Machine Learning in Public Health

Lecture 6: Linear Model Selection and Regularization

Dr. Yang Feng

Linear Model Selection and Regularization

Consider the linear regression model.

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

- One typically fits this model using least squares.
- When many β_j's are 0, we might want to use another fitting procedure instead of least squares to yield better prediction accuracy and model interpretability
  - *Best Subset Selection*
  - *Forward Regression*
  - *Ridge Regression*
  - *LASSO*

Best Subset Selection

- To perform best subset selection, we fit a separate least squares regression for each possible combination of the p predictors.
- Potential problem: selecting the best model from among too many possibilities considered by best subset selection is not trivial.
- Best subset selection is usually implemented by:

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.

Choose the model in Step 3)

There are essentially two ideas in this step

- directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in Chapter 5.
- indirectly estimate test error by making an adjustment to the training error (e.g., adjusted R^2, AIC, BIC and C_p).

Criteria for comparing models

For a candidate model M_d with d variables, we define the following criteria.

- Adjusted R^2:

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

Q: how is adjusted R^2 different from R^2?

- Trade-off between goodness-of-fit (RSS) and model complexity (d).

- Mallow's C_p

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

where $\hat{\sigma}^2$ is the estimated variance of random error term ϵ using the full model (i.e., p predictors).

Criteria for comparing models

- Akaike information criterion (AIC)

$$AIC = -2\log L + 2d,$$

where $\log L$ is the log-likelihood for the current model M_d. For linear regression, we have

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

- Bayesian information criterion(BIC)

$$AIC = -2\log L + (\log n)d$$

For linear regression, we have

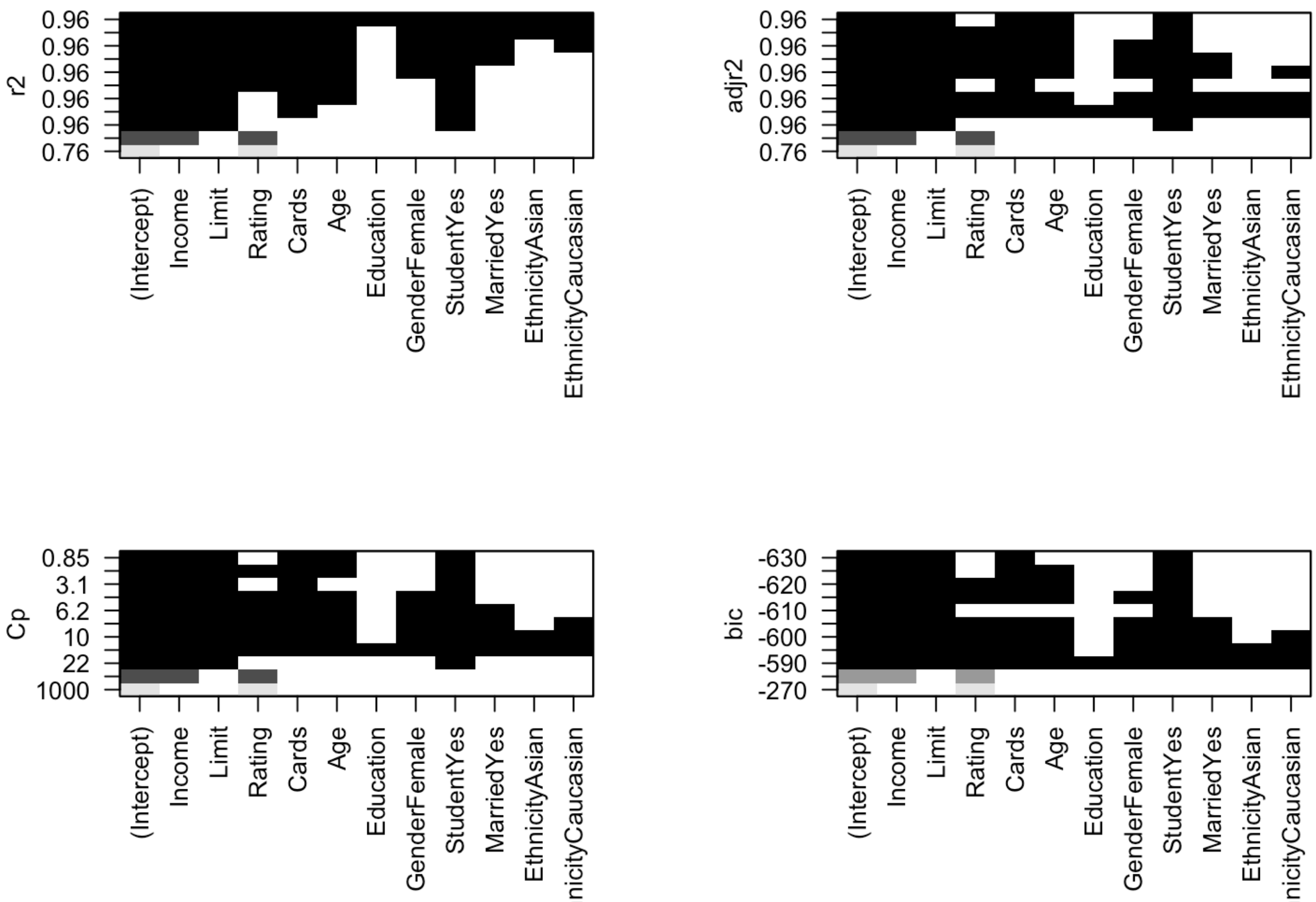
$$BIC = \frac{1}{n}\left(RSS + \log(n)d\hat{\sigma}^2\right)$$

