Ridge Regression and the Lasso

Scott Powers

STaRT@Rice 2024

Outline

Part I: Theory (60 minutes)

- Ridge regression
- The lasso
- Comparison
- Cross-validation

Part II: Application (30 minutes)

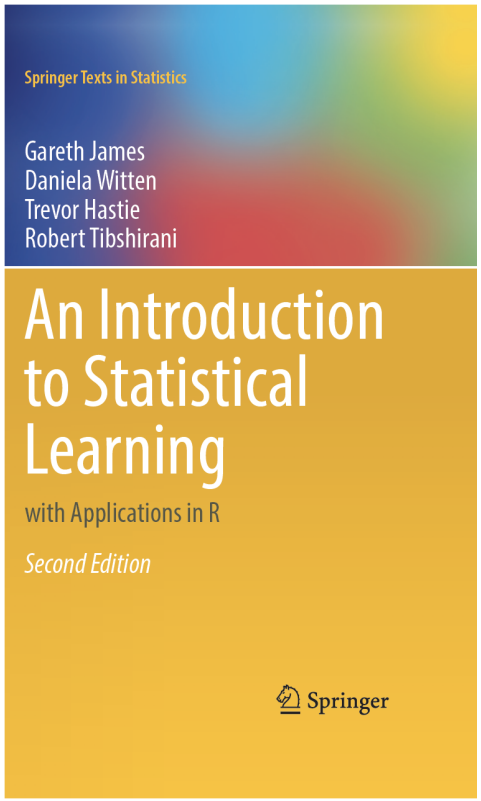
- The glmnet package in R
- Application to basketball

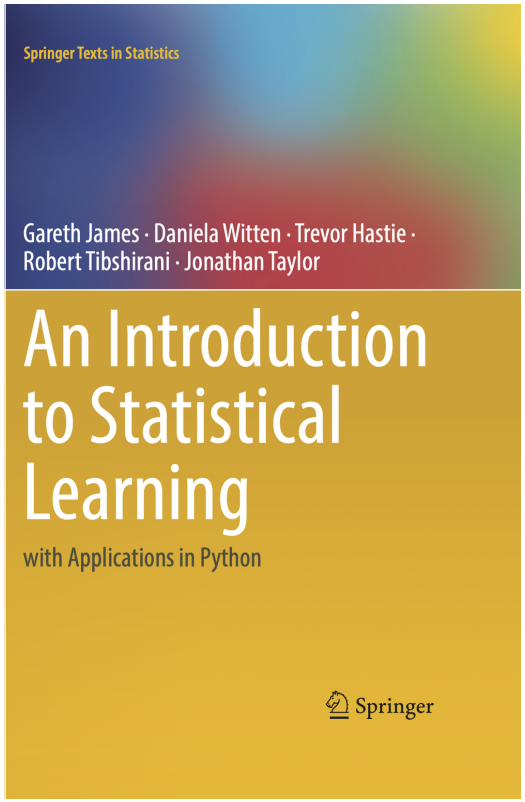
Part III: Practice (90 minutes)

- R tutorial

Questions are highly encouraged!

An Introduction to Statistical Learning





- Available for FREE online: `statlearning.com`
- Today we'll cover mostly Section 6.2 (in R)

Linear regression

Standard linear model:

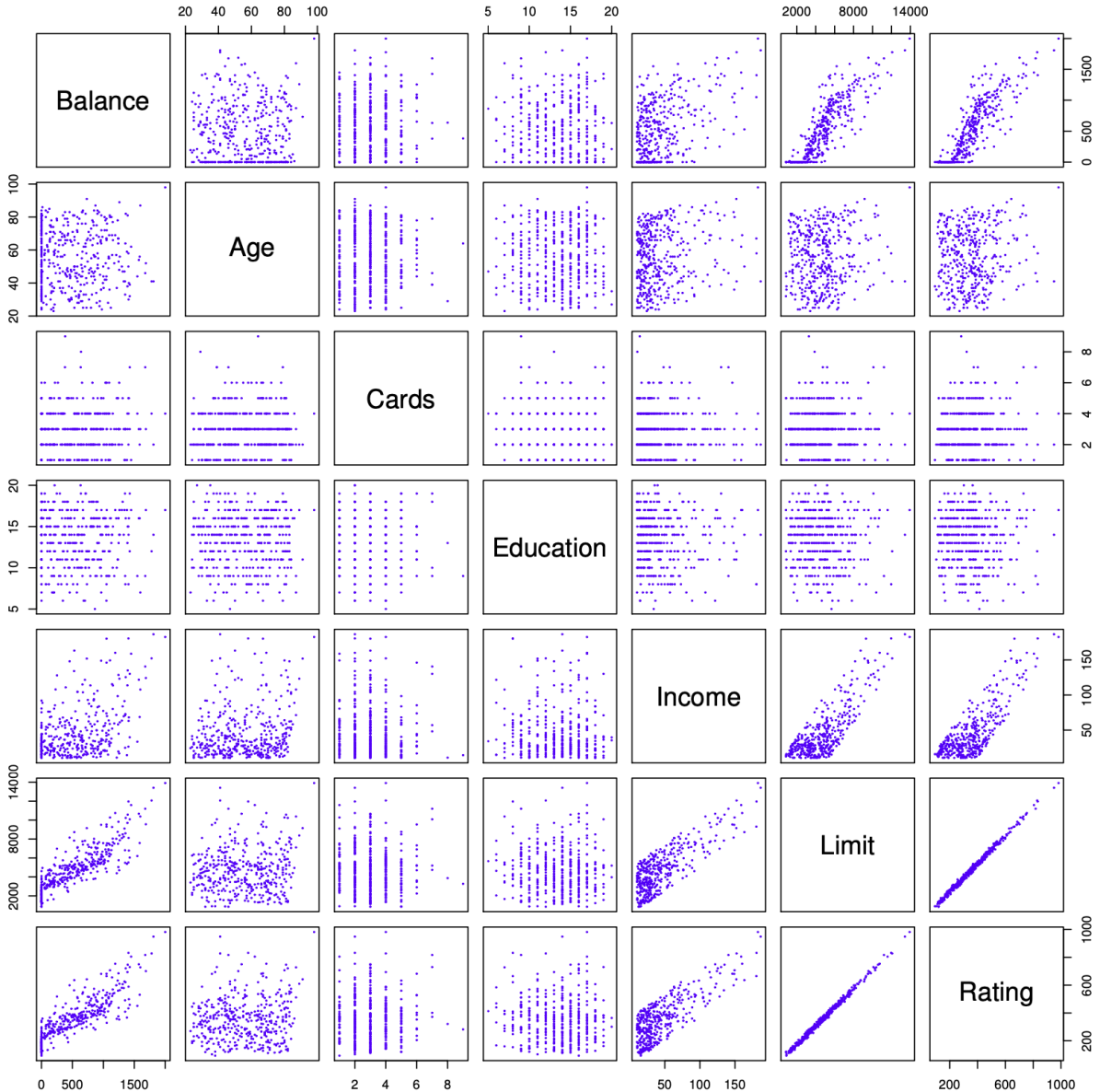
$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

