STATISTICAL LEARNING

CHAPTER 6: LINEAR MODEL SELECTION AND REGULARIZATION

INSTRUCTOR: SEOKHO LEE

HANKUK UNIVERSITY OF FOREIGN STUDIES

2015 SPRING

- The standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{6.1}$$

- Extension from the standard linear model
  - In Chapter 7 we generalize (6.1) in order to accommodate nonlinear, but still additive, relationships
  - In Chapter 8 we consider even more general non-linear models
  - However, linear model has distinct advantages in terms of inference and is often surprisingly competitive

- In this chapter, the simple linear model will be improved by replacing plain least squares fitting with some alternative fitting procedure
  - **Prediction Accuracy**
    - If n is not much larger than p, the there can be a lot of variability in the least squares fit, resulting in overfitting and poor predictions
    - If $p > n$, there is no longer a unique least squares coefficient estimate: the variance is *infinite*. By **constraining** or **shrinking** the estimated coefficients, we can often substantially reduce the variance
  - **Model Interpretability**
    - Including the *irrelevant* variables leads to unnecessary complexity in the resulting model
    - It is better to consider some approaches for automatically performing **feature selection** or **variable selection**

- Three important classes of methods
  - **Subset Selection**
    - Identify a subset of the p predictors that we believe to be related to the response
  - **Shrinkage** or **Regularization**
    - Involve all p predictors but the estimated coefficients are shrunken toward zero relative to the least squares estimates
    - Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero
  - **Dimension Reduction**
    - **Project** the p predictors into an M-dimensional subspace, where $M < p$.
    - This is achieved by computing M different **linear combinations**, or **projections**, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares

Best Subset Selection

- **Best subset selection** identifies the **best model** from models with all possible combination of the p predictors
- It is not trivial to consider all 2^p possibilities unless p is very small

Algorithm 6.1 Best subset selection

```
1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$,
    (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 $\mathcal{M}_k$. Here best is defined as having the smallest RSS, or equivalently largest $R^2$
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$, AIC, BIC or adjusted $R^2$
```

Best Subset Selection

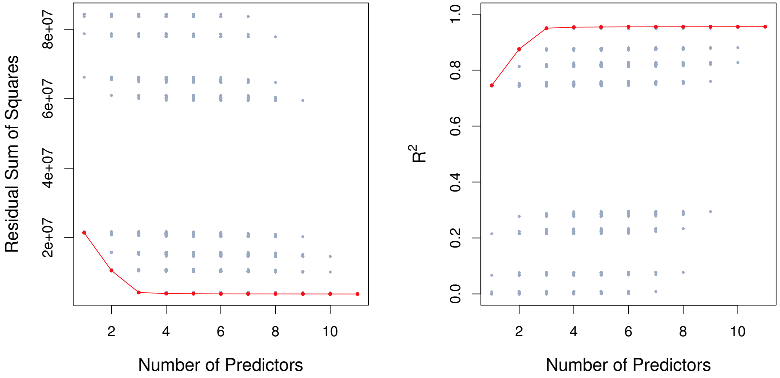

Figure 6.1: For each possible model containing a subset of the ten predictors in the Credit data set, the RSS and R^2 are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and R^2. Though the data set contains only ten predictors, the x-axis ranges from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

Best Subset Selection

- In the case of logistic regression, instead of ordering models by RSS in Step 2 of Algorithm 6.1, we instead use the **deviance**, a measure that plays the role of RSS for a broader class of models
- In general, there are 2^p models that involve subsets of p predictors
  - If $p = 10$, there are 1,024 possible models to be considered
  - If $p = 20$, there are over one million possibilities!
  - **Branch-and-bound** techniques are sometimes used to reduce some choice. But they also only work for least squares linear regression

Forward Stepwise Selection

Algorithm 6.2 Forward stepwise selection

```
1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors
2. For $k = 0, 1, \ldots, p-1$,
    (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor
    (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest $R^2$
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$, AIC, BIC or adjusted $R^2$
```

- Forward stepwise selection is not guaranteed to find the best model

- Forward stepwise selection can be applied evening the high-dimensional setting where $n < p$; however, in this case, it is possible to construct sub models $\mathcal{M}_0, \ldots, \mathcal{M}_{n-1}$ only

Forward Stepwise Selection

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

Table 6.1: The first four selected models for best subset selection and forward stepwise selection on the Credit data set. The first three models are identical but the fourth models differ.

Backward Stepwise Selection

Algorithm 6.3 Backward stepwise selection

1. Let $\mathcal{M}_p$ denote the *full model*, which contains all p predictors

2. For $k = p, p-1, \ldots, 1$,

    (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 $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p, AIC, BIC or adjusted R^2

- Backward stepwise selection is not guaranteed to find the best model

- Backward stepwise selection requires that the number of samples n is larger than the number of variables p

Hybrid Approaches

- As another alternative, hybrid versions of forward and backward stepwise selection are available
- Variables are added to the model sequentially, in analogy to forward selection
- However, after adding each new variable, the method may also remove any variable stat no longer provide an improvement in the model fit

Choosing the Optimal Model

- We need a way to determine which mode is **best** for the implementation of best subset selection, forward selection, and backward selection
    - RSS and R^2 are not suitable because (1) the model containing all of the predictors will always be selected, and (2) these are related to training error, not test error
- There are two common approaches to select the best model with respect to test error
    - We can indirectly estimate test error by making an **adjustment** to the training error to account for the bias due to overfitting
    - We can **directly** estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in Chapter 5

C_p, AIC, BIC, and Adjusted R^2

- From Chapter 2, the training set MSE is generally an underestimate of the test MSE

- A number of techniques for **adjusting** the training error for the model size are available (say d is the number of predictors in the model): $\mathbf{C_p}$, **Akaike information criterion** (**AIC**), **Bayesian information criterion** (**BIC**), **adjusted** $\mathbf{R^2}$