A few questions

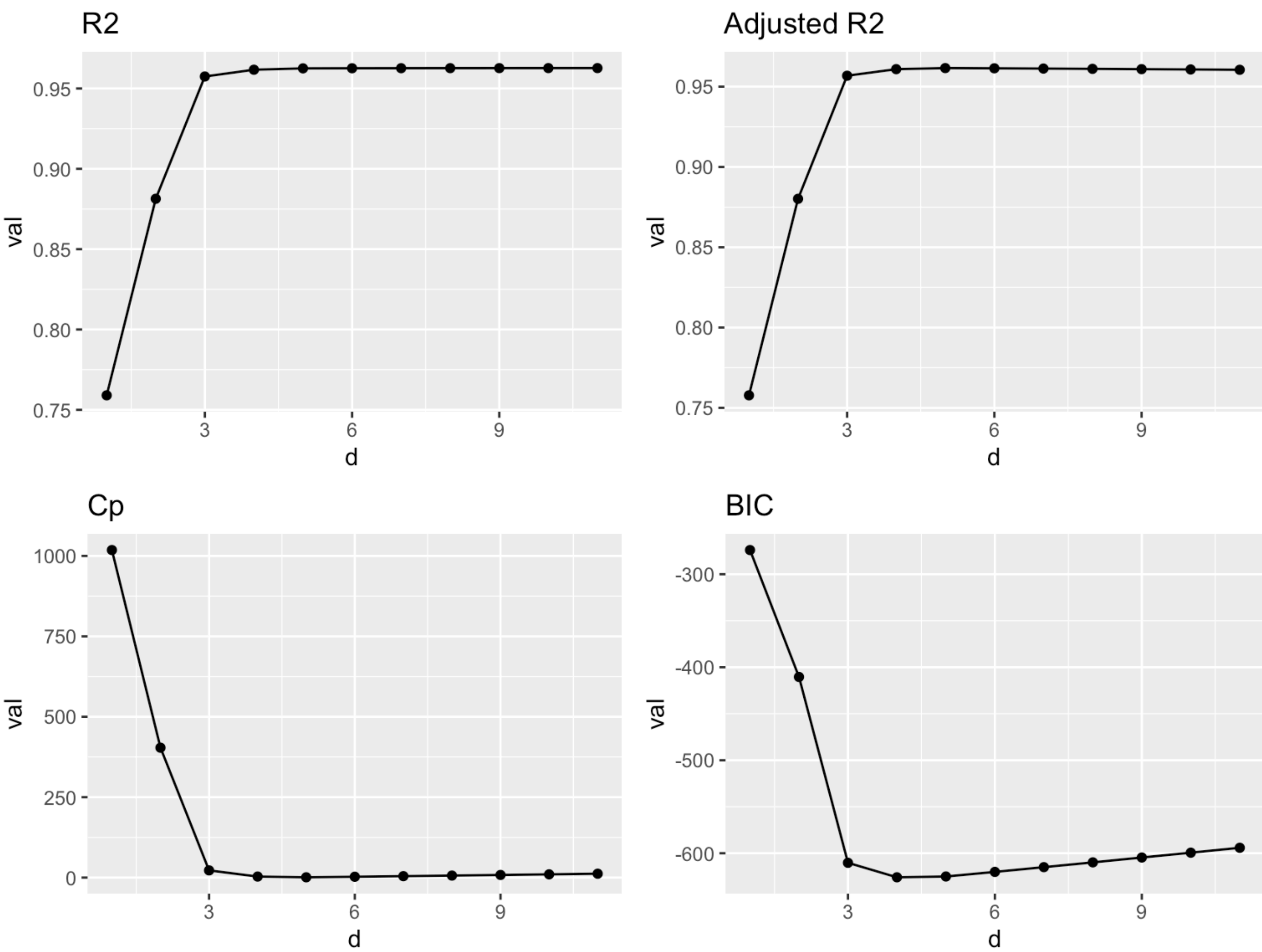
- When n is bigger than $7, \log n > 2$. This means BIC penalizes larger models heavier compared to AIC. So which criterion encourages smaller models?
- Among adjusted R^2, AIC, C_p, BIC and Cross-Validation, which one is easier to be generalized beyond least squares linear regression?