Least squares (the typical fitting procedure):

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

- Today we will discuss alternative fitting procedures
- Considerations:
  - Prediction accuracy
  - Model interpretability

Preview: The Credit data set



Credit: ISLR, page 84

Alternative #1: Best subset selection

Idea: Select the best model from among the 2^p possible models (according to which variables are included in the model)

Algorithm:

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model predicts the sample mean for each observation.
2. For $k = 1, 2, ..., p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Best is defined as having the highest R^2.
3. Select a single best model from among $\mathcal{M}_0, ..., \mathcal{M}_p$ using using adjusted R^2 or some alternative criterion (e.g. C_p, BIC).

- Computationally infeasible for $p > 40$ (trillions of models)

Example: Best subset selection on Credit data set

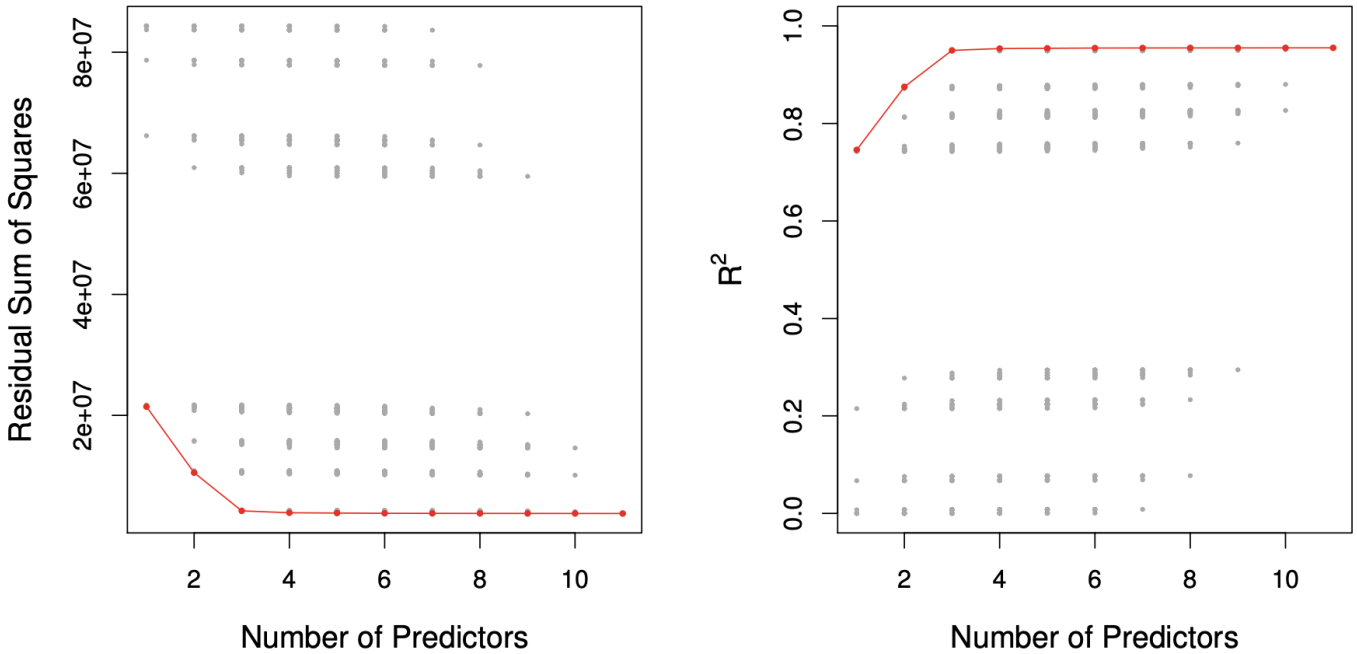


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Credit: ISLR, page 229

Example: Best subset selection on Credit data set

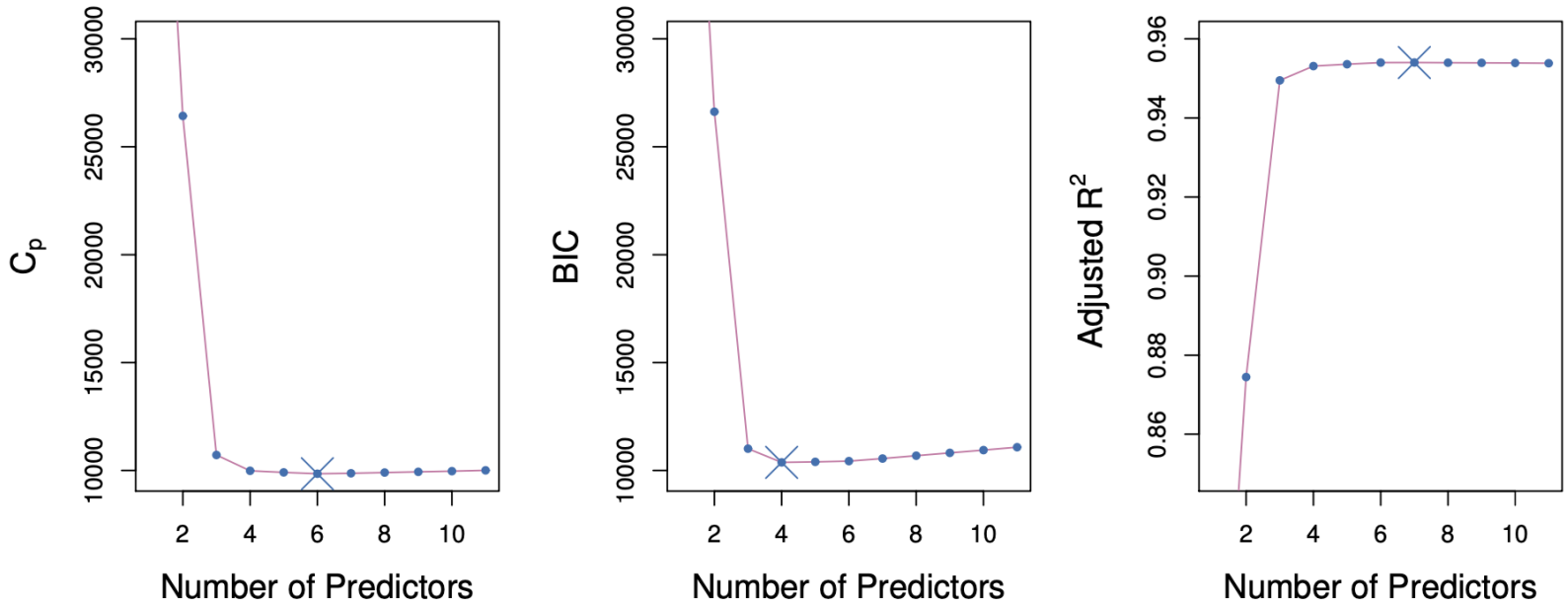


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

Credit: ISLR, page 233

Alternative #2: Backward stepwise selection

Idea: Approximate best subset selection with something feasible

Algorithm:

1. Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.
2. For $k = p, p-1, ..., 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Best is defined as having the highest R^2.
3. Select a single best model from among $\mathcal{M}_0, ..., \mathcal{M}_p$ using using adjusted R^2 or some alternative criterion (e.g. C_p, BIC).

- Similar to repeatedly dropping predictor with highest p-value

Alternative #3: Ridge regression

Least squares:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

Ridge regression:

$$\hat{\beta}_{\lambda}^{R} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$\lambda \geq 0$ is a *tuning parameter*

- Ridge regression introduces a *shrinkage penalty*
- Computation is very fast (similar to least squares)