- C_p estimate of test MSE

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2) \tag{6.2}$$

  - $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement in (6.1)
  - Essentially, the C_p statistic adds a penalty of $2d\hat{\sigma}^2/n$ to the training MSE$(= \text{RSS}/n)$ for adjustment
  - The C_p statistic is known as an unbiased estimate of test MSE
  - We choose the model with the lowest C_p value

C_p, AIC, BIC, and Adjusted R^2

- The AIC criterion is defined for a large class of models fit by maximum likelihood

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2)$$

  - We choose the model with the lowest AIC value
  - For least squares models with Gaussian error, C_p and AIC are proportional to each other, and so the results from both ways are the same

- BIC is derived from a Bayesian point of view

$$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)d\hat{\sigma}^2) \tag{6.3}$$

  - We choose the model with the lowest BIC value
  - BIC replace the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term
    - Since $\log n > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p

C_p, AIC, BIC, and Adjusted R^2

- The adjusted R^2 is another popular approach for selecting among a set of models that contain different numbers of predictors

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)} \tag{6.4}$$

  - The adjusted R^2 is a modified version of $R^2 = 1 - \text{RSS}/\text{TSS}$
  - We choose the model with the largest adjusted R^2 value
  - Maximizing the adjusted R^2 is equivalent to minimizing $\text{RSS}/(n-d-1)$

- See Figure 6.2 for Credit data set
  - C_p selects the model having 6 predictors income, limit, rating, cards, age, student. BIC selects 4 predictors income, limit, cards, student. The adjusted R^2 selects 7 predictors of the same 6 predictors with additional gender

- C_p, AIC, and BIC all have rigorous theoretical justification that are beyond the scope of this lecture
  - The adjusted R^2 is not as well motivated in statistical theory as AIC, BIC, and C_p
  - All of these measures are simple to use and compute
  - AIC, BIC, and C_p can also be defined for more general types of models

C_p, AIC, BIC, and Adjusted R^2

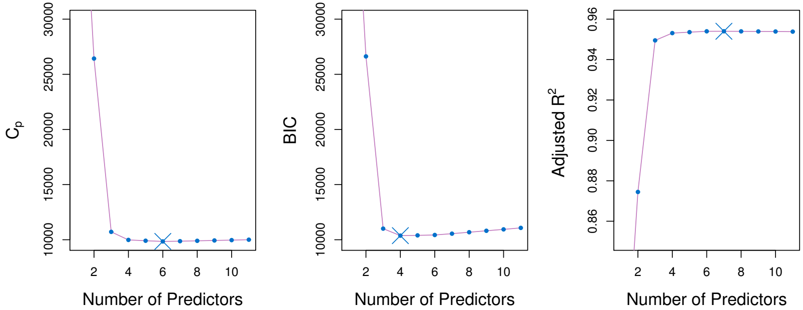

Figure 6.2: C_p, BIC, and adjusted R^2 are shown for the best models of each size for the Credit data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other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

Validation and Cross-Validation

- Advantage of the validation set and cross-validation methods relative to AIC, BIC, C_p, and adjusted R^2
  - It provides a direct estimate of the test error
  - It makes fewer assumptions about the true underlying model
  - It can also be used in a wider range of model selection tasks, even in cases where it is harder to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2
- Disadvantage
  - Performing cross-validation is ("was" in many cases nowadays) computationally burdensome for many problems with large p and/or large n
- See Figure 6.3 for Credit data set
  - Validation set approach: train set for 3/4 observations, validation set for 1/4. 10-fold cross-validation
  - Both approaches select the model of 6 predictors
  - If we repeated the validation set approach using a different split, then the different model may be selected
  - **One-standard-error rule**

Validation and Cross-Validation

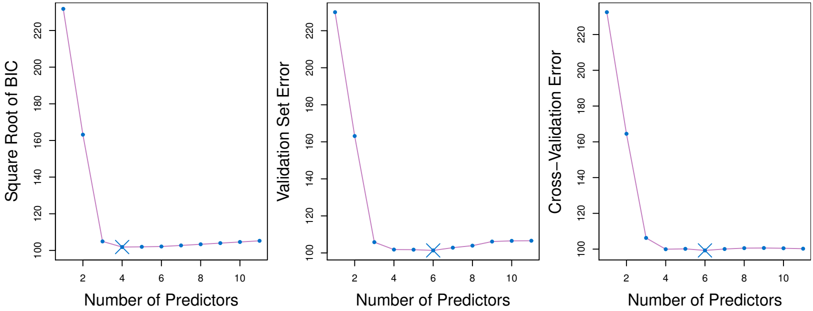

Figure 6.3: For the Credit data set, three quantities are displayed for the best model containing *d* predictors, for *d* ranging form 1 to 11. The overall best model, based on each of these quantities, is shown as a blue cross. Left: Squared root of BIC. Center: Validation set errors. Right: Cross-validation errors.

Shrinkage Methods

- We can fit a model containing all p predictors using a technique that **constrains** or **regularizes** the coefficient estimates
- This approach **shrinks** the coefficient estimates towards zero
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance
- Two best-known techniques for shrinking the regression coefficients towards zero are **ridge regression** and the **lasso**

Ridge Regression

- **Ridge regression** coefficient estimates $\hat{\beta}^R_\lambda$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{2}\beta_j^2 \;=\; \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (6.5)$$

  - Trade-off two criteria
    - Fitting data well by making the RSS small
    - Shrinking the estimates of β_j towards zero due to a **shrinkage penalty** term $\lambda\sum_{j=1}^{p}\beta_j^2$
  - $\lambda \geq 0$ is a **tuning parameter**
    - When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates
    - As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero
    - Selecting a good value for λ is critical; we use cross-validation
  - We do not want to shrink the intercept β_0

An Application to the Credit Data

- See Figure 6.4 for Credit data set
  - Plots are shown as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$
  - The notation $\|\beta\|_2$ denotes the ℓ_2 **norm** (pronounced "ell two") of a factor and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$, which measures the distance of β from zero
  - $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$ is the relative ℓ_2 norm of $\hat{\beta}_\lambda^R$ relative to $\hat{\beta}$