CV - no term for model complexity, only need to get the error,
thus it works better/more generalized for some other non-parametric models bc it does not require a likelihood function

Optimal model for each size



Optimal model for each size: four methods



Coefficients corresponding to optimal models

```
coef(best_subset, 1:11)
```

```
## [[1]]
## (Intercept)      Rating
## -354.238144    2.501169
##
## [[2]]
## (Intercept)      Income      Rating
## -548.260759   -7.684998    4.002145
##
## [[3]]
##  (Intercept)       Income        Limit   StudentYes
## -456.4610300   -7.8639273    0.2723742  429.2673608
##
## [[4]]
##  (Intercept)       Income        Limit        Cards   StudentYes
## -523.0518688   -7.7925423    0.2712276   23.0613375  437.2257637
##
## [[5]]
##  (Intercept)       Income        Limit        Cards          Age   StudentYes
## -474.4151238   -7.6386077    0.2695746   23.6202781   -0.8761518  435.8910167
##
## [[6]]
##  (Intercept)       Income        Limit       Rating        Cards          Age
## -483.1236387   -7.6550380    0.2425861    0.4057927   21.4430273   -0.8818302
##   StudentYes
##  434.7457520
##
## [[7]]
##  (Intercept)       Income        Limit       Rating        Cards          Age
## -479.7062751   -7.6587166    0.2432506    0.3961782   21.4556382   -0.8844975
## GenderFemale   StudentYes
##   -5.9442936  434.5896004
##
## [[8]]
##  (Intercept)       Income        Limit       Rating        Cards          Age
## -476.0559413   -7.6549983    0.2430891    0.3995257   21.4452607   -0.8982854
## GenderFemale   StudentYes   MarriedYes
##   -5.6678976  433.4931082   -5.4990089
##
## [[9]]
##       (Intercept)             Income              Limit             Rating
```

```
##       -478.2999819        -7.6539473         0.2430925         0.3992281
##              Cards               Age      GenderFemale        StudentYes
##         21.4787522        -0.8909104        -5.6975997       433.3446404
##         MarriedYes EthnicityCaucasian
##         -5.5205462         3.6022736
##
## [[10]]
##        (Intercept)            Income             Limit            Rating
##       -482.4570451        -7.6484920         0.2412729         0.4253337
##              Cards               Age      GenderFemale        StudentYes
##         21.4309499        -0.8814066        -5.7321331       433.2456953
##         MarriedYes    EthnicityAsian EthnicityCaucasian
##         -6.0516622         6.2314770         6.8553322
##
## [[11]]
##        (Intercept)            Income             Limit            Rating
##       -477.5268904        -7.6464664         0.2415817         0.4201142
##              Cards               Age         Education      GenderFemale
##         21.4194258        -0.8833370        -0.3266788        -5.7043018
##         StudentYes        MarriedYes    EthnicityAsian EthnicityCaucasian
##        433.3529352        -6.1769790         6.3441993         6.8727948
```

```
coef(best_subset, 4)
```

```
##  (Intercept)       Income        Limit        Cards   StudentYes
## -523.0518688   -7.7925423    0.2712276   23.0613375  437.2257637
```