Example: Ridge regression on Credit data set

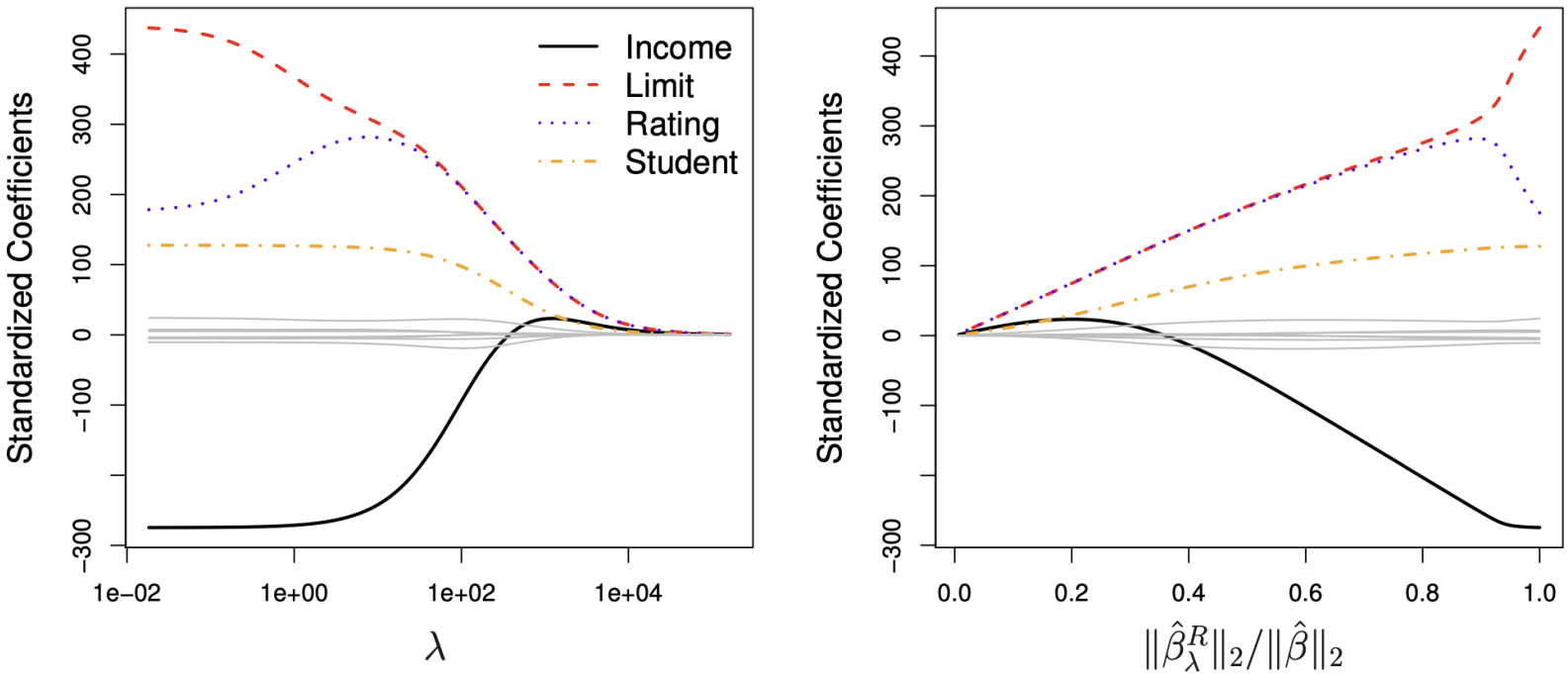


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Credit: ISLR, page 238

The bias-variance trade-off

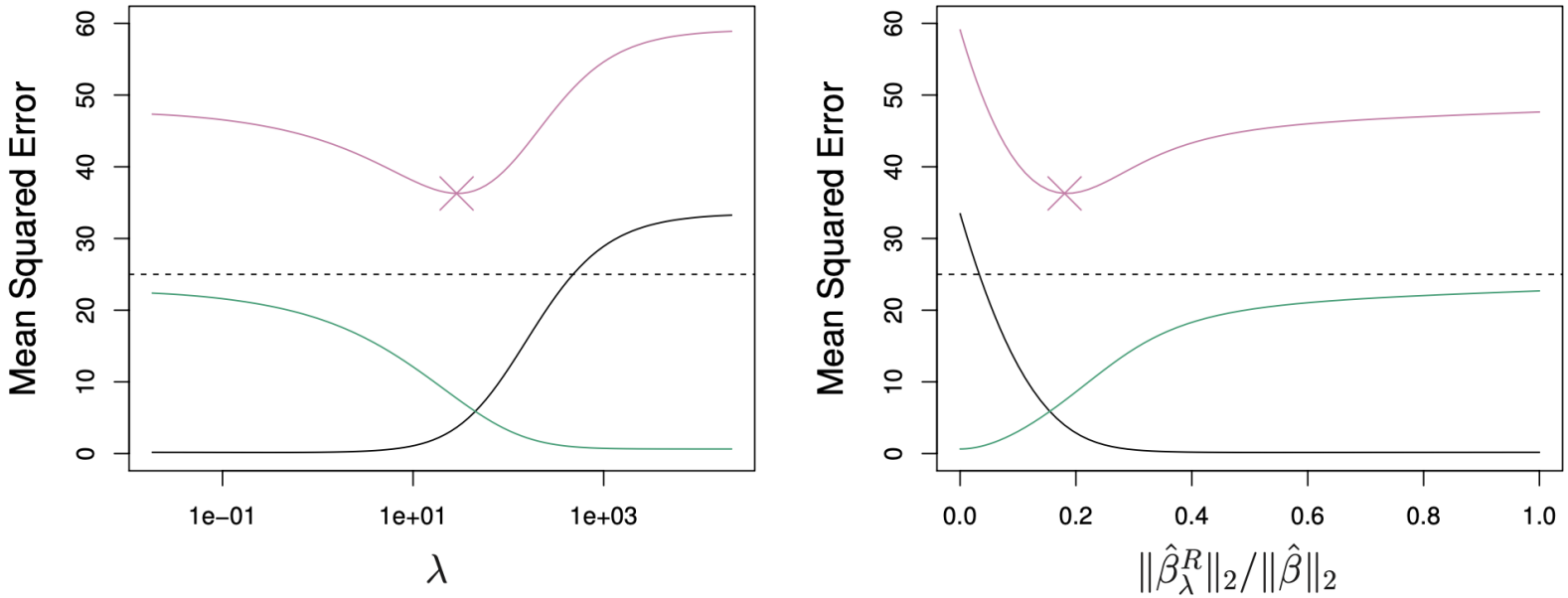


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Credit: ISLR, page 240

Exercise (ISLR, page 284)

Suppose we estimate the coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

(a) As we increase λ from 0, the training RSS will: (choose correct answer)

i. Decrease initially, and then eventually increase in a U shape.
ii. Increase initially, and then eventually decrease in an inverted U.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

The lasso

$$\hat{\beta}_\lambda^L = \arg\min_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Achieves *variable selection* (drops some variables from model)
- Not as fast as ridge regression, but also very fast