- The standard least squares coefficient estimates discussed in Chapter 3 are **scale equivariant**
  - Multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. $X_j\hat{\beta}_j$ will remain the same
  - The ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant
  - It is best to apply ridge regression after **standardizing the predictors**, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}} \tag{6.6}$$

An Application to the Credit Data

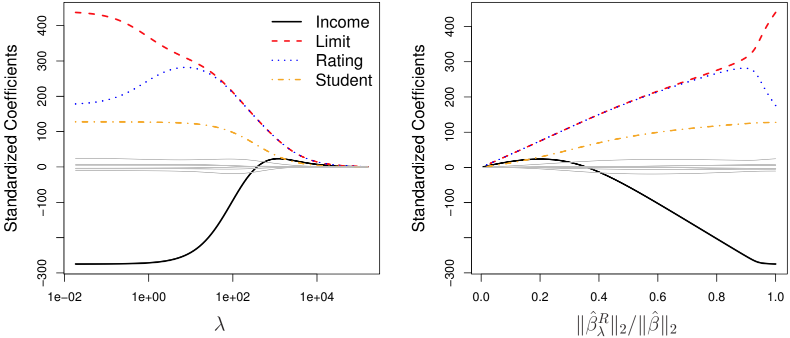

Figure 6.4: The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

Why Does Ridge Regression Improve Over Least Squares?

- Ridge regression's advantage over least squares is rooted in the **bias-variance trade-off**
  - As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias
  - See Figure 6.5 for a simulated data set containing $n = 50$ and $p = 45$
- If $p > n$, the least squares estimates will be extremely variable and do not even have a unique solution
  - The ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance
- Ridge regression also has substantial computational advantages over best subset selection
  - Best subset selection requires searching through 2^p models
  - One can show that the computations required to solve (6.5), *simultaneously for all values of* λ, are almost identical to those for fitting a model using least squares

Why Does Ridge Regression Improve Over Least Squares?

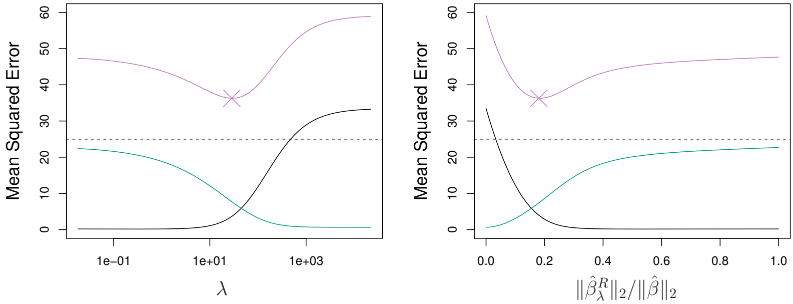

Figure 6.5: Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses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

The Lasso

- Ridge regression will include all p variables in the final model
  - The penalty $\lambda \sum \beta_j^2$ in (6.5) will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - No variable selection is achieved: Increasing the value of λ will tend to reduce the magnitude of the coefficients, but will not result in exclusion of any of the variables
  - It may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in setting in which the number of variables p is quite large

- The **lasso** overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \;=\; \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (6.7)$$

The Lasso

- ℓ_1 penalty in lasso
  - The only difference from the ridge regression is that β_j^2 term in (6.5) has been replaced by $|\beta_j|$ in the lasso penalty (6.7)
  - In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell one") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum |\beta_j|$
  - The ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large
  - Much like best subset selection, the lasso performs **variable selection**
  - We say that the lasso yields **sparse** model–that is, models that involve only a subset of the variables
- See Figure 6.6 for Credit data set

The Lasso

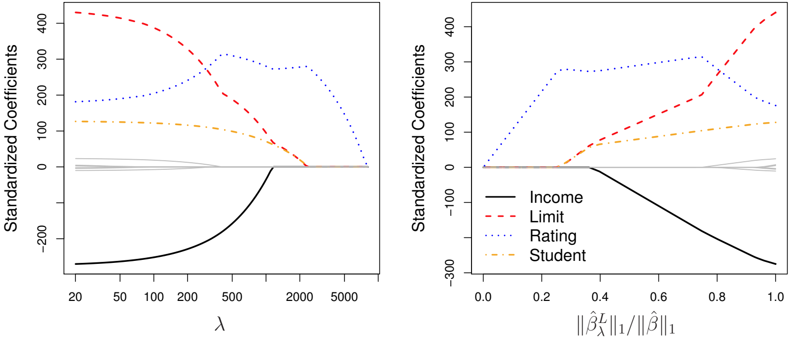

Figure 6.6: The standardized lasso coefficients on the Credit data set are shown as a function of λ and $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Another Formulation for Ridge Regression and Lasso

- Equivalent optimization
  - Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \qquad (6.8)$$

  - Ridge regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s \qquad (6.9)$$

  - Best subset selection

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s \qquad (6.10)$$

- We can think of (6.8) as follow. When we perform the lasso we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to the constraint that there is a **budget** s for how large $\sum_{j=1}^{p} |\beta_j|$ can be. (6.9) and (6.10) has the similar interpretation

The Variable Selection Property of the Lasso

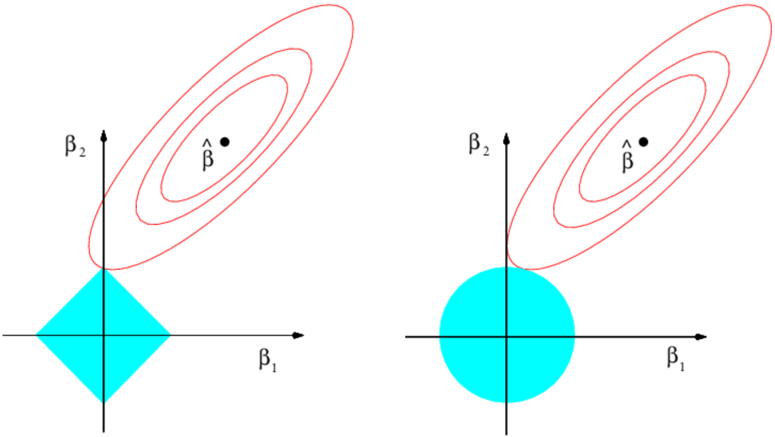

Figure 6.7: Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \le s$ and $\beta_1^2 + \beta_2^2 \le s$, while the red ellipses are the contours of the RSS.

