Statistical Machine Learning and Its Applications

Lecture 5: Linear Model Selection and Regularization

KAIST Mark Mintae Kim

Department of Industrial & Systems Engineering
KAIST

OUTLINE

- Overview
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
- Considerations in High Dimensions

OUTLINE

- Overview
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
- Considerations in High Dimensions

OVERVIEW

- Recall the linear model

$$y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

- Despite its simplicity, the linear model has distinct advantages in terms of its **interpretability** and often shows good **predictive performance**.
- In this lecture, we discuss some ways in which the simple linear model can be improved, by **replacing ordinary least squares fitting** with some alternative fitting procedures.

WHY CONSIDER ALTERNATIVES TO LEAST SQUARES?

- **Prediction Accuracy**
  - If the true relationship between the response and the predictors is approximately linear, the least squares estimates will have **low bias**
  - What about **variance**?
    - $n \gg p$: Low variance → No problem
    - $n \approx p$: Variance gets higher
    - $p > n$: Variance is infinite (no unique least squares coefficient estimates)

    → **How can we reduce variance and improve accuracy?**
  - By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias
- **Model interpretability**
  - Often, many of the variables used in a multiple linear regression model are in fact not associated with the response.
    - Including such irrelevant variables leads to unnecessary complexity in the resulting model.
    - How can we remove these irrelevant variables?
  - By performing automatic **feature selection**

THREE CLASSES OF METHODS

- **Subset Selection**
  - We identify a subset of the p predictors (= variables) that we believe to be related to the response.
  - We then fit a model using least squares on the reduced set of predictors.

- **Shrinkage**
  - We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero.
  - This shrinkage (a.k.a. regularization) has the effect of reducing variance and can also perform variable selection.

- **Dimension Reduction**
  - We project the p predictors into a M-dimensional subspace, where $M < p$.
  - This is achieved by computing M different linear combinations, or projections, of the predictors.
  - Then these M projections are used as predictors to fit a linear regression model by least squares.

OUTLINE

- Overview
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
- Considerations in High Dimensions

SUBSET SELECTION: BEST SUBSET SELECTION

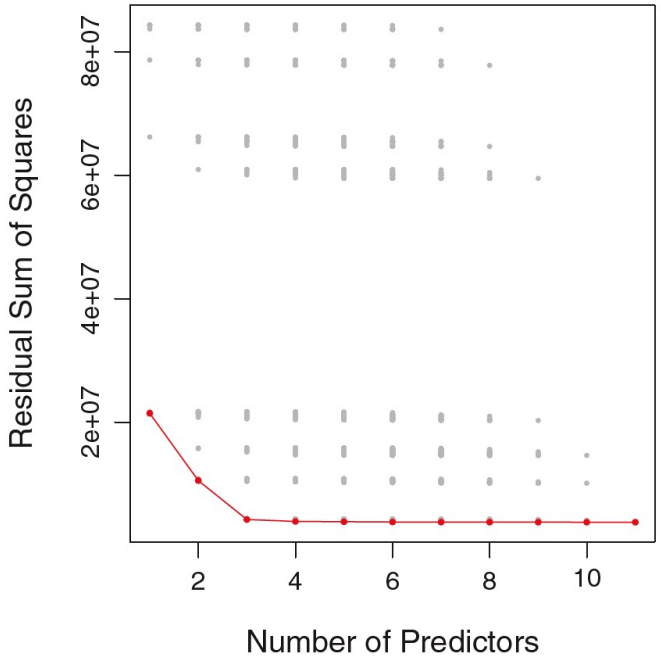
- **Which variables to keep and which to drop?**
- Idea
  - Exhaust all possible combinations of inputs
  - With p variables, there are 2^p many distinct combinations.
  - Identify the best model among these models.

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p$$

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.

EXAMPLE: CREDIT DATA SET



- Residual sum of squares $(RSS) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- **Problem**: RSS always decreases as the number of features included in the models increases
- Which model should we select?
  - We should focus "test error" rather than "training error"
- Thus, in step 3, **cross-validation** should be used
  - More on this later (+ C_p, AIC, BIC, and Adjusted R^2)

COMMENTS ON BEST SUBSET SELECTION

- **Advantage**: Simple, easy to implement and conceptually clear.

- **Disadvantage:** Computationally expensive
  - With p variables, there are 2^p many distinct combinations

- Solution: Stepwise selection methods
  - Forward stepwise selection
  - Backward stepwise selection

SUBSET SELECTION: FORWARD STEPWISE SELECTION

- Begin with a model containing no predictors
- Then, add predictors to the model, one-at-a-time, until all of the predictors are in the model.
- At each step the variable that gives the greatest additional improvement to the fit is added to the model.

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (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.

COMMENTS ON FORWARD STEPWISE SELECTION

- **Advantage**:
  - 1. Computationally efficient
    - Only need to consider $1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$
    - When $p = 20$
      - Best subset selection: $2^{20} = 1{,}048{,}576$ models
      - Forward Stepwise selection: $1 + \frac{20(20+1)}{2} = 211$ models
  - 2. Can work for even in the high-dimensional setting i.e., $n < p$
    - But, it is possible to only construct submodels $M_0, \ldots, M_{n-1}$, since each submodel is fit using least squares, which will not yield a unique solution if $n \leq p$

- **Disadvantage**: Not guaranteed to find the best subset (Once an input is in, it does not get out)

# 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

SUBSET SELECTION: BACKWARD STEPWISE SELECTION

- Like forward stepwise selection, **backward stepwise selection** provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then **iteratively removes the least useful predictor**, one-at-a-time.

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.

COMMENTS ON BACKWARD STEPWISE SELECTION

- **Advantage**: Computational efficiency
  - Same as forward stepwise selection

- **Disadvantage**
  - 1) Not guaranteed to find the best subset (like forward stepwise selection)
    - Once an input is out, it does not get in.
  - 2) Cannot work for $n < p$
    - We start with the model with all p predictors

CHOOSING THE OPTIMAL MODEL

- Best subset, Forward stepwise, Backward stepwise all generate multiple models
- **Recall**: Training error will always decrease as the number of features included in the models increases
- Our goal is choose a model with low **test error**, not a model with low training error.
  - Recall that training error is usually a poor estimate of test error.
- Two approaches
  - 1. **Directly** estimating the test error
    - Validation/cross-validation approach (Learned about this in Lecture 4)
  - 2. **Indirectly** estimating the test error by making an adjustment to the training error
    - Idea: Account for the bias due to overfitting
    - Adjusted R^2, AIC, BIC or C_p

CHOOSING THE OPTIMAL MODEL

- Best subset, Forward stepwise, Backward stepwise all generate multiple models (based on training RSS)
- **Recall**: Training RSS will always decrease as the number of features included in the models increases
- Our goal is choose a model with low **test error**, not a model with low training error.
  - Recall that training error is usually a poor estimate of test error.

- Two approaches
  - 1. **Directly** estimating the test error
    - Validation/cross-validation approach (Learned about this in Lecture 4)
  - 2. **Indirectly** estimating the test error by making an adjustment to the training error
    - Idea: Account for the bias due to overfitting
    - Adjusted R^2, AIC, BIC or C_p

ADJUSTED R^2

- **Intuition**: Once all of the useful variables have been included in the model, adding additional *noise* variables will lead to only a small decrease in RSS
- **Idea**: Pays a price for the inclusion of unnecessary variables in the model by dividing RSS by $(n - d - 1)$

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

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- The larger the better: Maximizing [Adjusted R^2] = Minimizing $[RSS/(n - d - 1)]$
- Adding noise variables → Increase in d → Increase $[RSS/(n - d - 1)]$ → Decrease in adjusted R^2

AIC, BIC, AND C_p

- Some other ways of penalizing RSS
  - The smaller the better
  - The penalty increases as the number of predictors d in the model increases

