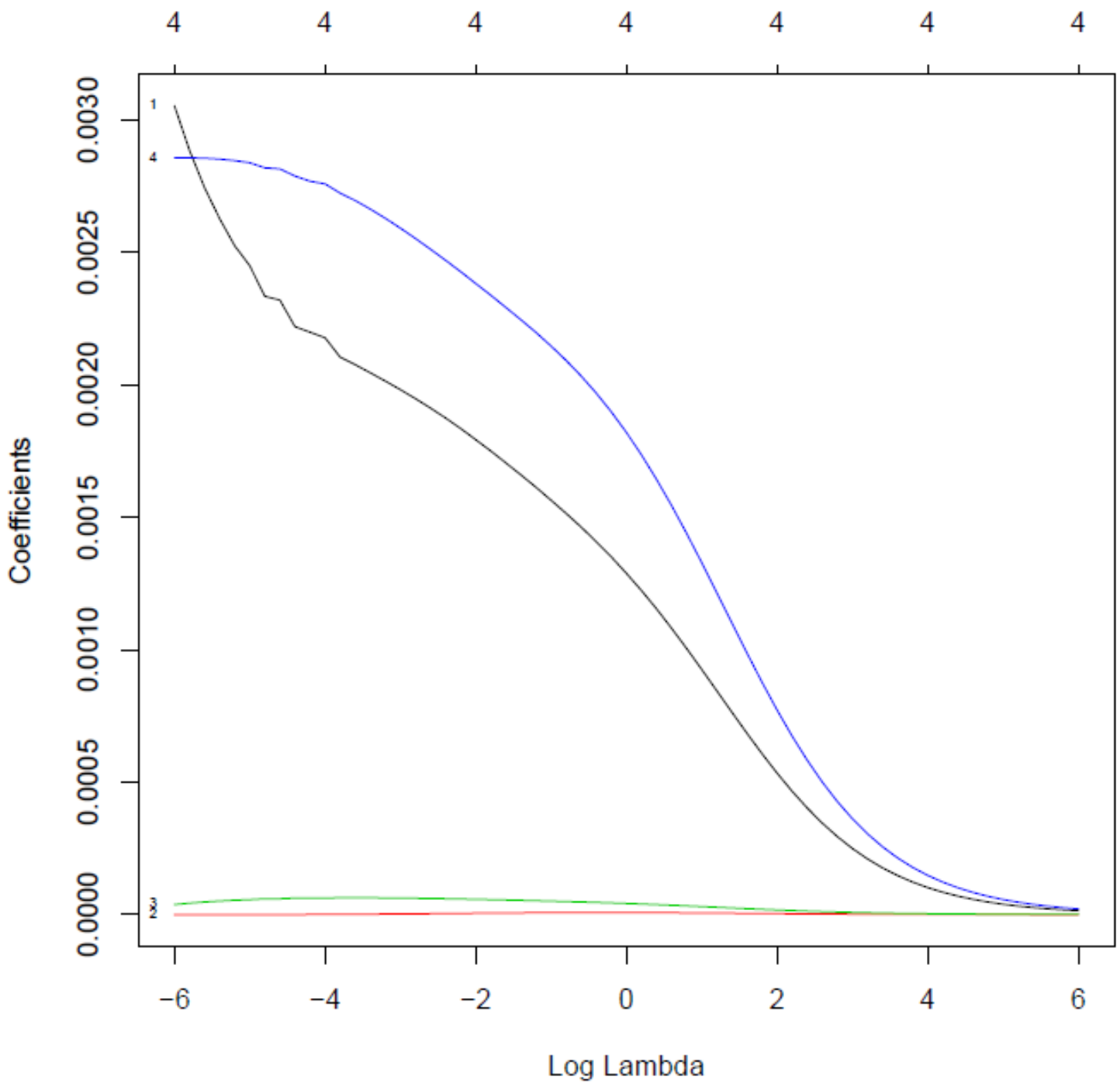
The package **`glmnet`** contains the function $glmnet()$ which performs the fit of a ridge regression model (if **`alpha`** $= 0$) and a Lasso regression model (if **`alpha`** $= 1$).

Predictors are standardised before the fit, but fitted coefficients and intercept are recalculated to the original scale.