Comparing the Lasso and Ridge Regression

- Neither ridge regression nor the lasso will universally dominate the other
- In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set
- See Figure 6.8 for a large number of true predictors and Figure 6.9 for 2 true predictors

Comparing the Lasso and Ridge Regression

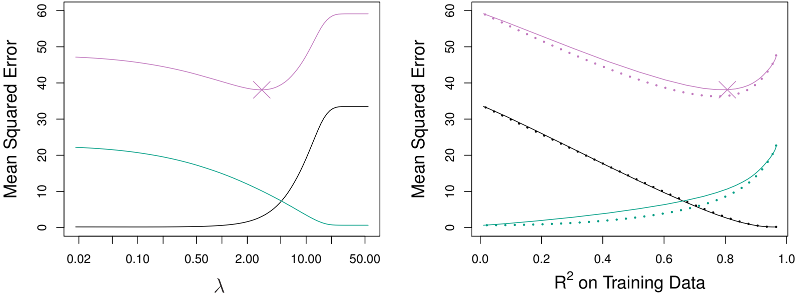

Figure 6.8: Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.

Comparing the Lasso and Ridge Regression

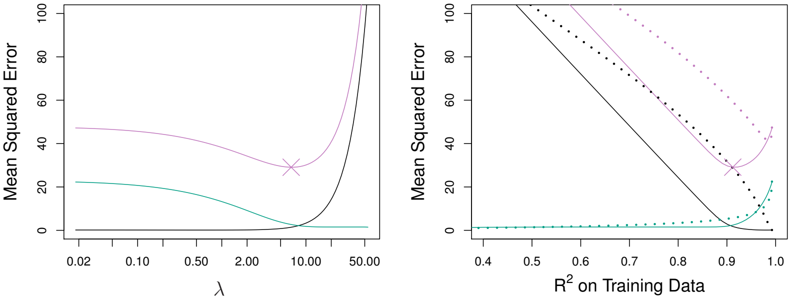

Figure 6.9: Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response. Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.

A Simple Special Case for Ridge Regression and the Lasso

- Consider a simple example with $n = p$ case
  - Suppose $\mathbf{X} = \text{diag}(\mathbf{1}_p) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ or $y_j = x_j + \epsilon_j$ for $j = 1, 2, \ldots, p$.
  - Usual least squares problem

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 \tag{6.11}$$

  - Ridge regression problem

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{6.12}$$

  - Lasso problem

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{6.13}$$

A Simple Special Case for Ridge Regression and the Lasso

- Solutions for a simple example with $n = p$ case (See Figure 6.10)
    - Least square solution
      $$\hat{\beta}_j = y_j$$
    - Ridge solution
      $$\hat{\beta}_j^R = y_j/(1+\lambda) = \hat{\beta}_j/(1+\lambda) \tag{6.14}$$
    - Lasso solution (**soft-thresholding**)
      $$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2 \\ 0 & \text{if } |y_j| \leq \lambda/2 \end{cases} = \begin{cases} \hat{\beta}_j - \lambda/2 & \text{if } \hat{\beta}_j > \lambda/2 \\ \hat{\beta}_j + \lambda/2 & \text{if } \hat{\beta}_j < -\lambda/2 \\ 0 & \text{if } |\hat{\beta}_j| \leq \lambda/2 \end{cases} \tag{6.15}$$

- The fact that some lasso coefficients are shrunken entirely to zero explains why the lasso performs feature selection

- Ridge regression more or less shrinks every dimension of the data by the same proportion, whereas the lasso more or less shrinks all coefficients toward zero by a smaller amount, and sufficiently small coefficients are shrunken all the way to zero

A Simple Special Case for Ridge Regression and the Lasso

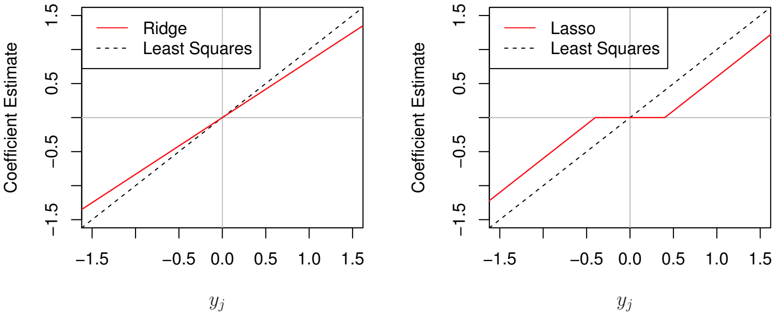

Figure 6.10: The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal. Left: The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates. Right: The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

Bayesian Interpretation for Ridge Regression and the Lasso

- **Posterior distribution**

$$p(\beta|X,Y) \propto f(Y|X,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta)$$

  - Gaussian model: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$ with a Gaussian noise ϵ
  - **Prior distribution**: $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$

- Ridge regression: $g = \mathrm{N}(0, 2/\lambda)$

- Lasso: $g = \mathrm{Laplace}(0, 1/\lambda)$ (or double-exponential distribution)

Bayesian Interpretation for Ridge Regression and the Lasso

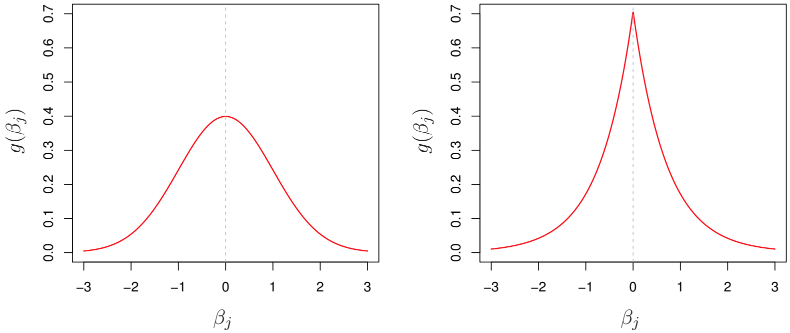

Figure 6.11: Left: Ridge regression is the posterior mode for β under a Gaussian prior. Right: The lasso is the posterior mode for β under a double-exponential prior.