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

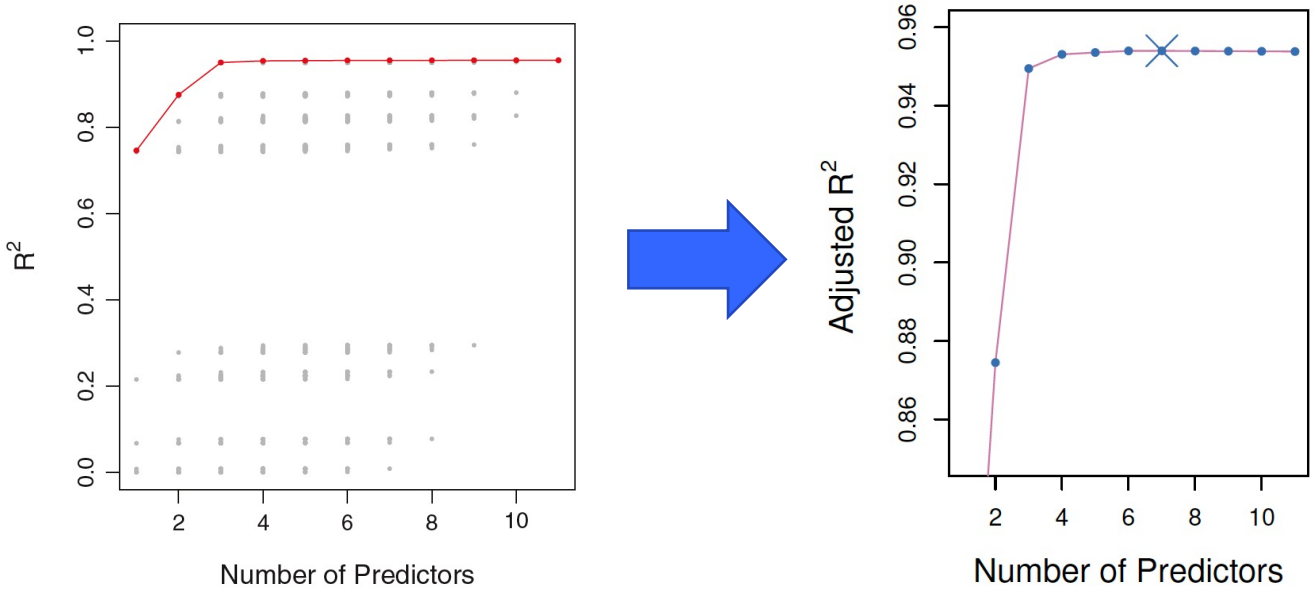
$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

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

(More severe penalty for large models)

  - $\hat{\sigma}^2$: An estimate of the variance of the error ϵ associated with each response measurement
  - BIC results in the selection of smaller models than C_p or AIC

EXAMPLE: CREDIT DATA



Using all the predictors is not the best anymore

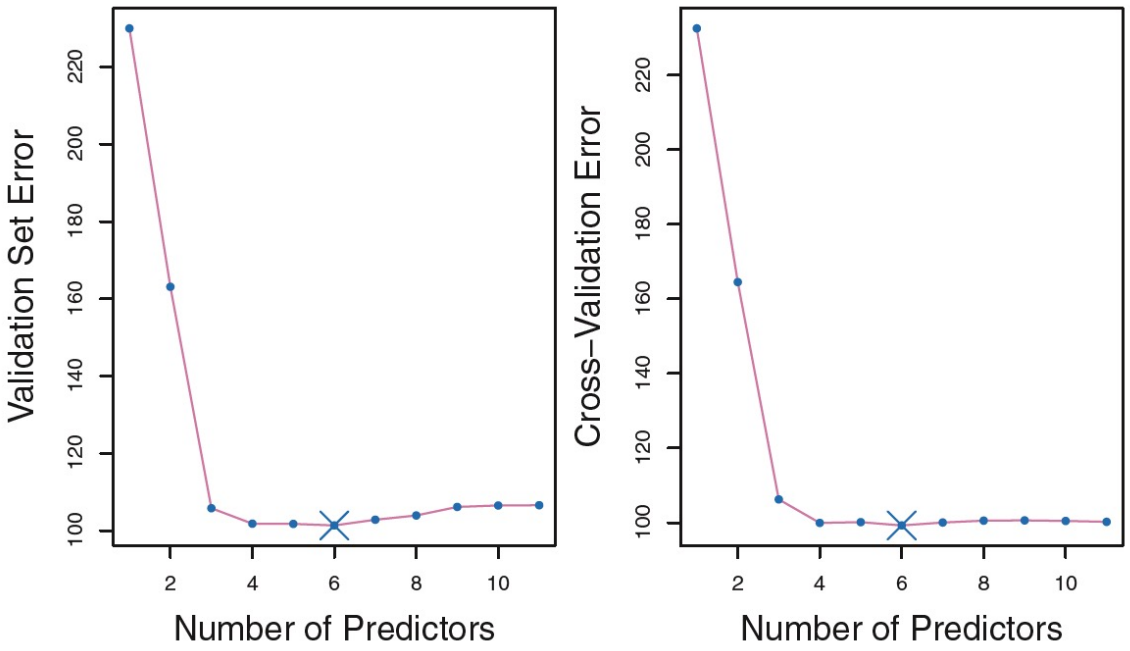
CHOOSING THE OPTIMAL MODEL

- Best subset, Forward stepwise, Backward stepwise all generate multiple models (based on training RSS)
- **Recall**: Training RSS will always decrease as the number of features included in the models increases
- Our goal is choose a model with low **test error**, not a model with low training error.
  - Recall that training error is usually a poor estimate of test error.

- Two approaches
  - 1. **Directly** estimating the test error
    - Validation/cross-validation approach (Learned about this in Lecture 4)
  - 2. **Indirectly** estimating the test error by making an adjustment to the training error
    - Idea: Account for the bias due to overfitting
    - Adjusted R^2, AIC, BIC or C_p

VALIDATION AND CROSS-VALIDATION

- Directly estimate the test error
- Each of the procedures returns a sequence of models M_k indexed by model size $k = 0,1,2, \dots p$.
- Our goal is to select $\hat{k}$, and return model $M_{\hat{k}}$.
- We compute the validation set error or the cross-validation error for each model M_k under consideration, and then select the k for which the resulting estimated test error is smallest.



- Models with 4,5,6 predictors are roughly equivalent
- Keep in mind the **Occam's razor**: Choose the simplest model if they are similar by other criterion.

OUTLINE

- Overview
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
- Considerations in High Dimensions

SHRINKAGE METHODS

- The subset selection methods use least squares to fit a linear model that contains a *subset* of the predictors
- Shrinkage methods use all p predictors, but *constrain* or *regularize* the coefficient estimates, or equivalently shrink the coefficient estimates to zero
- It turns out that shrinking estimated coefficients towards zero can significantly reduce their variance
- Two best-known techniques are **ridge regression** and **lasso**

- **Overview**
  - Linear regression

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$$

  - Ridge regression

L2 norm

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^{\mathbf{2}}), \text{ where } \|\boldsymbol{\beta}\|^{\mathbf{2}} = \sum_{j=1}^{p} \beta_j^2$$

  - Lasso

L1 norm

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda|\boldsymbol{\beta}|), \text{ where } |\boldsymbol{\beta}| = \sum_{j=1}^{p} |\beta_j|$$

RIDGE REGRESSION

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

- In contrast, the *ridge regression* coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2} \leftarrow \textbf{Shrinkage penalty}$$

  - The **first term** measures goodness of fit, the smaller the better.
  - The **second term** is called shrinkage penalty, which shrinks β_j towards 0.
    - $\lambda \geq 0$ is a tuning parameter (or hyperparameter) that controls the model complexity
      - **If λ is large**, then we want more parameters to be close to 0 → less flexibility → bias increase, variance decrease
      - **If λ is small**, then ridge gets similar to OLS → more flexibility → bias decrease, variance increase
      - Need to determine separately (Cross-validation is used)