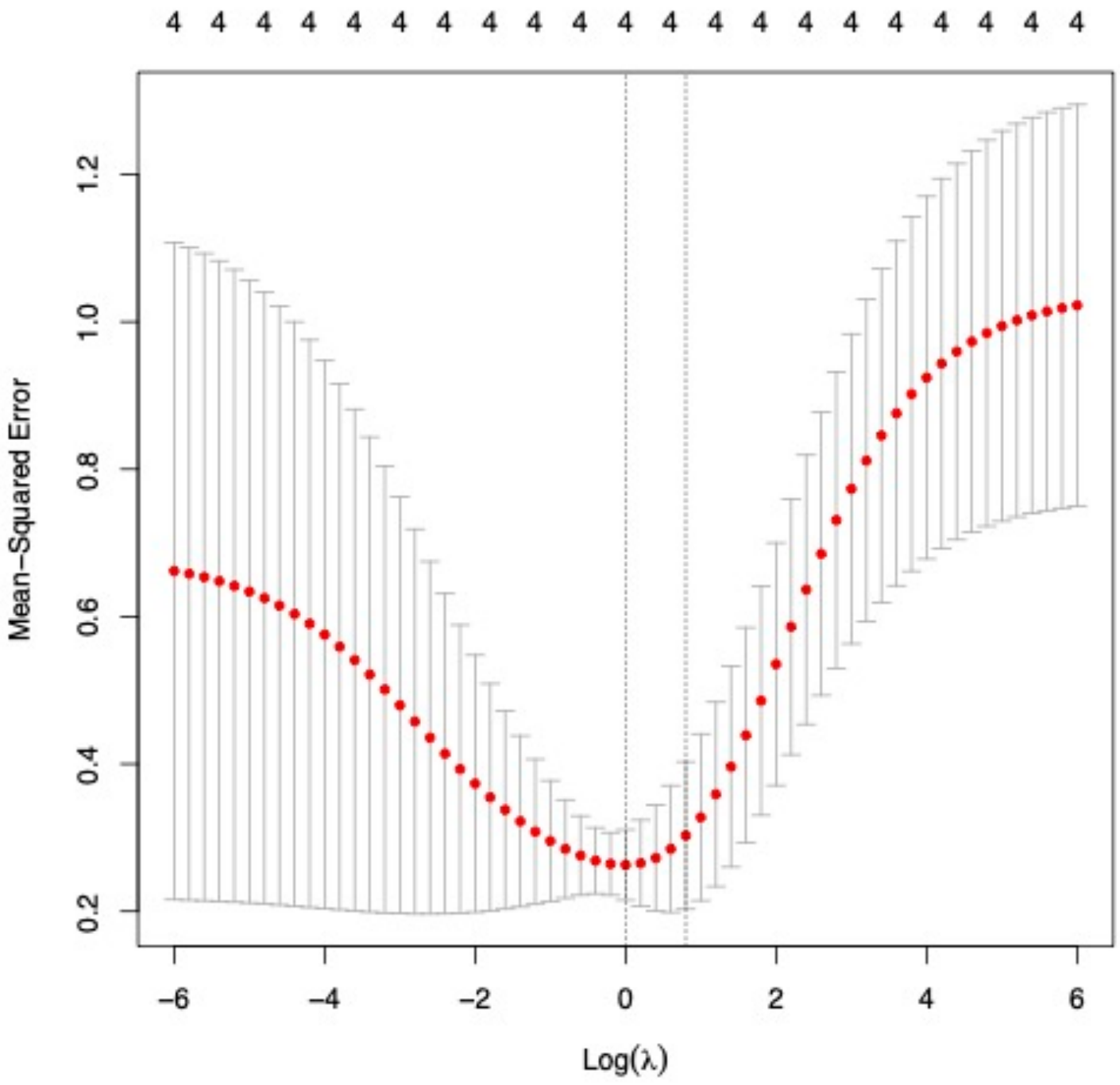
The fit is simultaneously calculated for a range of values of λ. As λ varies, a trajectory of parameters is created, called the **ridge trace**.

It plots the coefficients on the original scale, where magnitude does not necessarily represent variable importance.

```
require(bestglm)
require(glmnet)

x=as.matrix(manpower[,1:4])
y=log(manpower$Hours)
lambda=exp(seq(-6,6,0.2))
fit=glmnet(x,y,alpha=0,lambda=lambda)
plot(fit,xvar="lambda",label=T)
```




```
require(bestglm)
require(glmnet)

x=as.matrix(manpower[,1:4])
y=log(manpower$Hours)
lambda=exp(seq(-6,6,0.2))

set.seed(1) # For reproducibility
cv.fit=cv.glmnet(x,y,alpha=0,nfolds=5,lambda=lambda)
plot(cv.fit)
```



The **cross-validation plot** displays the value of λ which yields the smallest estimated generalisation error. The error bars represent one standard deviation of the CV error estimate, estimated via the standard deviation of the prediction errors across the $m = 5$ folds.

CV error estimates within one standard deviation are deemed interchangeable. A "conservative" **choice of shrinkage is then given by the largest λ within one standard deviation of the minimum**, see the right-hand vertical line.