Selecting the Tuning Parameter

- Tuning parameter selection
  - Select the value for the tuning parameter λ in (6.5) and (6.7) giving the best model, or equivalently
  - Select the value for the constraint s in (6.9) and (6.8)
- Use cross-validation
  1. Choose a grid of λ values
  2. Compute the cross-validation error for each value of λ
  3. Select the tuning parameter value for which the cross-validation error is smallest
  4. The model is refit using all of the available observations and the selected value of λ

Selecting the Tuning Parameter

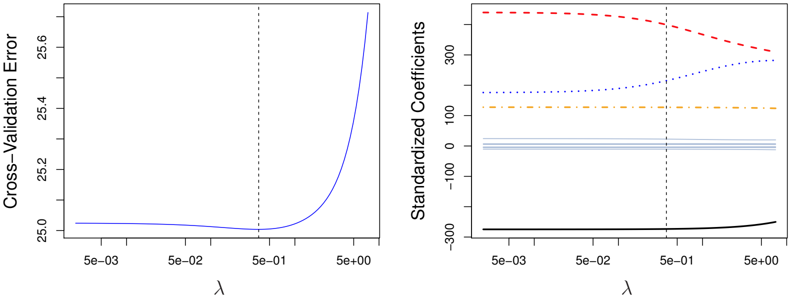

Figure 6.12: Left: Cross-validation errors that result from applying ridge regression to the Credit data set with various value of λ. Right: The coefficient estimates as a function of λ. The vertical dashed lines indicate the value of λ selected by cross-validation.

Selecting the Tuning Parameter

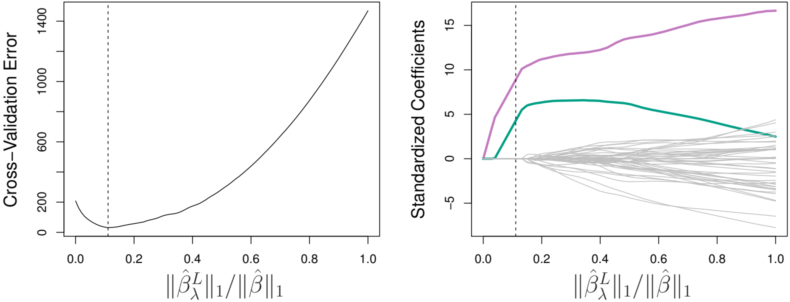

Figure 6.13: Left: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9. Right: The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.

Dimension Reduction Methods

- In **dimension reduction** methods, we explore a class of approaches that *transform* the predictors and then fit a least squares model using the transformed variables

- **Linear combinations** of original p predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \quad m = 1, \ldots, M(< p) \tag{6.16}$$

- Fit the linear model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n \tag{6.17}$$

- If the constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ are chosen wisely, then such dimension reduction approaches can often outperform least squares regression

Dimension Reduction Methods

- Dimension reduction for the problem
  - (6.1) requires estimating $p + 1$ coefficients $\beta_0, \beta_1, \dots, \beta_p$
  - (6.17) requires estimating $M + 1$ coefficients $\theta_0, \theta_1, \dots, \theta_M$
  - If $M < p$, the dimension of the problem is reduced from $p + 1$ to $M + 1$

- Dimension reduction serves to constrain the estimates
  - Notice that, from (6.16),

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \left(\sum_{m=1}^{M} \theta_m \phi_{jm} \right) x_{ij}$$

  - Estimating θ is equivalent to estimating β with the constraints

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm} \qquad (6.18)$$

    given ϕ's
  - If $M = p$, and all the Z_m are linearly independent, then (6.18) poses no constraints. In this case, no dimension reduction occurs, and so fitting (6.17) is equivalent to performing least squares on the original p predictors

Dimension Reduction Methods

- This constraint on the form of the coefficients has the potential to bias the coefficient estimates
- In situations where p is large relative to n, selecting a value of $M \ll p$ can significantly reduce the variance of the fitted coefficients
- All dimension reduction methods work in two steps
  1. Obtain the transformed predictors $Z_1, Z_2, \ldots, Z_M$
  2. Fit the model with these M predictors
- How to choose $Z_1, Z_2, \ldots, Z_M$, or equivalently, to select the ϕ_{jm}'s?
  - Principal components regression: use principal components
  - Partial least squares

Principal Components Regression

- **Principal component analysis** (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables
- PCA is one of important tools for **unsupervised learning** in Chapter 10
- Here we describe its use as a dimension reduction technique for regression

An Overview of Principal Components Analysis

- The **first principal component** direction
  - The first principal component direction is that along which the observations *vary the most*
  - The first principal component direction defines the line that is *as close as possible* to the data
  - If we **project** all observations onto this direction, then the resulting projected observations (**first principal component score**) would have the largest possible variance
  - The projected observations are *as close as possible* to the original data