EXAMPLE: CREDIT DATA

Ordinary least square
$(\lambda = 0)$

Ordinary least square
$(\lambda = 0)$

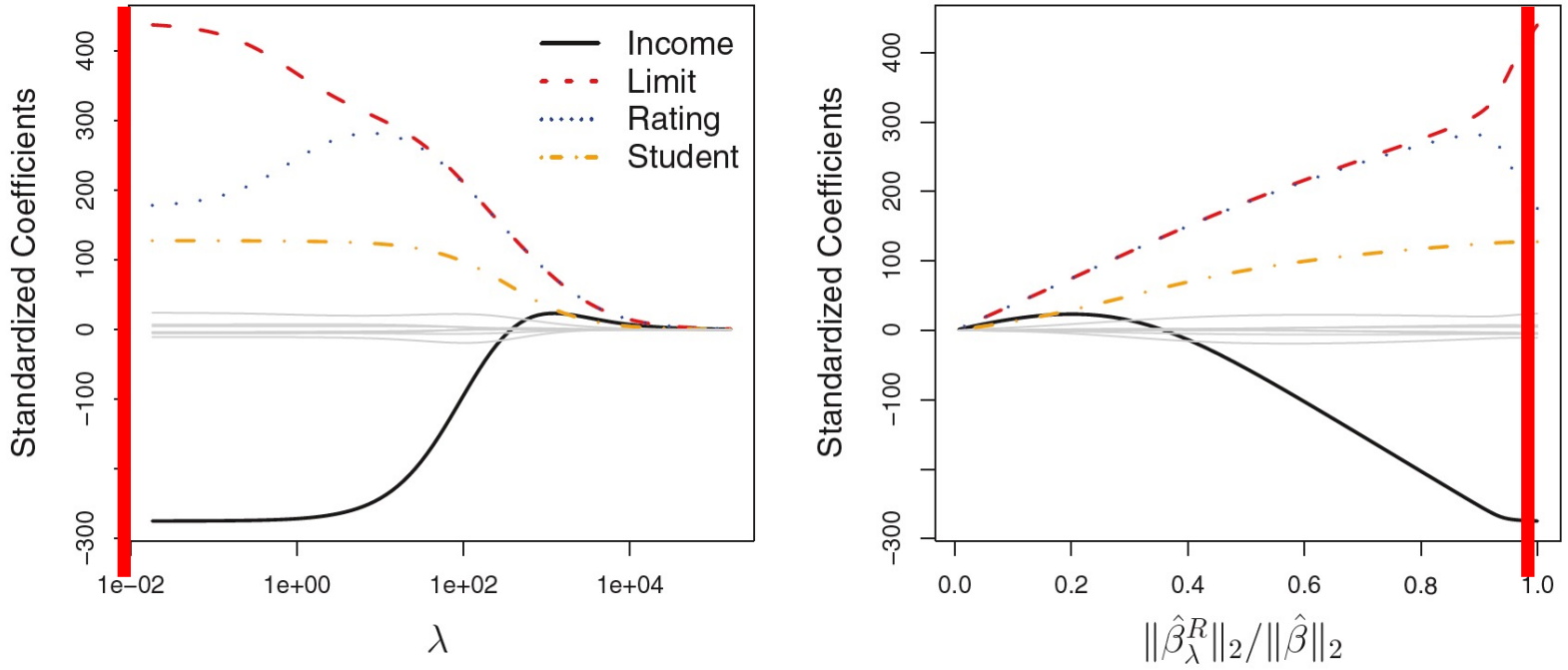


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

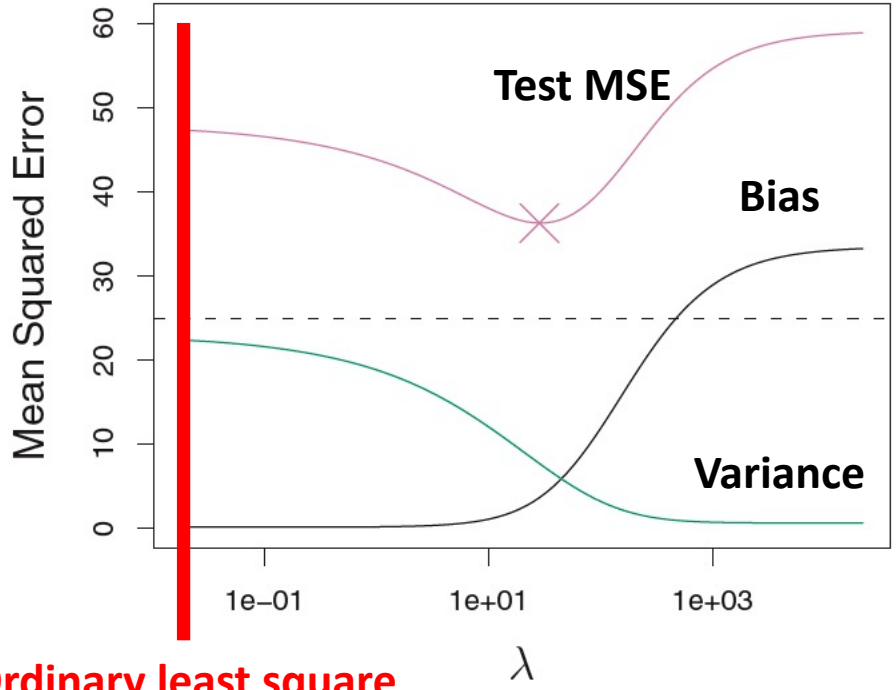
RIDGE REGRESSION: SCALING OF PREDICTORS

- The **standard least squares** coefficient estimates are *scale equivariant*
- Multiplying X_{ij} by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $\frac{1}{c}$.
  - $\beta_j X_{ij} = (cX_{ij})(\frac{\beta_j}{c})$
- The **ridge regression** coefficient estimates can change substantially when multiplying a given predictor by a constant
  - Due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after *standardizing the predictors* by scaling the predictors by their standard deviations

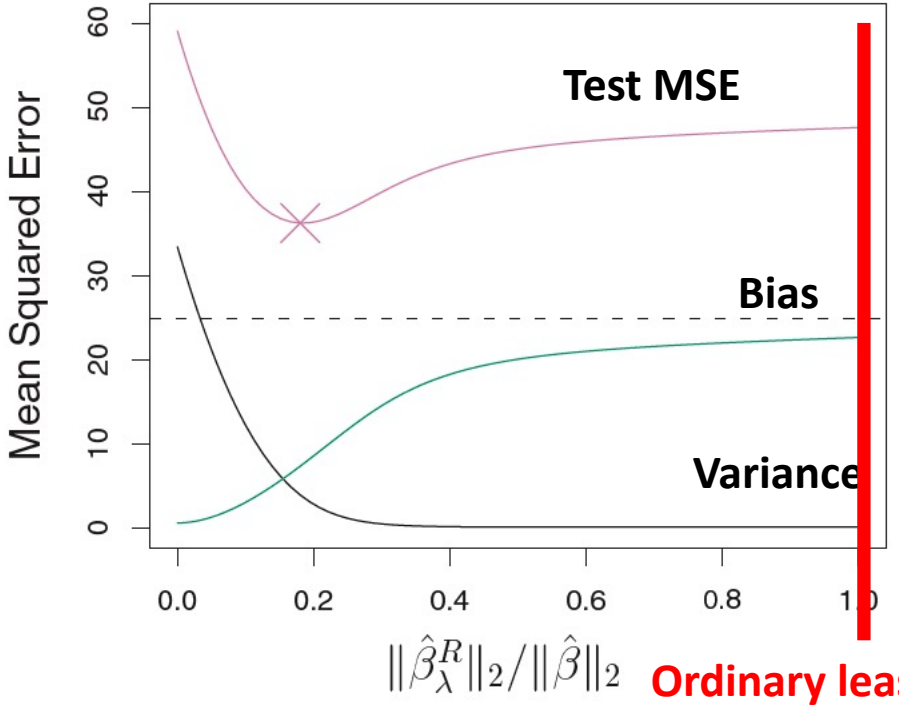
$$\tilde{X}_{ij} = \frac{X_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)}} \quad \text{, where } \bar{X}_j = \sum_{i=1}^{n} X_{ij}$$

WHY DOES RIDGE REGRESSION IMPROVE OVER LEAST SQUARES?

Test MSE $= (\boldsymbol{bias})^2 + \boldsymbol{variance}$



Ordinary least square
$(\boldsymbol{\lambda = 0})$



Ordinary least square
$(\boldsymbol{\lambda = 0})$

Variance is high, but no bias

CLOSED-FORM SOLUTION FOR THE RIDGE REGRESSION

- Recall the closed-form solution for the linear regression model

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \parallel \boldsymbol{y} - \boldsymbol{X\beta} \parallel^2}{\partial \boldsymbol{\beta}} = \mathbf{0} = -2\boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta})$$

$$\Longrightarrow \quad \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta}) = \mathbf{0}$$

$$\Longrightarrow \quad \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X\beta} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

$$\Longrightarrow \quad \widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

- Loss function of ridge regression

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

$$= \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial(\parallel \boldsymbol{y} - \boldsymbol{X\beta} \parallel^2 + \lambda\|\boldsymbol{\beta}\|^2)}{\partial \boldsymbol{\beta}} = \mathbf{0} = -2\boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta}) + \mathbf{2}\lambda\boldsymbol{\beta}$$

$$\Longrightarrow \quad \boldsymbol{X}^{\boldsymbol{T}} (\boldsymbol{y} - \boldsymbol{X\beta}) = \lambda\boldsymbol{\beta}$$

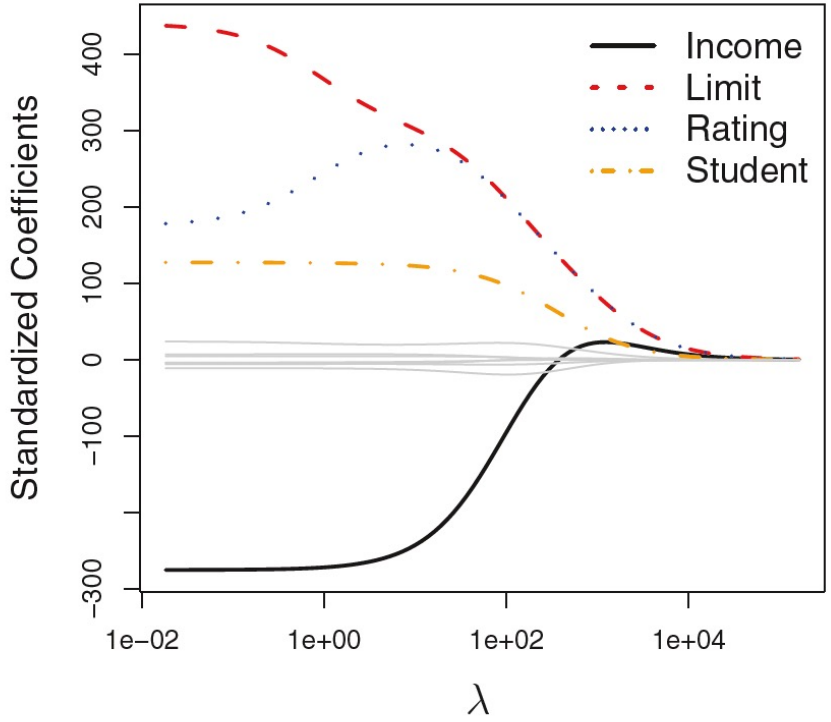
$$\Longrightarrow \quad \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X\beta} = \lambda\boldsymbol{\beta}$$

Always invertible

$$\Longrightarrow \quad \widehat{\boldsymbol{\beta}} = \boxed{\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)}^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

COMMENTS ON RIDGE REGRESSION

- Works best when least squares estimates have high variance
- Computationally efficient (Best subset selection (2^p models))
- If $p > n$,
  - Ridge regression can still perform well
    - By trading off a small increase in bias for a large decrease in variance
  - Ordinary least square estimates do not have a unique solution
- Lacks interpretability
  - Doesn't actually perform variable selection
  - Final model will include all predictors
    - If all we care about is prediction accuracy, this isn't a problem
    - It does, however, pose a challenge for model interpretation
- If we want a technique that actually performs variable selection, what needs to change?