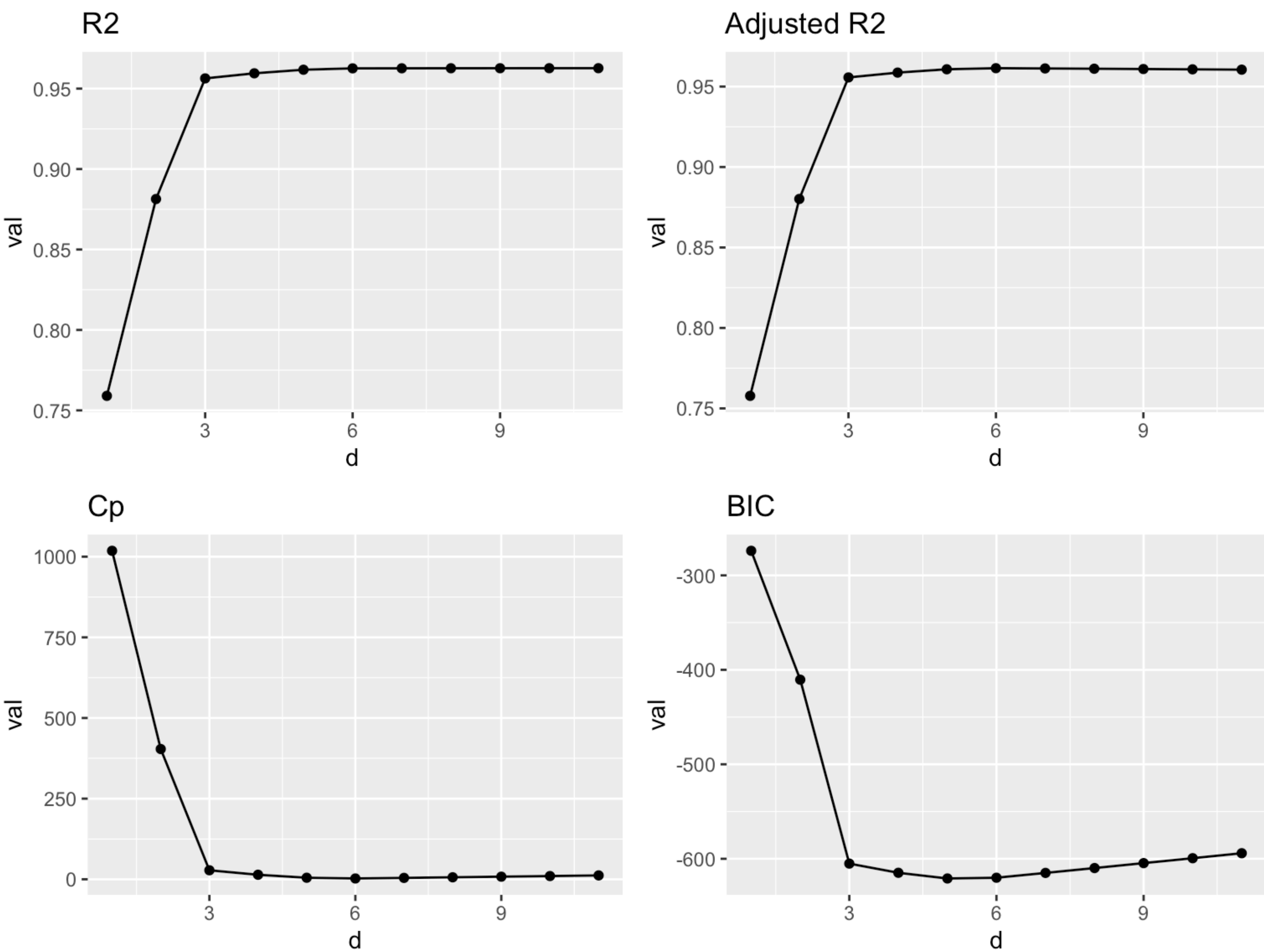
Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