Example: The lasso on Credit data set

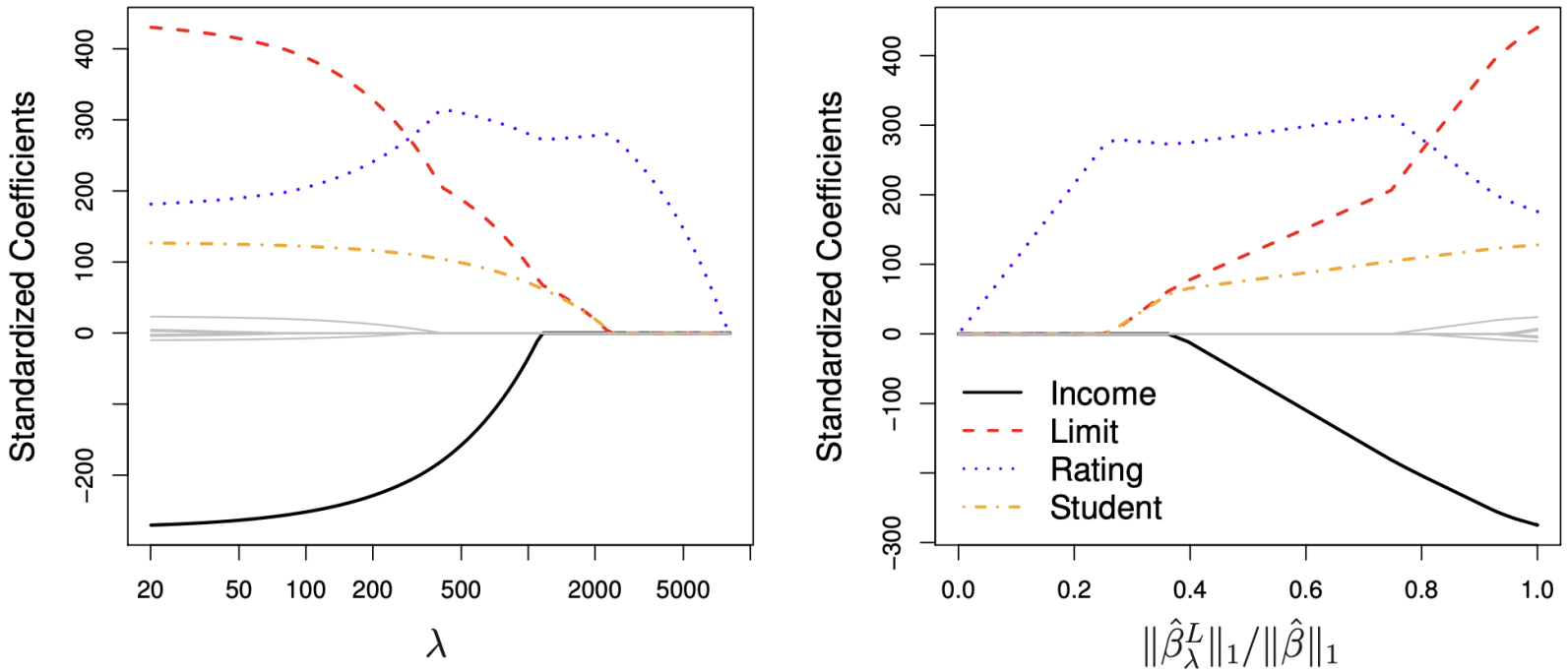


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.

Credit: ISLR, page 243

Exercise (ISLR, page 284)

The lasso, relative to least squares, is: (choose correct answer)

i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

Another formulation

Ridge regression:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum \beta_j x_{ij} \right)^2 \right\} \text{ s.t. } \sum_{j=1}^{p} \beta_j^2 \leq s$$

The lasso:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum \beta_j x_{ij} \right)^2 \right\} \text{ s.t. } \sum_{j=1}^{p} |\beta_j| \leq s$$