THE LASSO

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.
- The main idea is the same as the ridge regression
  - Minimize RSS plus an additional penalty that rewards small (sum of) coefficient values
- The Lasso overcomes the disadvantage of the ridge regression.
  - No variable selection in the ridge regression (Lacks interpretability)
- *The Lasso* coefficient estimates $\hat{\beta}^L$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \boxed{\lambda \sum_{j=1}^{p}|\beta_j|} \leftarrow \textbf{Shrinkage penalty}$$

- L2 norm in the ridge regression → L1 norm in the Lasso
- As with ridge, lasso also shrinks coefficients towards zero.
- Then, what is the difference?

THE LASSO

- However, unlike the L2 penalty, L1 has the effect of forcing some coefficients to be exactly equal to zero
- Hence, much like best subset selection, the Lasso performs **variable selection**
- We say that the lasso yields **sparse** models, i.e., models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for the Lasso is critical
  - Cross-validation is again the method of choice
- **Why does L1 penalty yield sparse models?**

THE VARIABLE SELECTION PROPERTY OF THE LASSO

- **Reformulation**: The Lasso and ridge regression coefficient estimates solve the following problems

- For each value of λ, there exists a value for s such that
  - Ridge regression

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

  - Lasso

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

THE LASSO PICTURE: COMPARING CONSTRAINT FUNCTIONS

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

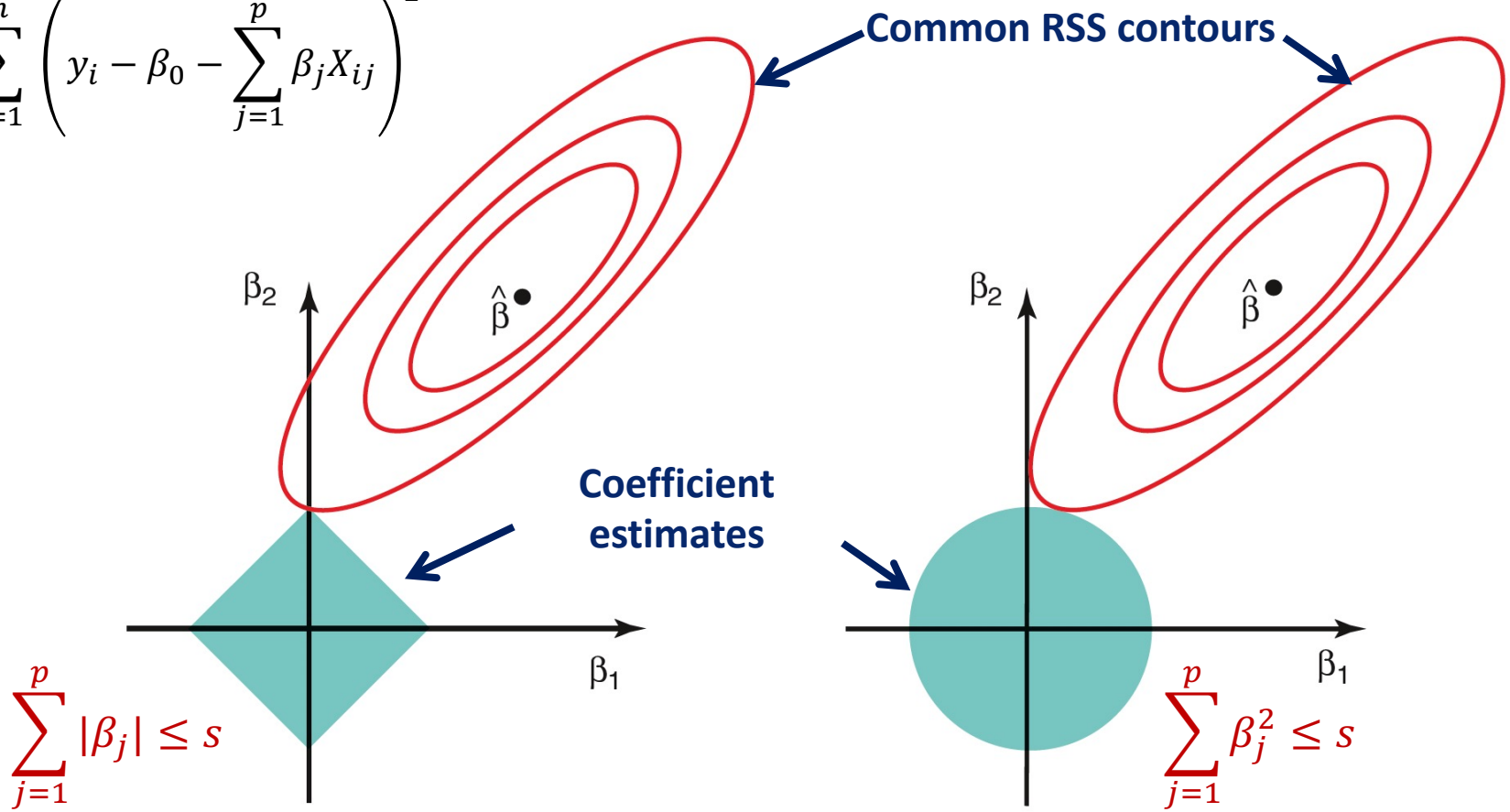


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \le s$ *and* $\beta_1^2 + \beta_2^2 \le s$, *while the red ellipses are the contours of the RSS.*