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

```
##  (Intercept)       Income        Limit       Rating        Cards   StudentYes
## -531.3130997   -7.8084975    0.2465294    0.3715033   21.0647478  436.1851935
```

Optimal model for each size: four methods



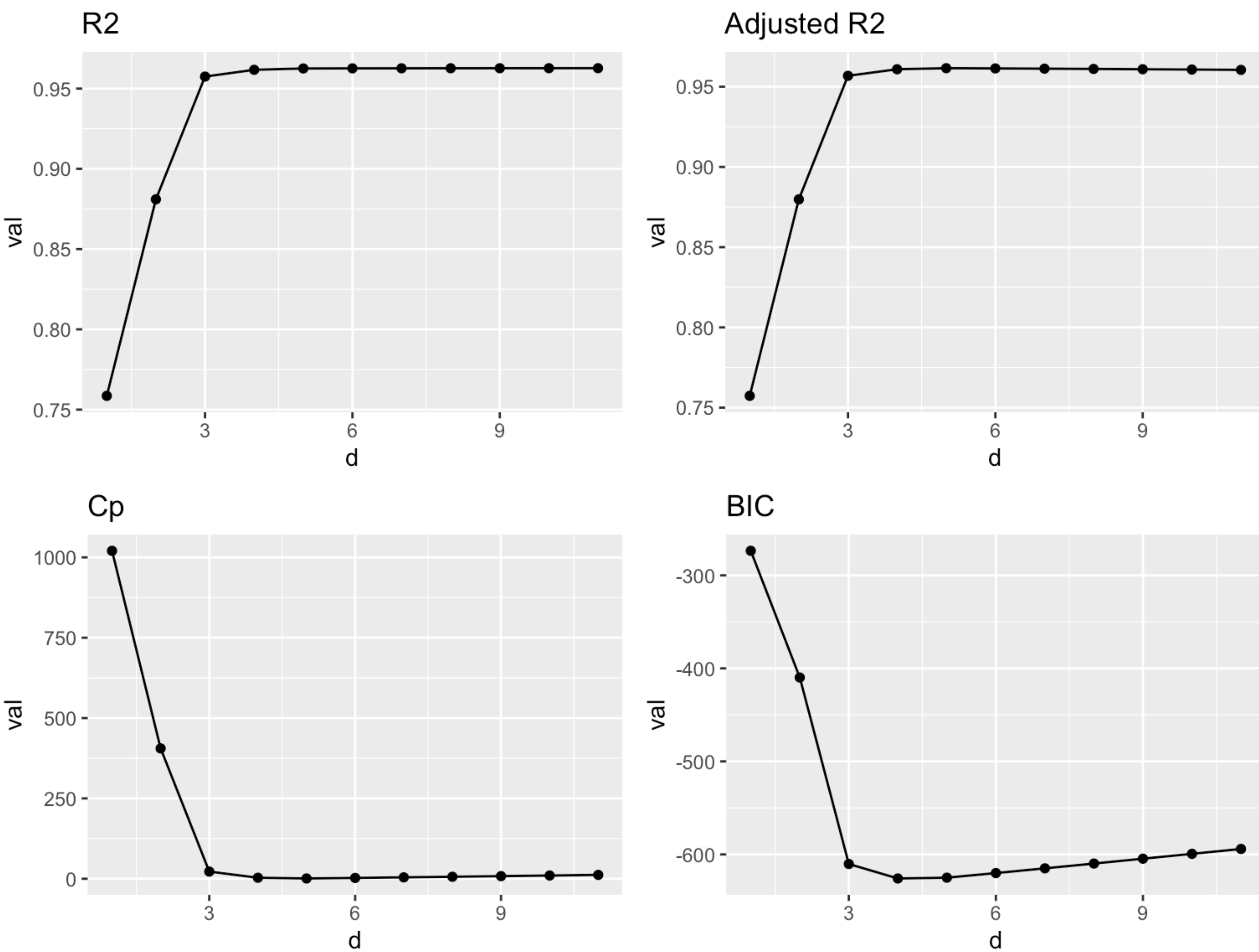
Backward Stepwise Selection

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.

```
##  (Intercept)       Income        Limit        Cards   StudentYes 
## -523.0518688   -7.7925423    0.2712276   23.0613375  437.2257637
```

Optimal model for each size: four methods



Comparison

- In either forward or backward selection, we search through $1 + p(p+1)/2$ models. This is a huge saving compared with best subset selection.
- However, forward stepwise selection and backward stepwise selection might miss the optimal subset of features. This is a price we have to pay for computational advantages.
- There are hybrid approaches that combine forward and backward selection.