Best subset selection:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum \beta_j x_{ij} \right)^2 \right\} \text{ s.t. } \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \leq s$$

Variable selection by the lasso

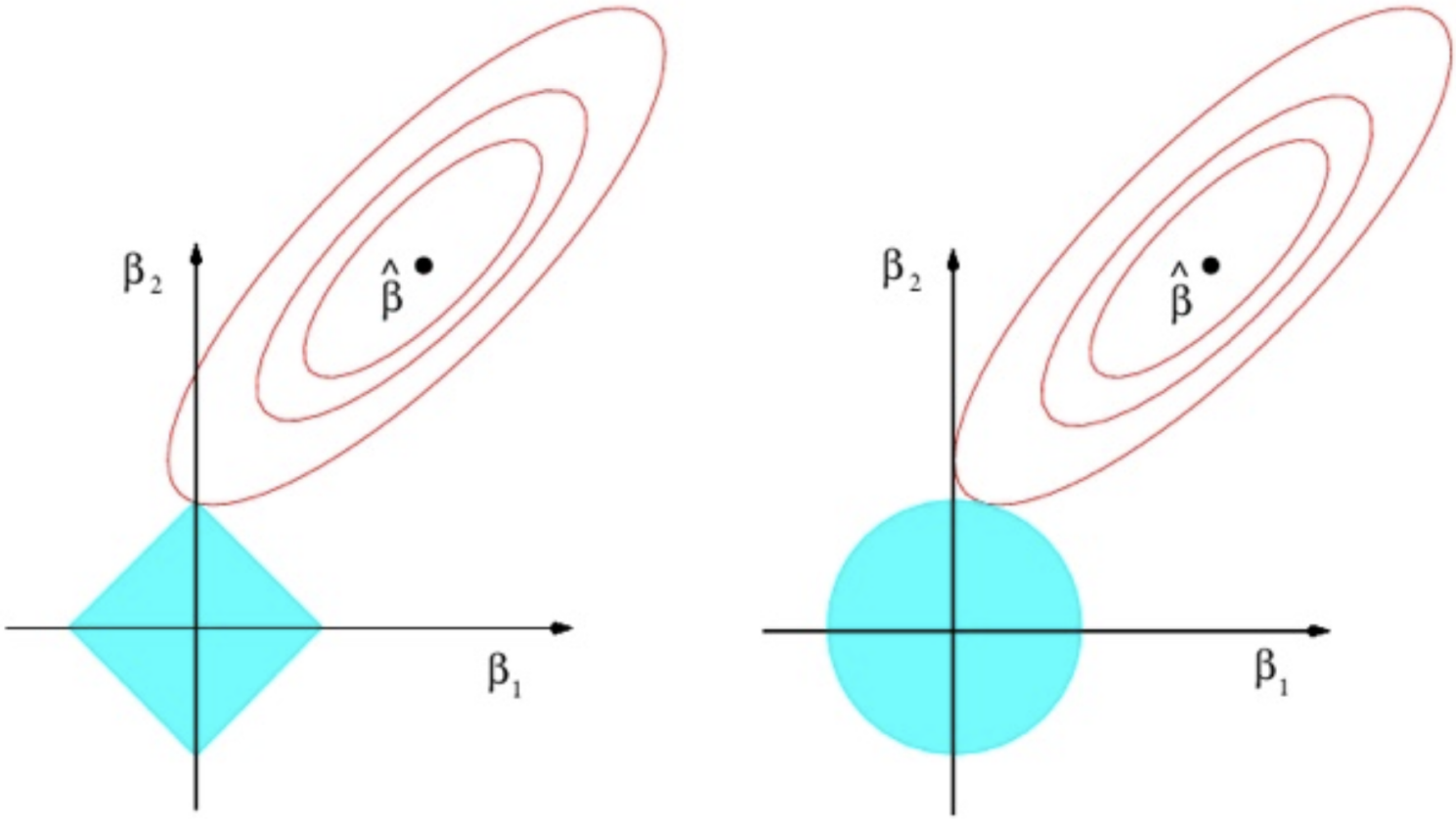


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Credit: ISLR, page 245

A simple special case

$n = p$, one coefficent β_j corresponding to each observation y_j

Least squares:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 \quad \Rightarrow \quad \hat{\beta}_j = y_j$$

Ridge regression:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \Rightarrow \quad \hat{\beta}_j^R = y_j/(1+\lambda)$$