- See Figure 6.14
  - The first principal component score is

$$Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}}) \qquad (6.19)$$

  - $\phi_{11} = 0.839$ and $\phi_{21} = 0.544$ are the principal component loadings, defining the first principal component direction
  - Such ϕ_{11} and ϕ_{21} maximizes $\text{Var}(\phi_{11} \times (\text{pop} - \overline{\text{pop}}) + \phi_{21} \times (\text{ad} - \overline{\text{ad}}))$ under the constraint $\phi_{11}^2 + \phi_{21}^2 = 1$
  - The first principal component score for the ith observation is

$$z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}}) \qquad (6.20)$$

An Overview of Principal Components Analysis

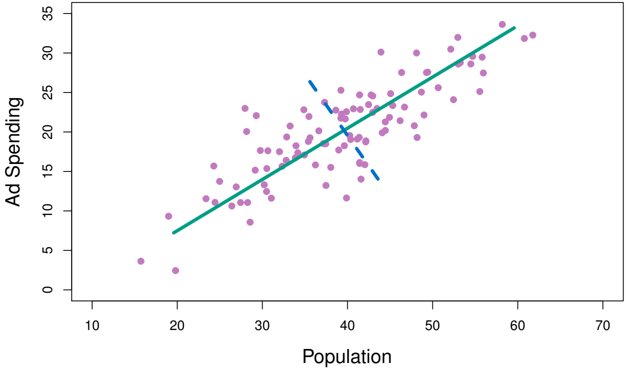

Figure 6.14: The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.

An Overview of Principal Components Analysis

- The second principal component direction
  - It is a linear combination of the variables that is uncorrelated with Z_1, and has largest variance subject to this constraint
  - For Figure 6.14,

$$Z_2 = 0.544 \times (\text{pop} - \overline{\text{pop}}) - 0.839 \times (\text{ad} - \overline{\text{ad}})$$

  - The second principal component score for the ith observation is

$$z_{i2} = 0.544 \times (\text{pop}_i - \overline{\text{pop}}) - 0.839 \times (\text{ad}_i - \overline{\text{ad}})$$

An Overview of Principal Components Analysis

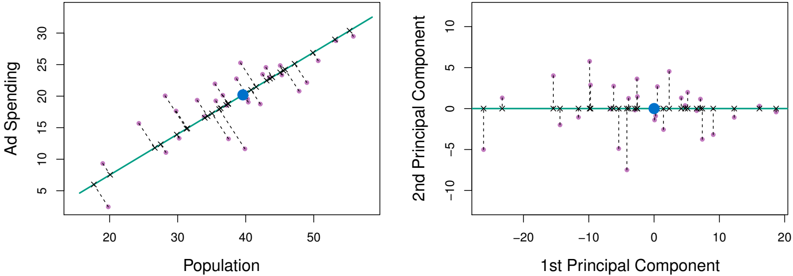

Figure 6.15: A subset of the advertising data. The mean pop and ad budgets are indicated with a blue circle. Left: The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all *n* of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents ($\overline{\text{pop}}$, $\overline{\text{ad}}$). Right: The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.

An Overview of Principal Components Analysis

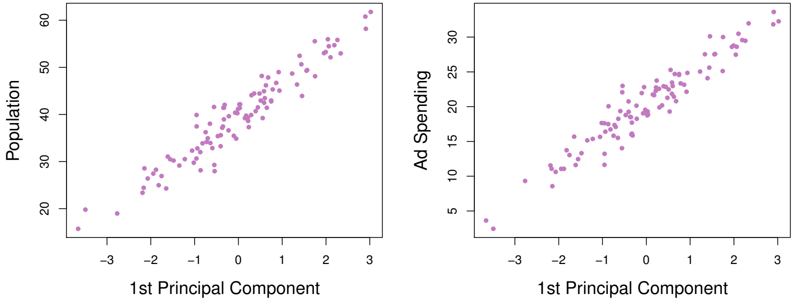

Figure 6.16: Plots of the first principal component scores z_{i1} versus pop and ad. The relationships are strong.

An Overview of Principal Components Analysis

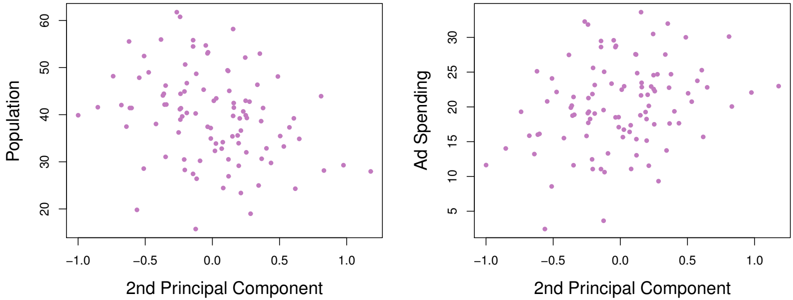

Figure 6.17: Plots of the second principal component scores z_{i2} versus pop and ad. The relationships are weak.

The Principal Components Regression Approach

- The **principal components regression** (PCR) approach fit a linear regression model with $Z_1, \ldots, Z_M$ as predictors using least squares
  - The key idea is that a small number of principal components suffice to explain most of the variability in the data as well as the relationship with the response
  - This assumes that *the directions in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with Y*
  - While this assumption is not guaranteed to be true, it often turns out to be a reasonable enough approximation to give good results
  - In the case where this assumption is true, estimating $M \ll p$ coefficients can mitigate overfitting
- Figure 6.18 does not favor PCR, and Figure 6.19 favors PCR

The Principal Components Regression Approach

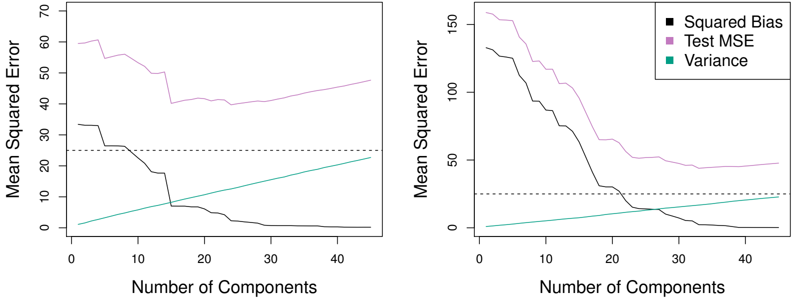

Figure 6.18: PCR was applied to two simulated data sets. Left: Simulated data from Figure 6.8. Right Simulated data from Figure 6.9.