The minimum λ can be found by:

```
require(bestglm)
```

```
require(glmnet)

x=as.matrix(manpower[,1:4])
y=log(manpower$Hours)
lambda=exp(seq(-6,6,0.2))

set.seed(1) # For reproducibility
cv.fit=cv.glmnet(x,y,alpha=0,nfolds=5,lambda=lambda)
lambda=cv.fit$lambda.min
lambda
```

```
[1] 1
```

The conservative choice would be:

```
require(bestglm)
require(glmnet)

x=as.matrix(manpower[,1:4])
y=log(manpower$Hours)
lambda=exp(seq(-6,6,0.2))

set.seed(1) # For reproducibility
cv.fit=cv.glmnet(x,y,alpha=0,nfolds=5,lambda=lambda)
lambda=cv.fit$lambda.1se
lambda
```

```
[1] 2.225541
```

The coefficients corresponding to these two model choices are

```
require(bestglm)
require(glmnet)

x=as.matrix(manpower[,1:4])
y=log(manpower$Hours)
lambda=exp(seq(-6,6,0.2))

fit=glmnet(x,y,alpha=0,lambda=lambda)

set.seed(1) # For reproducibility
cv.fit=cv.glmnet(x,y,alpha=0,nfolds=5,lambda=lambda)

fit$beta[,cv.fit$index[,1]]
fit$a0[cv.fit$index[,1]]
```

```
                 s30          s26
Load    1.288189e-03 1.004304e-03
Xray    7.607377e-06 6.482554e-06
BedDays 4.204629e-05 3.283089e-05
AreaPop 1.820190e-03 1.438385e-03
     s30      s26
```

```
7.282781 7.427188
```

Note that here the **intercepts** are reported on the original scale and hence **depend on the shrinkage.**

Variance reduction through shrinkage

Singular value decomposition of the centred input matrix $\boldsymbol{Z}$ gives us some additional insight into the nature of ridge regression.

We can compute the **singular value decomposition** (SVD) of the $N \times p$ centered data matrix $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

where $\boldsymbol{U}$ is a $N \times p$ matrix with orthonormal columns that span the column space of $\boldsymbol{Z}$, $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix, and $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with elements d_j ordered such that $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$.

Using the SVD we can rewrite the expression for $\hat{\boldsymbol{\beta}}$ as follows

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{y} \\
&= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + \lambda \boldsymbol{V}\boldsymbol{V}^\top)^{-1} \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{y} \\
&= (\boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})\boldsymbol{V}^\top)^{-1} \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{y} \\
&= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{y}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{y}.
\end{aligned}$$

We note that $\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}$ is a diagonal matrix with elements given by

$$\frac{d_j^2}{d_j^2 + \lambda}$$

and the vector $\boldsymbol{U}^\top \boldsymbol{y}$ is the coordinates of the vector $\boldsymbol{y}$ in the basis spanned by the p-columns of $\boldsymbol{U}$.

Thus

$$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \sum_{j=1}^{n} \boldsymbol{U}_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) \boldsymbol{U}_j^\top \boldsymbol{y}$$

and it results that the inner products $\boldsymbol{U}_j^\top \boldsymbol{y}$ are scaled by the factors $\frac{d_j^2}{d_j^2 + \lambda}$ in the ridge regression.

Note that the hat matrix $\boldsymbol{H}_\lambda$, defined via $\boldsymbol{H}_\lambda \boldsymbol{y} = \hat{\boldsymbol{y}}$ depends on λ and is equal to

$$\boldsymbol{H}_\lambda = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top = \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^\top.$$

The fitted values for ridge regression satisfy

$$\mathrm{Cov}(\hat{\boldsymbol{y}}) = \sigma^2 \boldsymbol{H}_\lambda^2 = \sigma^2 \boldsymbol{U} \mathrm{diag}\left(d_1^4/(d_1^2+\lambda)^2, \ldots, d_p^4/(d_p^2+\lambda)^2\right) \boldsymbol{U}^\top.$$

Therefore, we can now see that shrinkage indeed reduces the predictive variance.

The effective number of parameters for ridge regression is

$$\mathrm{tr}(\boldsymbol{H}_\lambda) = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}.$$

Hence as λ increases from 0 to $+\infty$, the effective number of parameters decreases continuously and monotonically from p to 0.

Activity: Parameter estimates dependence on λ

Question *Submitted Mar 16th 2023 at 11:24:33 pm*

Consider the ridge regression problem

$$\hat{\beta} = \mathbf{argmin}_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}.$$

Show that this problem is equivalent to the problem

$$\hat{\beta}^c = \mathbf{argmin}_{\beta_c}\{\sum_{i=1}^{N}(y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda \sum_{j=1}^{p}(\beta_j^c)^2\}$$

Give the correspondence between β_0^c and the original β_0. Calculate $\hat{\beta}_0^c$ and $\hat{\beta}_0$. Explain how does $\hat{\beta}_0$ in the original problem depend on the choice of the shrinkage parameter λ.

asdf

Activity: Estimation in Ridge Regression

Question *Submitted Mar 16th 2023 at 11:27:45 pm*

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the standardized matrix $\boldsymbol{X}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$, and augment $\boldsymbol{y}$ with p zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero.