The lasso:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad \Rightarrow \quad \hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2 \\ 0 & \text{if } |y_j| \leq \lambda/2 \end{cases}$$

A simple special case

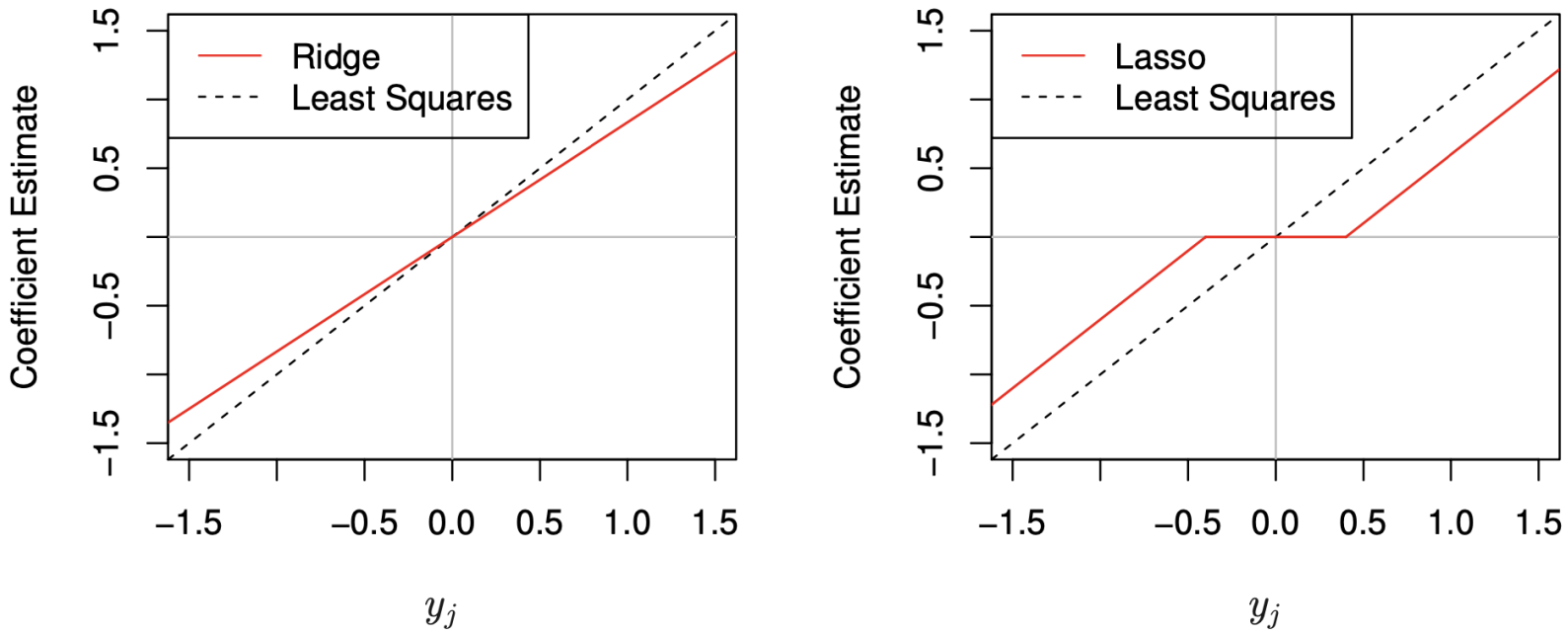


FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

Credit: ISLR, page 248

Exercise (ISLR, page 285)

(a) Assume $p = 1$. For some choice of y_1 and $\lambda > 0$, plot

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

as a function of β_1. Your plot should agree that (1) is minimized by

$$\hat{\beta}_j^R = y_j/(1+\lambda).$$

(b) Assume $p = 1$. For some choice of y_1 and $\lambda > 0$, plot

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

as a function of β_1. Your plot should agree that (2) is minimized by

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2 \\ 0 & \text{if } |y_j| \leq \lambda/2 \end{cases}.$$

Selecting the tuning parameter

How to select the tuning parameter λ?

Validation set approach:

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

Credit: ISLR, page 199

Leave-one-out cross-validation (LOOCV)

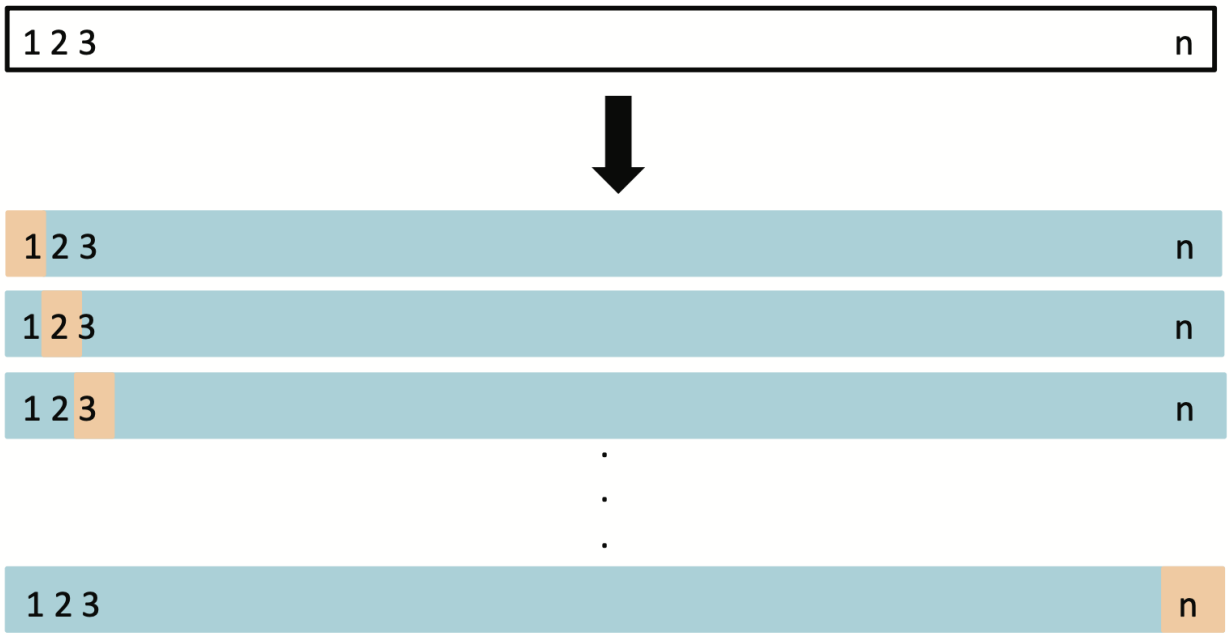


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSEs. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