EXAMPLE: CREDIT DATA

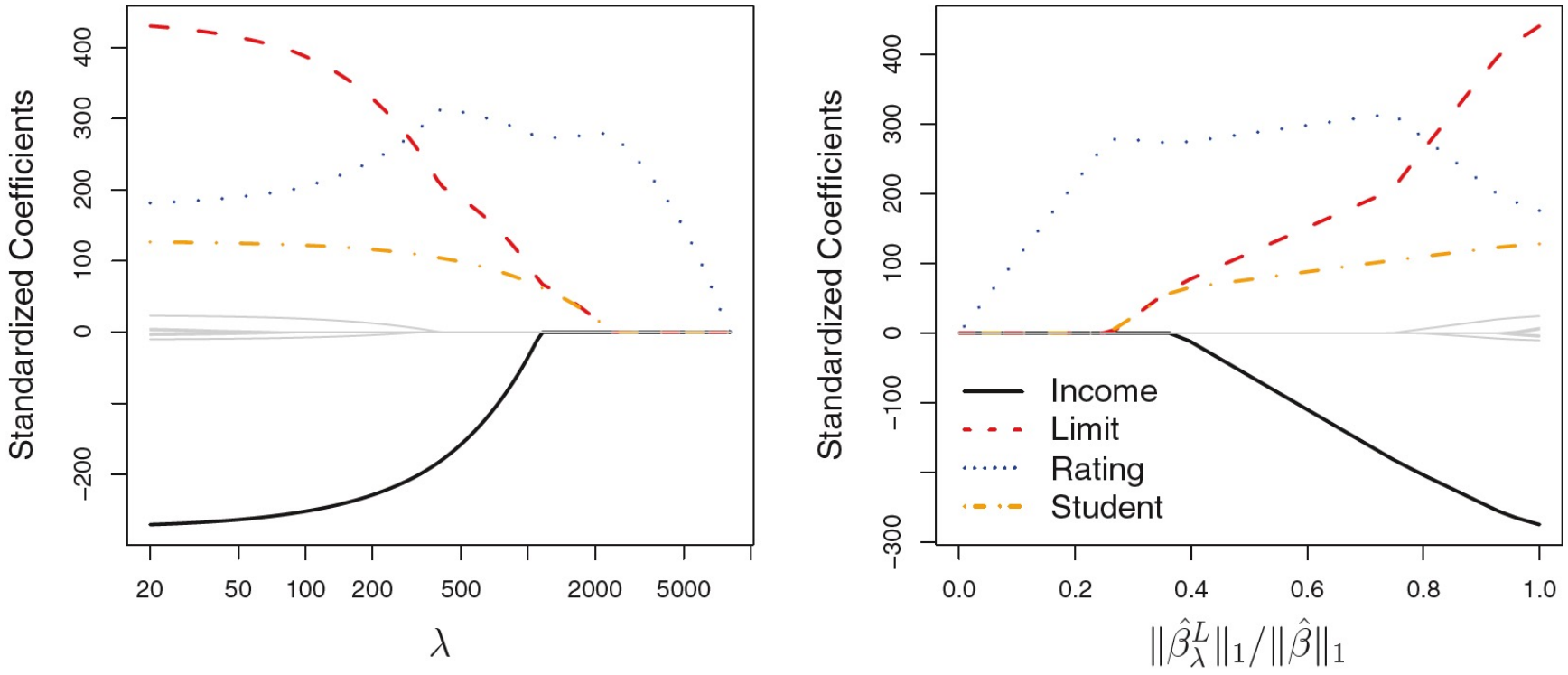


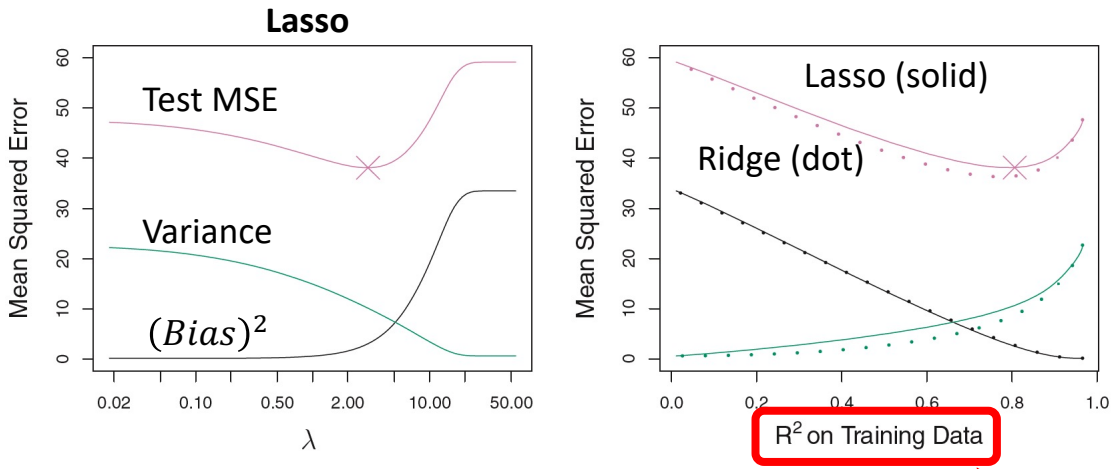
FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

RIDGE REGRESSION VS. THE LASSO

- Both significantly **reduce variance** at the expense of a **small increase in bias**
- **Interpretability**: Lasso > Ridge regression
- **Question**: What about the model accuracy? Which outperforms the other?

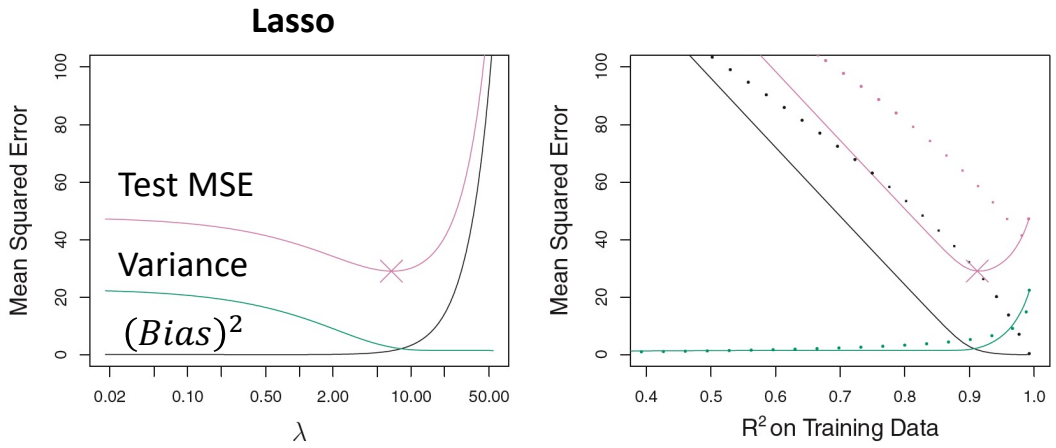
EXAMPLE: SIMULATED DATA

Simulated such that all 45 predictors are related to the response



Test MSE: Ridge < Lasso

Simulated such that 2 out of 45 predictors are related to the response



Test MSE: Ridge > Lasso

Ridge: $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2)$

Lasso: $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda|\boldsymbol{\beta}|)$

Using λ to compare different methods can be misleading since the meaning of λ can be different for different methods

CONCLUSION: RIDGE REGRESSION VS. THE LASSO

- Both significantly **reduce variance** at the expense of **a small increase in bias**
- **Interpretability**: Lasso > Ridge regression

- **Question**: What about the model accuracy? Which outperforms the other?
- **Answer**
  - In practice, neither ridge or Lasso dominates the other
  - When there are relatively many equally-important predictors, ridge regression will win
  - When there are small number of important predictors and many others that are not useful, the Lasso will win

- However, the number of useful features is never known a priori for real datasets
- Hence, **cross-validation** can be used to determine which approach is better on the real dataset.
  - We cannot use indirect estimates such as Adjusted R^2, AIC, BIC or C_p, because d for each λ is not known

OUTLINE

- Overview
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
- Considerations in High Dimensions

DIMENSION REDUCTION: OVERVIEW

- The methods that we have discussed so far in this lecture have *controlled variance* in two different ways
  - **1. Subset selection**
    - Selecting subset of variables
  - **2. Shrinkage method**
    - Shrinking the estimated coefficients toward zero
- All of these methods are defined using the original predictors $X_1, X_2, \dots, X_p$.
- **This implies that our data live in p-dimensional space, but what if not all p dimensions are equally useful?**
- We now explore a class of approaches that *transform* the predictors and then fit a least squares model using the transformed variables.
- We will refer to these techniques as *dimension reduction* methods.

DIMENSION REDUCTION: OVERVIEW

- **Big idea**: Transform the data before performing regression

$$\underbrace{[X_{i1}, X_{i2}, \dots, X_{i5}]}_{p\text{-dim}} \rightarrow \underbrace{[Z_{i1}, Z_{i2}]}_{m\text{-dim}}$$

- Then, instead of

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

- Let us solve the following!

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{m=1}^{M} \theta_m Z_{im} \right)^2$$

DIMENSION REDUCTION: DETAILS

- Let $Z_{i1}, Z_{i2}, \dots, Z_{iM}$ represent $M < p$ linear combinations of our original p predictors of ith observation X_i

$$Z_{im} = \sum_{j=1}^{p} \phi_{jm} X_{ij}$$

$\boldsymbol{X} \in \mathbb{R}^{n \times p}$

$\boldsymbol{Z} \in \mathbb{R}^{n \times M}$

$\boldsymbol{\phi} \in \mathbb{R}^{p \times M}$

  - $\phi_{m1}, \dots, \phi_{mp}$ are constants

- We can then fit the linear regression model using ordinary least squares (OLS)

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_{im} + \epsilon_i, \qquad i = 1, \dots, n$$

- The regression coefficients are now given by $\boldsymbol{\theta} = [\theta_0, \dots, \theta_M] \in \mathbb{R}^{(M+1)}$
  - Compare with the coefficients $\boldsymbol{\beta} = [\beta_0, \dots, \beta_p]^T \in \mathbb{R}^{(p+1)}$ in ordinary least squares

- A proper selection of $\phi_{1m}, \dots, \phi_{pm}$ can lead to a dimension reduced regression with $M + 1$ coefficients, which outperforms the original OLS regression with $p + 1$ coefficients
  - This is where the term *dimension reduction* comes from (Reduced from p to M)

DIMENSION REDUCTION: DETAILS

- Notice from $y_i = \theta_0 + \boxed{\sum_{m=1}^{M} \theta_m Z_{im}} + \epsilon_i$,

$$\sum_{m=1}^{M} \theta_m Z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} X_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} X_{ij} = \sum_{j=1}^{p} \beta_j X_{ij}$$

$$Z_{im} = \sum_{j=1}^{p} \phi_{jm} X_{ij} \qquad \beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

- Hence, the above model can be thought of as a special case of the original linear regression model.
- Instead of constraining the coefficients β_j as done by ridge and Lasso, β_j should take the form $\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$
  - This is another way to handle the bias-variance tradeoff.
  - This has the potential to bias the coefficient estimates, but when p is large, selecting $M \ll p$ can significantly reduce the variance of the fitted coefficients.
- **How can we fit a model on the dimension reduced data?**