Shrinkage Methods

- We can fit a model containing all p predictors using a technique that *constrains* or *regularizes* the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- Shrinking the coefficient estimates can significantly reduce their variance (with some cost in bias).
- Best known shrinkage methods: ridge regression and the lasso.
- The ridge regression coefficient estimates $\hat{\beta}_j^R$ are the values that minimize

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

  where $\lambda \geq 0$ is tuning parameter, and x_{ij} is the jth coordinate of the ith observation x_i.
- Lasso: find $\hat{R}_j^L$ that minimizes

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

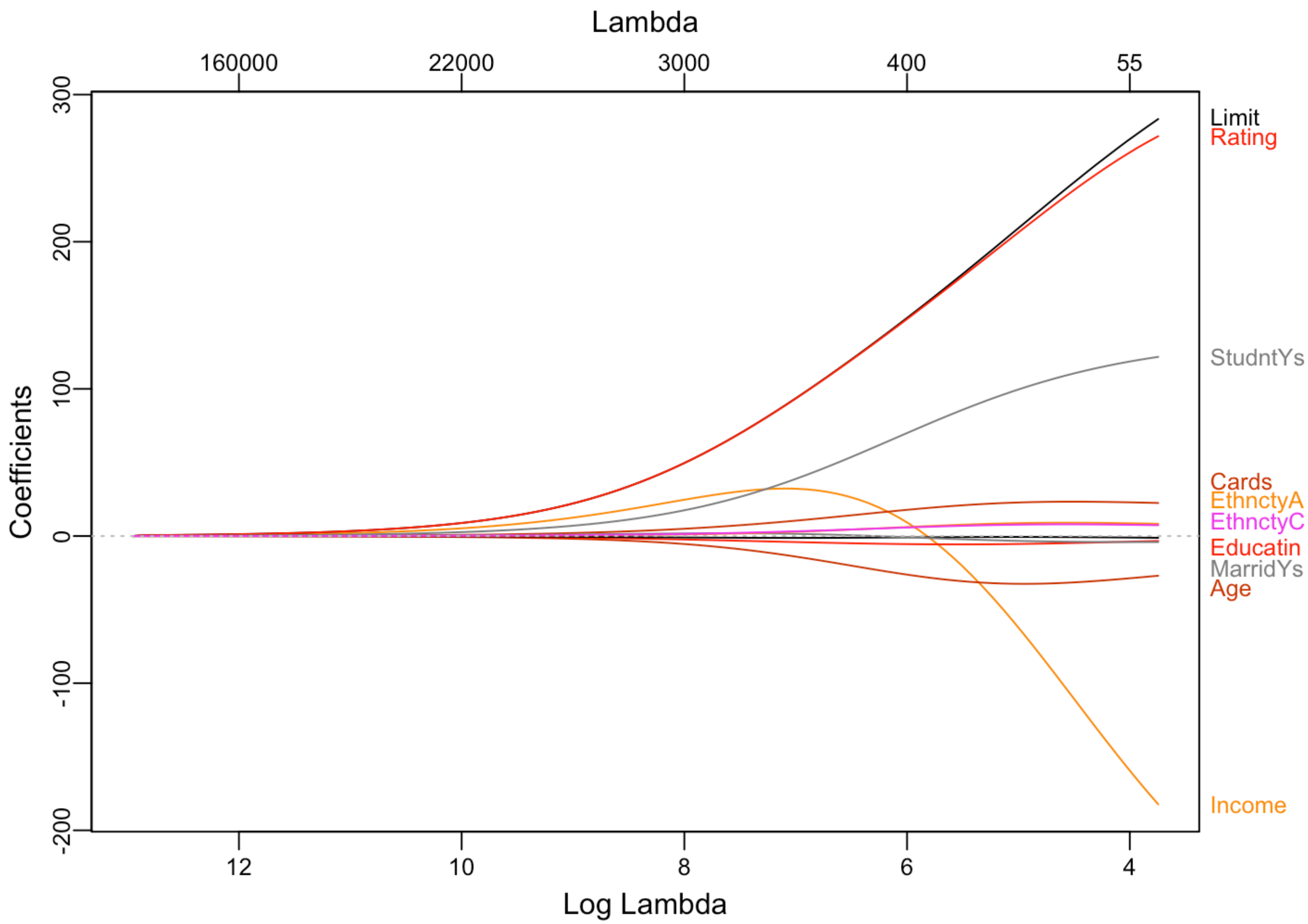
Importance of standardization