Credit: ISLR, page 201

k-fold cross-validation

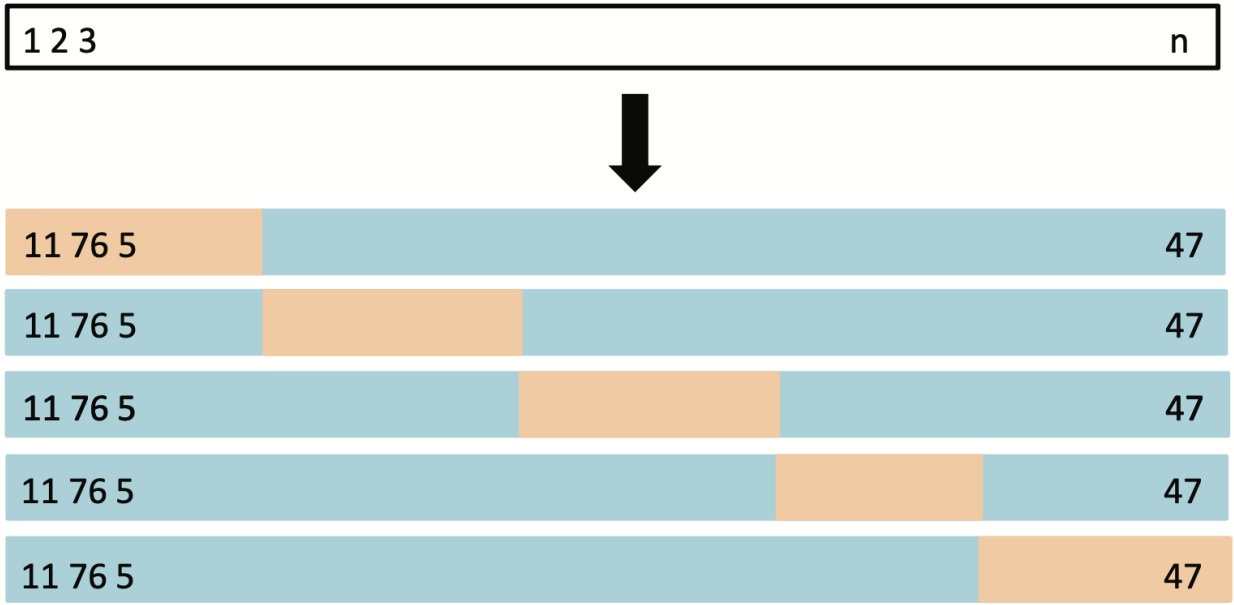


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

Credit: ISLR, page 203

Example: Cross-validation on the Credit data set

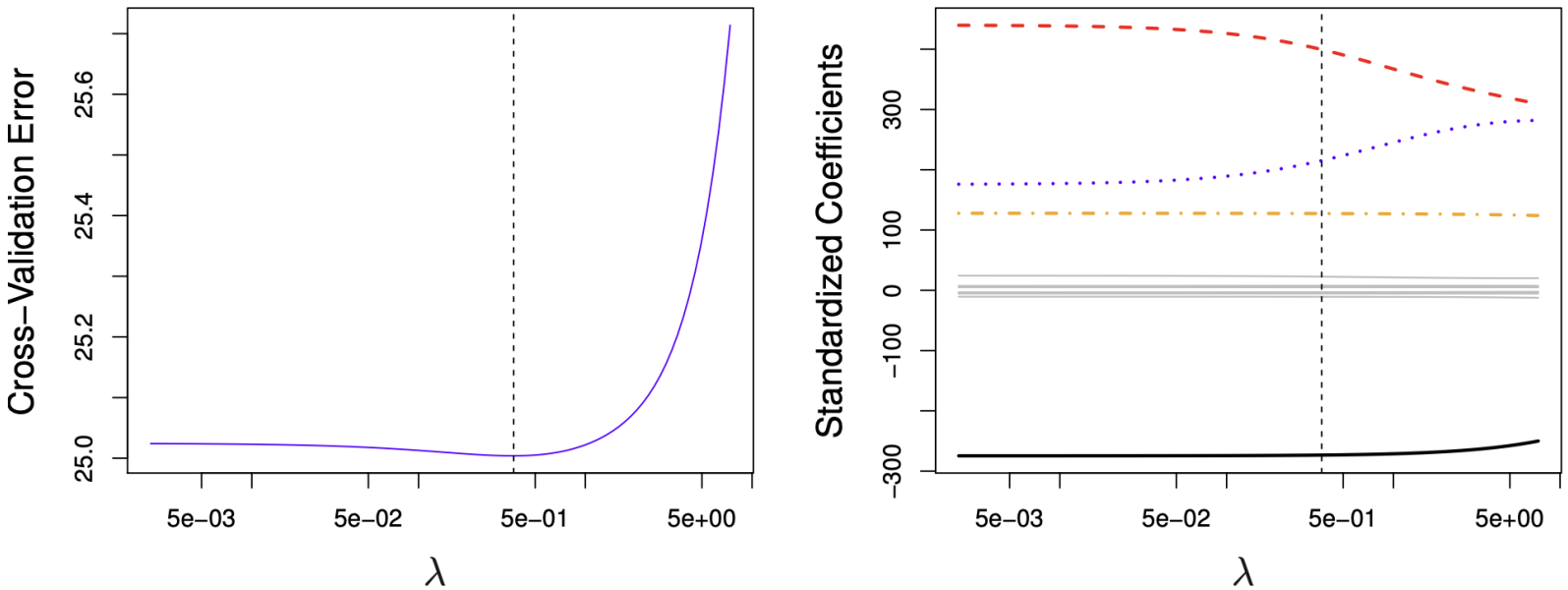


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various values of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

Credit: ISLR, page 251

Exercise (ISLR, page 220)

(a) Explain how k-fold cross-validation is implemented.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to:

   i. The validation set approach?

   ii. Leave-one-out cross-validation (LOOCV)?