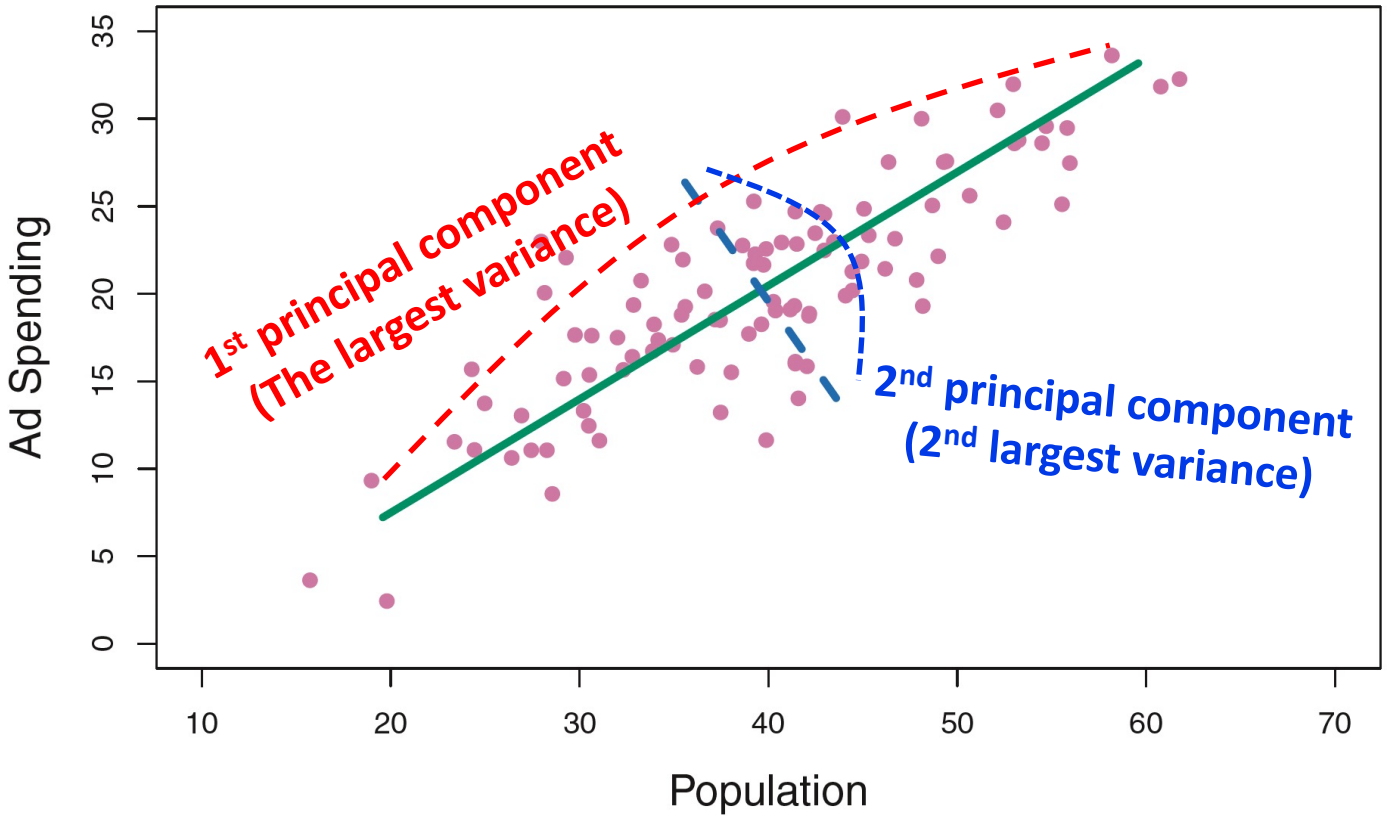
PRINCIPAL COMPONENTS REGRESSION

- We apply ***principal components analysis (PCA)*** to define the linear combinations of the predictors, for use in our regression.
  - More details on this in later lectures
  - Here, we use it as a **dimension reduction technique for regression**

- Given a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, **PCA derives linear combinations of the variables** such that
  - **1st principal component:** Defines a direction of the data along which the observations vary the most (largest variance)
  - **2nd principal component:** Defines a direction which is orthogonal to the first one, and along which varies the most (among directions that are orthogonal to the first principal component).
  - And so on…

- Generally, an mth principal component is defined such that it is orthogonal with the earlier $m-1$ principal components and captures most of the remaining variability

PICTURES OF PCA



Principal component loadings

$$Z_{i1} = \overset{\phi_{11}}{0.839} \times (\overset{X_{i1}}{pop_i} - \overline{pop}) + \overset{\phi_{21}}{0.544} \times (\overset{X_{i2}}{ad_i} - \overline{ad})$$

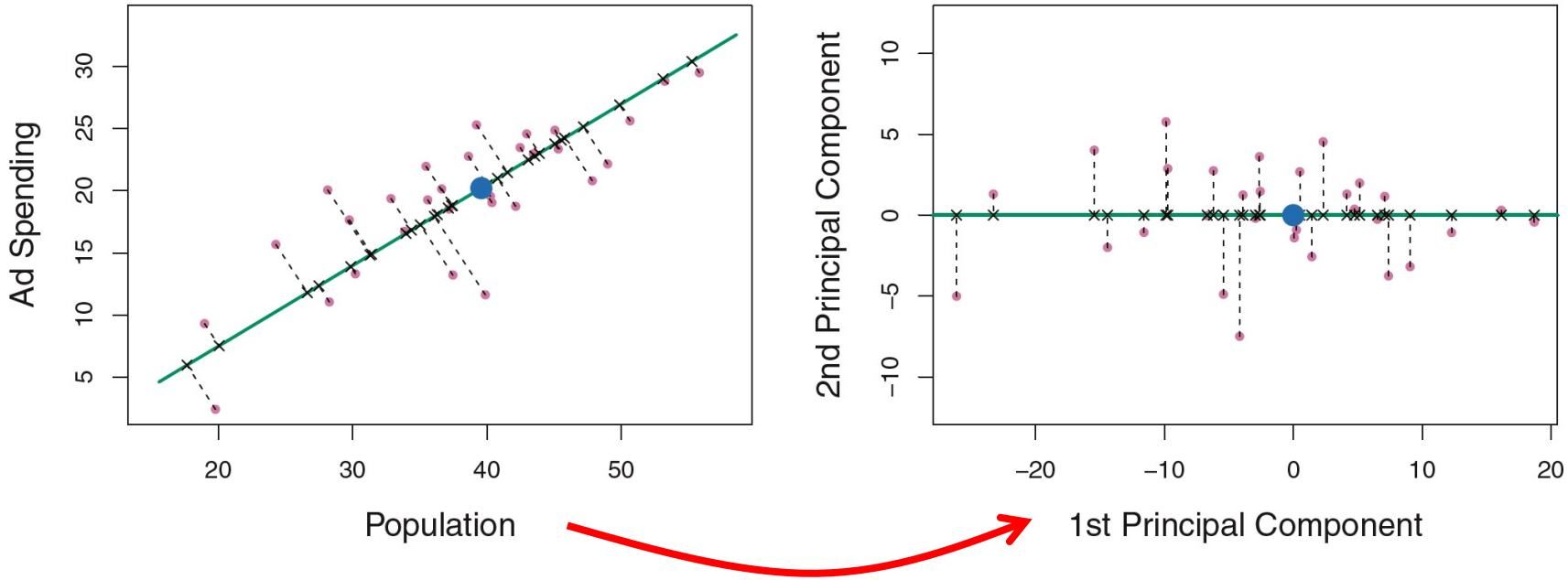
- Out of every possible linear combinations of pop and ad such that $\phi_{11}^2 + \phi_{21}^2 = 1$, this particular linear combination yields the **highest variance**
  - Why $\phi_{11}^2 + \phi_{21}^2 = 1$?
- Z_{i1}: Principal component score of the 1st principal component for the ith observation.

$$Z_{i2} = \overset{\phi_{12}}{0.544} \times (\overset{X_{i1}}{pop_i} - \overline{pop}) - \overset{\phi_{22}}{0.839} \times (\overset{X_{i2}}{ad_i} - \overline{ad})$$

- Z_{i2}: Principal component score of the 2nd principal component for the ith observation.