The Principal Components Regression Approach

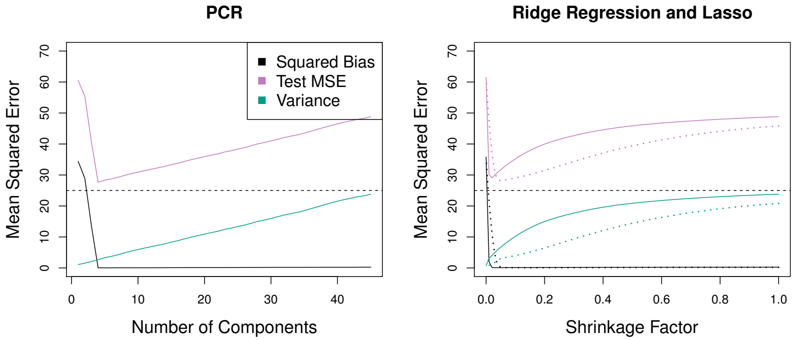

Figure 6.19: PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y. In each panel, the irreducible error Var(ϵ) is shown as a horizontal dashed line. Left: Results for PCR. Right: Results for lasso (solid) and ridge regression (dotted). The x-axis displays the shrinkage factor of the coefficients estimates, defined as the ℓ_2 norm of the shrunken coefficient estimates divided by the ℓ_2 norm of the least squares estimate.

The Principal Components Regression Approach

- PCR does *not* provide a feature selection
  - Note that each of M principal components is a linear combination of all p original features
  - In this sense, PCR is more closely related to ridge regression than to the lasso
  - In fact, ridge regression can be interpreted as a continuous version of PCR (See Section 3.5 of ESL for more detail)
- M is typically chosen by cross-validation

The Principal Components Regression Approach

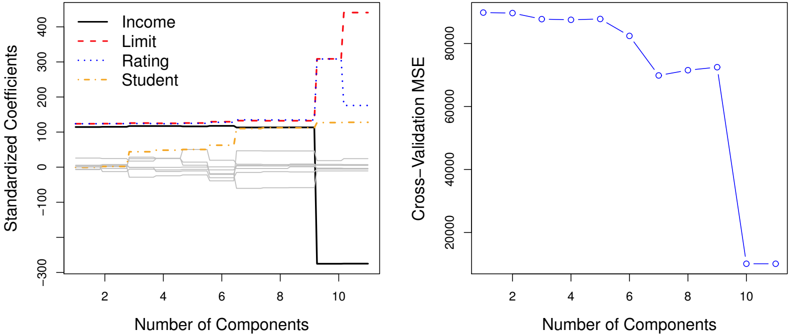

Figure 6.20: Left: PCR standardized coefficient estimates on the Credit data set for different values of M. Right: The ten-fold cross validation MSE obtained using PCR, as a function of M.

Partial Least Squares

- The PCR approach is unsupervised way
  - Linear combinations best represent the predictors $X_1, \ldots, X_p$
  - The response Y is not used to help determine the principal component directions
  - There is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response
- **Partial least squares** (PLS) is a supervised alternative to PCR
  - The PLS approach attempts to find directions that help explain both the response and the predictors
  - PLS procedure
    1. Standardize the p predictors
    2. Compute the first direction Z_1 by setting ϕ_{j1} in (6.16) equal to the coefficient from the simple linear regression of Y onto X_j; $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$
    3. Compute residuals $Y - Z_1$ and replace Y by this. Then find Z_2 by the same way to find Z_1
    4. Repeat the above to find $Z_1, Z_2, \ldots, Z_M$

Partial Least Squares

- PLS places the highest weight on the variables that are most strongly related to the response
- Residuals can be interpreted as the remaining information that have not been explained by the previous PLS directions
- PLS is popular in chemometrics, especially in the analysis of digitized spectrometry signals
- In practice PLS often performs no better than ridge regression or PCR

Partial Least Squares

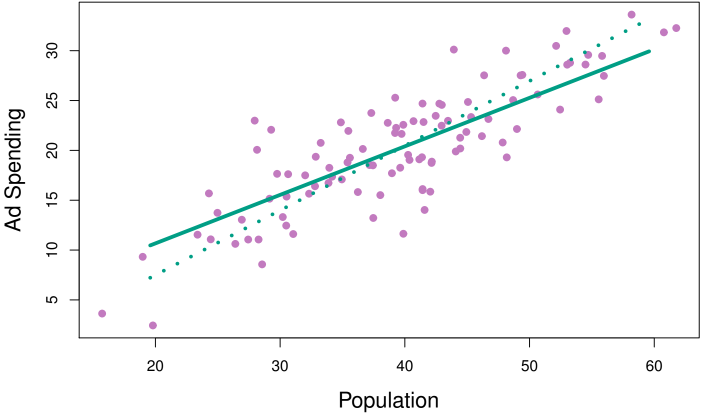

Figure 6.21: For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.

High-Dimensional Data

- Most traditional statistical techniques for regression and classification are intended for the *low-dimensional* setting, i.e., $n \gg p$
- Due to new technology development, it is commonplace to collect an almost unlimited number of feature measurements (p is very large)
- Examples
  - It is possible to collect measurements for half a million **single nucleotide polymorphisms** (SNPs; these are individual DNA mutation that are relatively common in the population) for inclusion in the predictive model. Then n is hundreds or thousands but $p \approx 500,000$
  - A marketing analyst interested in understanding people's online shopping patterns could treat as features all of the search terms entered by users of a search engine. This is sometimes known as the "bag-of-words" model. The same researcher might have access to the search histories of only a few hundred or a few thousands search engine users who have consented to share their information with the researcher. For a given user, each of the p search terms is scored present (0) or absent (1), creating a large binary feature vector. Then $n \approx 1,000$ and p is much larger

What Goes Wrong in High Dimensions?