We recognise this expression as the solution to the ridge regression problem.

asdf

Activity: Ridge estimator

Question *Submitted Mar 16th 2023 at 11:28:55 pm*

Show that $||\hat{\beta}^{\text{ridge}}||^2$ increases as $\lambda \to 0$.

asdf

Activity in R: Ridge Regression

The data set **hald** was first printed in the article by Woods, Steinour and Starke (1932), "Effect of Composition of Portland Cement on Heat Evolved during Hardening," *Industrial and Engineering Chemistry*, 24, pp. 1207-14.

With these data we are interested in predicting the response **y** (heat evolved in calories per gram of cement) in terms of the predictors **x1**, **x2**, **x3** and **x4**.

Construct pairwise scatter plots of the predictor variables. Do you think multicollinearity is present in these data?

Find the ridge estimators for the coefficients of the standardised predictors when $\lambda = 0.02$, using ordinary least squares regression on an augmented dataset. We augment the standardized matrix $\boldsymbol{X}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$, and augment $\boldsymbol{y}$ with p zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero.

Compare the parameter estimates with the estimates obtained from glmnet() function. Are they approximately the same? Note that to compare glmnet() to other methods authors of glmnet() suggest to standardise the response variable, too.

Activity

In this exercise we will analyse the properties of the ridge regression through an application to the `credit` dataset available in the `ISLR` package in `R`.

1. Load the dataset. Define the response vector by `y` and the design matrix `X` using the function `model.matrix()`. The function `model.matrix()` prepares the predictors to be included in the ridge regression via the `glmnet()` in the correct format (numerical or quantitative outputs only).
2. Define a grid for the tuning parameter λ, ranging from $\lambda = 10^{-2}$ to $\lambda = 10^{5}$ in decreasing order. Perform a ridge regression over the defined grid of λ values. This covers a range of scenarios from the null model containing only the intercept, to the least square fit. Use the function `plot(, xvar = "lambda")` to plot the ridge regression coefficients for the predictor `Income` as a function of λ.
3. When λ is small, ridge regression gives similar answers to ordinary regression. Check this assertion by comparing the estimates for the ordinary regression and ridge regression with the smallest λ considered.
4. When λ is large, ridge regression shrinks the parameter estimates when compared to the least squares estimates. Check this assertion by comparing the estimates for the ordinary regression and ridge regression with the largest λ considered.
5. Split the data equally into training and test sets using `set.seed(1)`. When λ is small, we get only small improvement in the test error over linear regression, while when λ is large we see a definite improvement, λ cannot be too large though. Check this assertions by computing the test MSE for the ordinary regression and the ridge regression penalty parameter fixed to $\lambda = 0.01, 7$ and 20.
6. In general, rather than arbitrarily choosing $\lambda = 7$, it would be better to use cross-validation to choose the tuning parameter λ. We can do this using the built-in cross validation function `cv.glmnet()`. By default the function performs 10-fold cross-validation, though it can be changed using the argument `folds`. Use 5-fold cross validation to select the optimal tuning parameter λ and plot the output (MSE as function of $\log(\lambda)$). What is the value of the tuning parameter than results in the smallest cross-validation error and what is the associated test MSE value? Fore reproducibility use `set.seed(2)`.
7. From the plot in the previous question, the $\lambda_{\min}$ seems to be suspiciously close to the boundary of the search grid. We therefore decide to re-run the cross-validation algorithm using the search grid that we initially defined. Do you observe any changes? Fore reproducibility use `set.seed(2)`.

Additional Activity

Question 1 *Submitted Mar 16th 2023 at 11:32:14 pm*

Which of the following are benefits of using ridge regression:

- [x] reduced model complexity;
- [x] smaller variance;
- [x] smaller mean squared prediction error.

Question 2 *Submitted Mar 16th 2023 at 11:32:23 pm*

Which of the following statements are ture of the parameter λ in the additional term $\lambda \sum_{j=1}^{p} \beta_j^2$ in the ridge regression is:

- [ ] when $\lambda = 0$ the ridge regression has no solution;
- [x] when $\lambda = \infty$ the fitted model is a constant model;
- [x] λ controls how much the coefficient values are penalized in the ridge regression estimation problem.

Question 3 *Submitted Mar 16th 2023 at 11:32:26 pm*

Is the following true: *The ridge trace is a plot displaying estimated coefficients in the ridge regression fit plotted for various values of the shrinkage parameter λ.*

- (•) Yes
- ( ) No