PICTURES OF PCA: PROJECTION

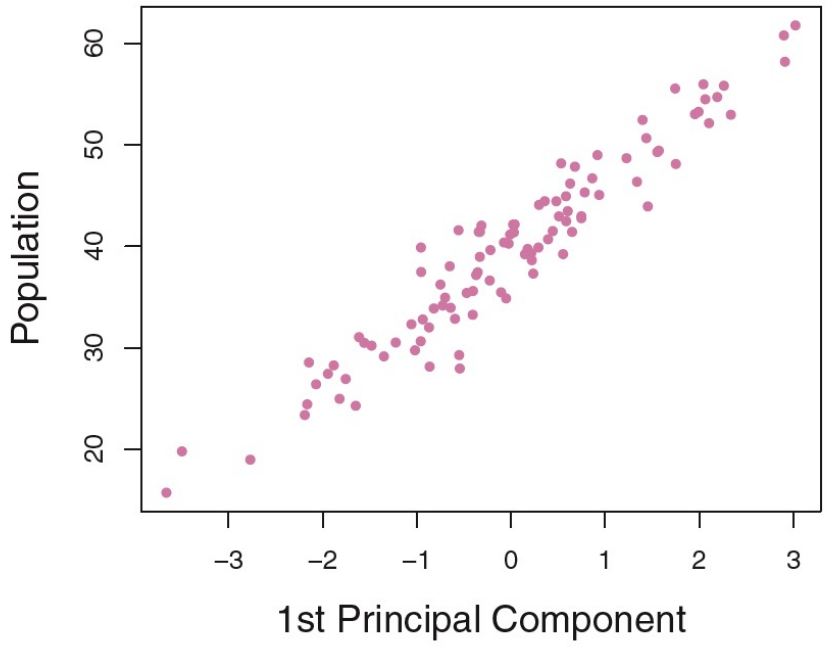
- **Projecting a point onto a line simply involves finding the location on the line which is closest to the point**
- The 1st principal component vector defines the line that is as close as possible to the data
- Difference with linear regression?

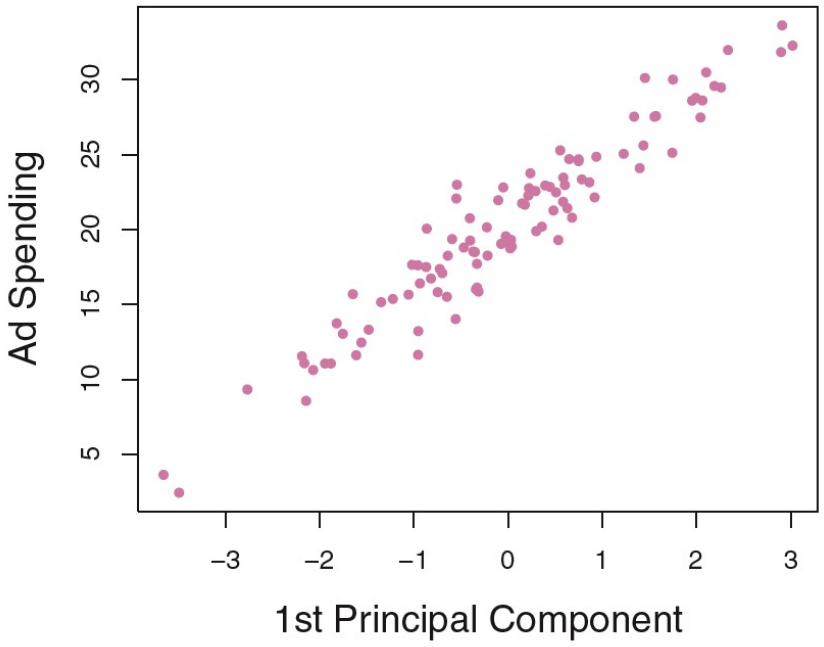


Rotated

$$Z_{i1} = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad})$$

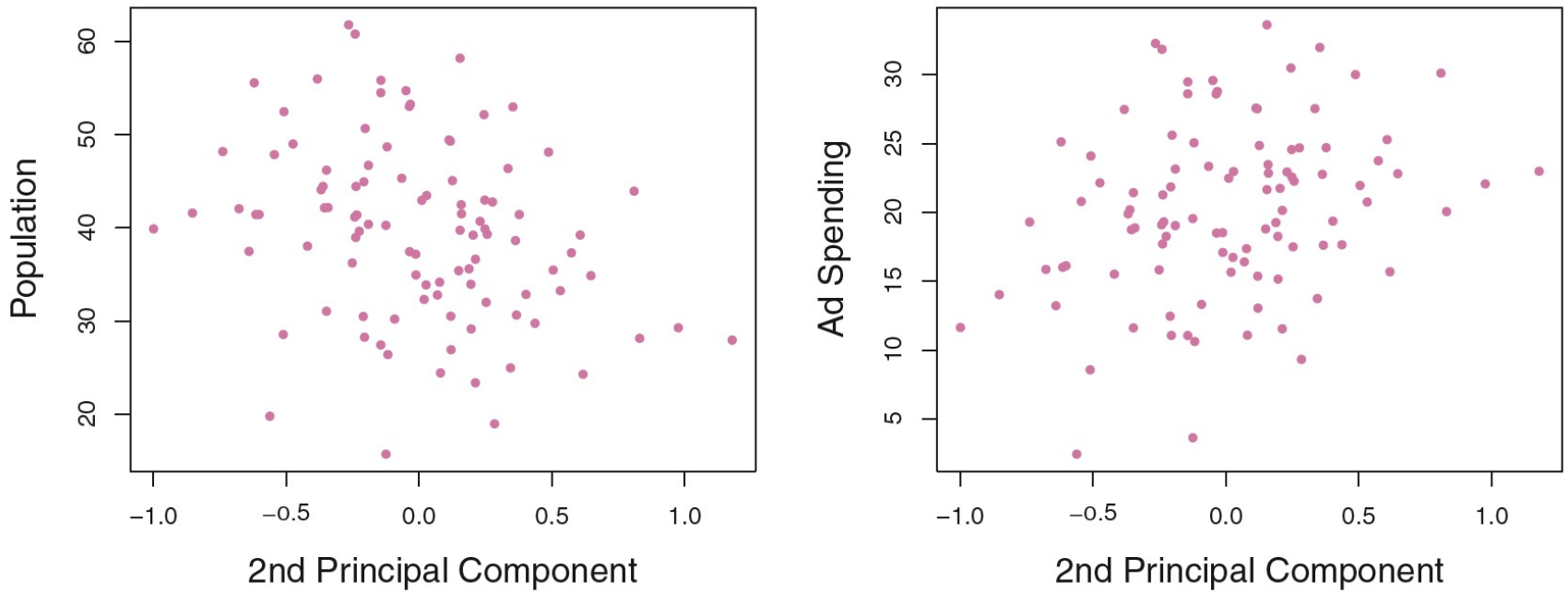
PICTURES OF PCA: 1ST PRINCIPAL COMPONENT





- The plots show a strong relationship between the 1st principal component and the two features (pop, ad).
- The 1st principal component appears to capture most of the information contained in the pop and ad predictors.

PICTURES OF PCA: 2ND PRINCIPAL COMPONENT



- There is little relationship between the 2nd principal component and these two predictors (pop, ad)
- In this case, we only need the 1st principal component to accurately represent the pop and ad budgets.

PROBLEMS WITH PCR

- PCR identifies linear combinations, or directions, that best represent the predictors $X_1, \ldots, X_p$.
- That is, principal components are selected based on predictors
  - This is ***unsupervised***, because the response Y is not used to help determine the principal component directions.
  - But, this is not what we are trying to predict!
- What if the values you're trying to predict aren't correlated with the first few components?
  - No guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.
- **Solution**: **Partial Least Squares (PLS)**

PARTIAL LEAST SQUARES (PLS)

- A supervised form of PCR
- Like PCR,
  - PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \dots, Z_M$ that are linear combinations of the original features $X_1, \dots, X_p$, and then fits a linear model via OLS using these M new features
- Unlike PCR,
  - PLS identifies these new features in a ***supervised way***, that is, it makes use of the response y in order to identify new features that not only approximate the old features well, but also that are related to the response.
- PLS approach attempts to find directions that help **explain both the response and the predictors**

PARTIAL LEAST SQUARES (PLS): DETAILS

- PLS computes the first direction Z_1 by setting each ϕ_{j1} equal to β_j in the simple linear regression

1st principal component $Z_{:,1}$

$$Z_{i1} = \sum_{j=1}^{p} \phi_{j1} X_{ij}$$

Linear regression

$$\beta = argmin_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2$$

- Recall that β_j represents the correlation between y and X_j.
- PLS places the highest weight on the variables that are most strongly related to the response.
- Subsequent directions are found by taking residuals and then repeating the above process