(c) What is the trade-off to consider when choosing k for k-fold cross-validation?

The glmnet R package

Elastic net regression:

$$\sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + (1-\alpha) \cdot \lambda \sum_{j=1}^{p} \beta_j^2 + \alpha \cdot \lambda \sum_{j=1}^{p} |\beta_j|$$

In R:

```
glmnet::glmnet(
  x,
  y,
  weights = NULL,
  alpha = 1,
  lambda = NULL,
  standardize = TRUE,
  intercept = TRUE,
  ...
)
```

```
glmnet::cv.glmnet(
  x,
  y,
  weights = NULL,
  lambda = NULL,
  nfolds = 10,
  foldid = NULL,
  parallel = FALSE,
  ...
)
```

Application: Basketball Analytics

Regularized Adjusted Plus-Minus (RAPM)

We have *stints* (periods of no substitutions) labeled $i = 1, ..., n$

- w_i is the length (in minutes) of stint i
- y_i is the score differential (home $-$ away) during stint i
- H_i is the set of home players on the floor during stint i
- A_i is the set of away players on the floor during stint i

Model:

$$Y_i/w_i = \beta_0 + \sum_{j \in H_i} \beta_j - \sum_{j' \in A_i} \beta_{j'} + \epsilon_i$$

Fitting procedure (ridge regression):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} w_i \left(y_i - \left(\beta_0 + \sum_{j \in H_i} \beta_j - \sum_{j' \in A_i} \beta_{j'} \right) \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Framing RAPM as ridge regression

We can re-frame this using

$$x_{ij} = \begin{cases} +1 & \text{if player } j \text{ is on the floor for the } \textit{home} \text{ team during stint } i \\ -1 & \text{if player } j \text{ is on the floor for the } \textit{away} \text{ team during stint } i \\ 0 & \text{if player } j \text{ is } \textit{off} \text{ the floor during stint } i \end{cases}$$

Ridge regression:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} w_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Sample X matrix for RAPM

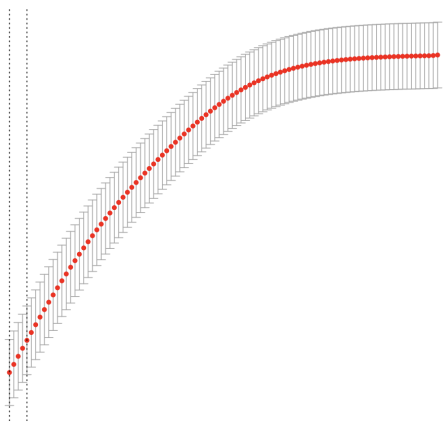
$$X = \begin{bmatrix}
+1 & +1 & +1 & +1 & +1 & 0 & \dots & -1 & -1 & -1 & -1 & -1 & \dots & 0 \\
+1 & +1 & +1 & +1 & 0 & +1 & \dots & -1 & -1 & -1 & -1 & -1 & \dots & 0 \\
+1 & +1 & +1 & +1 & 0 & +1 & \dots & -1 & -1 & -1 & 0 & 0 & \dots & 0 \\
0 & +1 & +1 & +1 & 0 & +1 & \dots & 0 & -1 & -1 & 0 & 0 & \dots & 0 \\
0 & +1 & +1 & +1 & +1 & 0 & \dots & 0 & -1 & -1 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & +1 & +1 & 0 & \dots & 0 & -1 & -1 & -1 & -1 & \dots & 0 \\
+1 & 0 & 0 & +1 & +1 & 0 & \dots & -1 & -1 & -1 & -1 & -1 & \dots & 0 \\
+1 & 0 & 0 & 0 & +1 & +1 & \dots & -1 & 0 & 0 & -1 & -1 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & +1 & +1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\
+1 & 0 & 0 & +1 & +1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\
+1 & 0 & 0 & +1 & 0 & +1 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0
\end{bmatrix}$$

- This matrix is very *sparse* (most of the entries are zero)

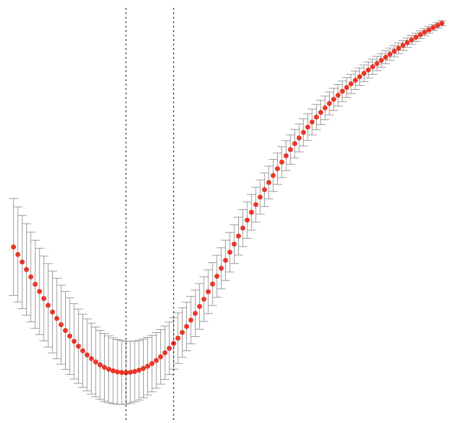
3 glmnet lessons I learned the hard way

```
glmnet::glmnet(
  x,
  y,
  weights = NULL,
  alpha = 1,
  lambda = NULL,
  standardize = TRUE,
  intercept = TRUE,
  ...
)
```

1. For ridge regression, use `alpha = 0`.
2. Check whether the default `lambda` range includes the optimal value for you



Example of bad λ range



Example of good λ range

3. For RAPM, `standardize = FALSE`.

Google Colab

Ground rules:

1. After opening the Google Colab notebook, select "File" $\rightarrow$ "Save a Copy in Drive". This will open a new tab with a version you can edit and save. If you keep the original open in a separate tab, you can refresh it to see updates I post.
2. Next, select "Runtime" $\rightarrow$ "Change runtime type", and change the runtime type from Python 3 to R.
3. Now you can get started! Raise your hand or ask a neighbor when you feel stuck or lost. The first two code blocks take several minutes to run, so start thinking ahead while they run.