- When $p > n$, least squares in Chapter 3 cannot (or rather *should not*) be performed
  - Regardless of whether or not there truly is a relationship between the features and the response, least squares will yield a set of coefficient estimates that result in a perfect fit to the data, such that the residuals are zero (See Figure 6.22)
  - *Even though the features are completely unrelated to the response*, the training set MSE decreases to 0 as p increases. On the other hand, the MSE on an *independent test set* becomes extremely large as p increases (See Figure 6.23)
  - For model selection, the C_p, AIC, and BIC approaches are not appropriate in the high-dimensional setting, because estimating σ^2 is problematic

What Goes Wrong in High Dimensions?

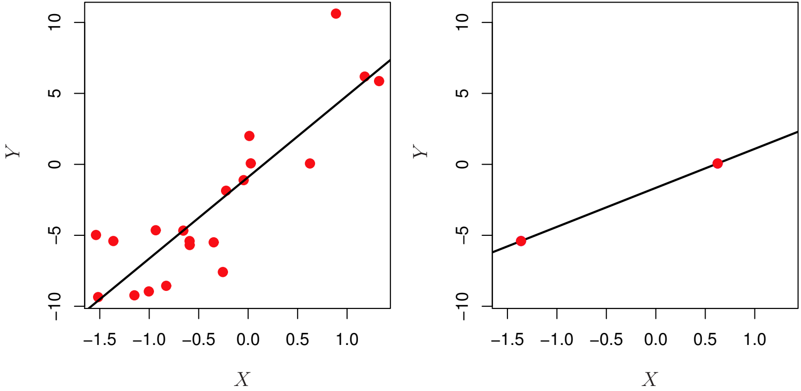

Figure 6.22: Left: Least squares regression in the low-dimensional setting. Right: Least squares regression with $n = 2$ observations and two parameters to be estimated (an intercept and a coefficient.

What Goes Wrong in High Dimensions?

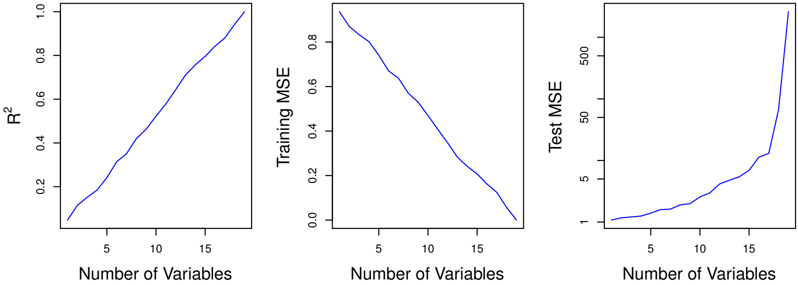

Figure 6.23: On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model. Left: The R^2 increases to 1 as more features are included. Center: The training set MSE decreases to 0 as more features are included. Right: The test set MSE increases as more features are included.

Regression in High Dimensions

- Fitting *less flexible* least squares models is useful in high-dimensional problem
  - Forward stepwise selection, ridge regression, the lasso, and principle components regression
  - These approaches avoid overfitting by using a less flexible approach than least squares
- See Figure 6.24
- **Curse of dimensionality**
  - The performance of statistical methods deteriorates as the dimension grows
  - In general, adding additional signal features that are truly associated with the response will improve the fitted model
  - Adding noise features that are not truly associated with the response will lead to a deterioration in the fitted model
  - Abundant information is a double-edged sword: they can lead to improved predictive model if these features are relevant, but will lead to worse results if the features are not relevant

Regression in High Dimensions

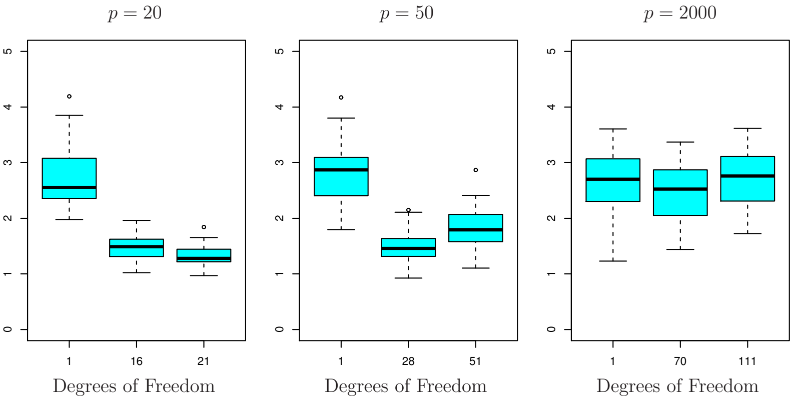

Figure 6.24: The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The box plots show the test MSEs that result using three different values of the tuning parameter λ in (6.7). For ease of interpretation, rather than reporting λ, the degrees of freedom are reported; for the lasso this turns out to be simply the number of estimated non-zero coefficients. When $p = 20$, the lowest test MSE was obtained with the smallest amount of regularization. When $p = 50$ the lowest test MSE was achieved when there is a substantial amount of regularization. When $p = 2,000$ the lasso performed poorly regardless of the amount of regularization, due to the fact that only 20 of the 2,000 features truly are associated with the outcome.

Interpreting Results in High Dimensions

- In the high-dimensional setting, the multicollinearity problem is extreme
    - Any variable in the model can be written as a linear combination of all of the other variables in the model
    - This means that we can never know exactly which variables truly are predictive of the outcome, and we can never identify the best coefficients for use in the regression
    - There can be many best models. We must be careful not to overstate the results obtained, and to make it clear that what we have identified is simply *one of many possible models* for prediction
- Do not use traditional measures of model fit on the training data as evidence of a good model fit. Validation must be based on the test data

Lab1: Subset Selection Methods

Lab2: Ridge Regression and the Lasso

Lab3: PCR and PLS Regression