EXAMPLE: ADVERTISING DATA

- Advertising data
  - Response: Sales / Predictors: Population and Ad

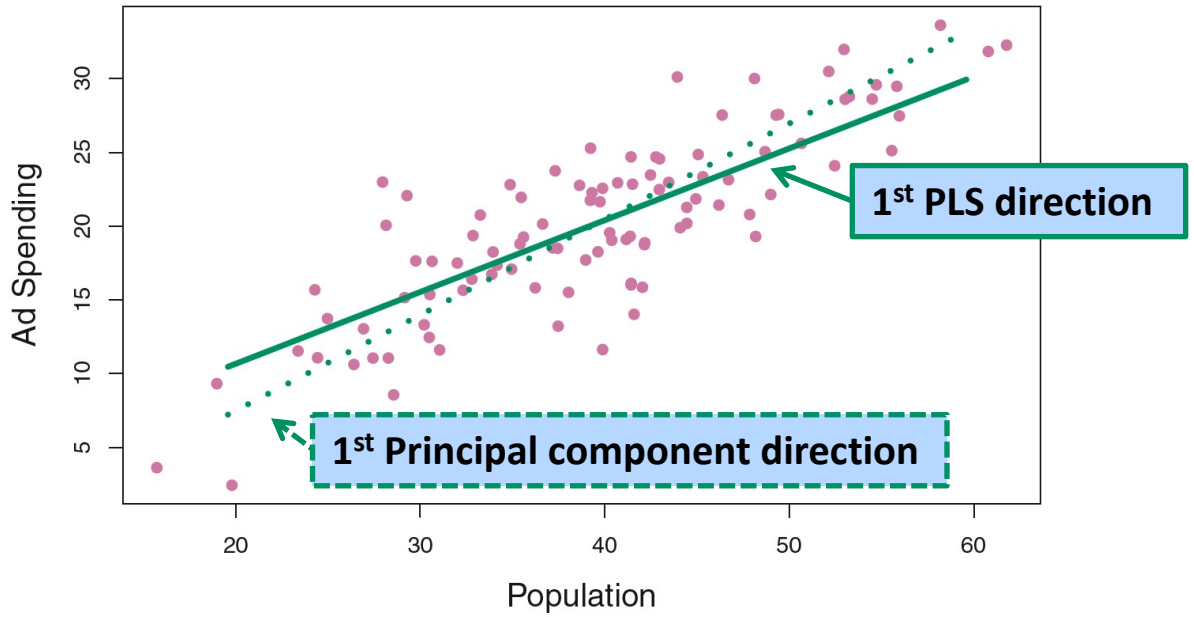


FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

- PLS has chosen a direction that has **less change in the ad-dim per unit change in the pop-dim**
- This implies that **pop is more highly correlated with the response** than ad

OUTLINE

- Overview
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
- Considerations in High Dimensions

CONSIDERATIONS IN HIGH DIMENSIONS

- Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting, where $n \gg p$
  - Ex) Patient data
    - Given: Age, Gender, BMI ($p = 3$)
    - Predict: Blood pressure of thousands of patients ($n = \text{thousands}$)

- However, there are also cases where p can be extremely large ($p \gg n$) → High-dimensional
  - Ex 1) DNA data
    - Half a million single nucleotide polymorphisms (SNP (단일염기변형)) of patients to predict the blood pressure
  - Ex 2) Search data
    - Search terms in “bag-of-words” of few thousands of users who have agreed to share their information to predict clicks

WHAT GOES WRONG IN HIGH DIMENSIONS?

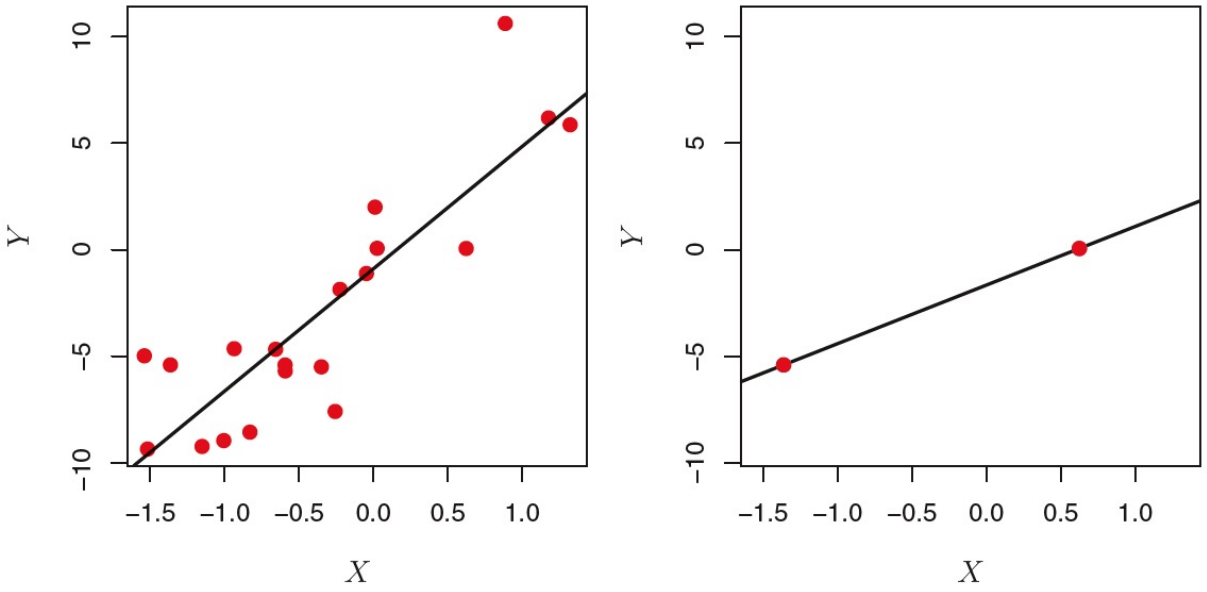
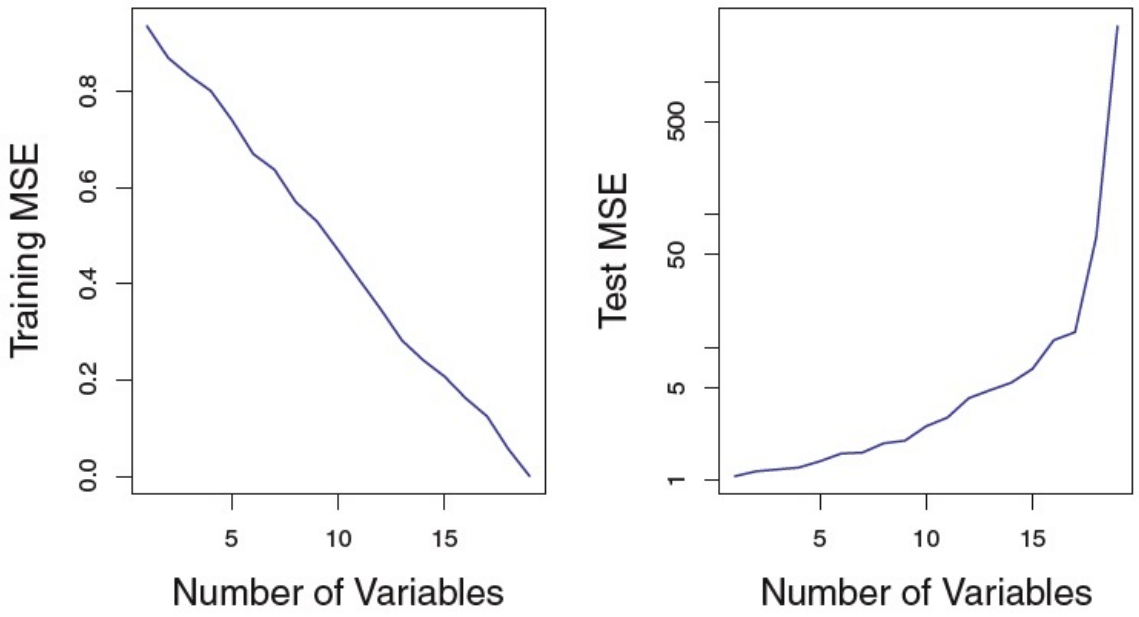


FIGURE 6.22. Left: *Least squares regression in the low-dimensional setting.* Right: *Least squares regression with $n = 2$ observations and two parameters to be estimated (an intercept and a coefficient).*

Perfect fit will almost certainly lead to overfitting of the data

$n = 20$
$p = 1 \sim 20$



We should use methods like forward stepwise selection, ridge regression, lasso and PCR

INTERPRETING RESULTS IN HIGH DIMENSIONS

- We must be quite cautious in the way that we report the results obtained
- **Multi-collinearity**: Predictors are correlated with each other
  - Any variable in the model can be written as a linear combination of all of the other variables in the model
- Multi-collinearity is more severe in high-dimensional data
  - We can never know exactly which variables (if any) truly are predictive of the outcome
  - We can hope to assign large regression coefficients to variables that are correlated with the variables that truly are predictive of the outcome
- **Example:** Assume forward stepwise selection selected 17 variables from half a million SNPs.
  - It would be incorrect to conclude that these 17 SNPs predict blood pressure more effectively than the other SNPs not included in the model.
  - There are likely to be many sets of 17 SNPs that would predict blood pressure just as well as the selected model.

In high-dimensional setting, we must be careful not to overstate the results

CONCLUSION

- How to regularize the model when we have many predictors?
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection (Forward/Backward)
- Shrinkage Methods
  - Ridge Regression
  - The Lasso
- Dimension reduction
  - Principal components regression (PCR)
  - Partial least squares (PLS)
- Considerations in High Dimensions
  - Cautions about overstating the result

Coming up next:
Moving Beyond Linearity