- In contrast to the usual least squares approach, standardizing the predictors is important in shrinkage methods. Why?
- To standardize the predictors:

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

- λ is an important penalty parameter that controls the amount of shrinkage.
- What happens when $\lambda = 0$ and $\lambda \to \infty$?
- How do we choose the tuning parameter λ ? Cross-validation
- Lasso tends to give sparser models compared to ridge (better for model interpretability), and it tends to perform better when the true model is sparse.
- But we do not know which is better for prediction accuracy.

Ridge regression does NOT give you sparse models



Bias-variance trade-off for ridge regression

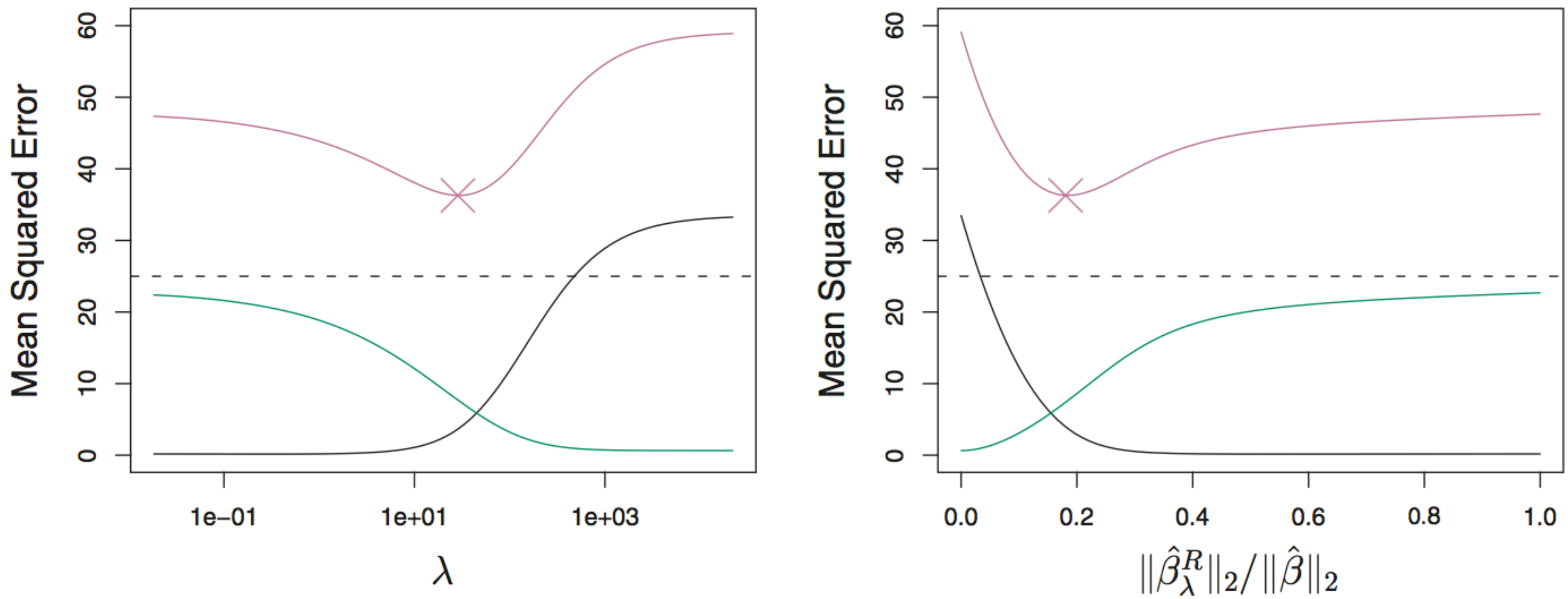
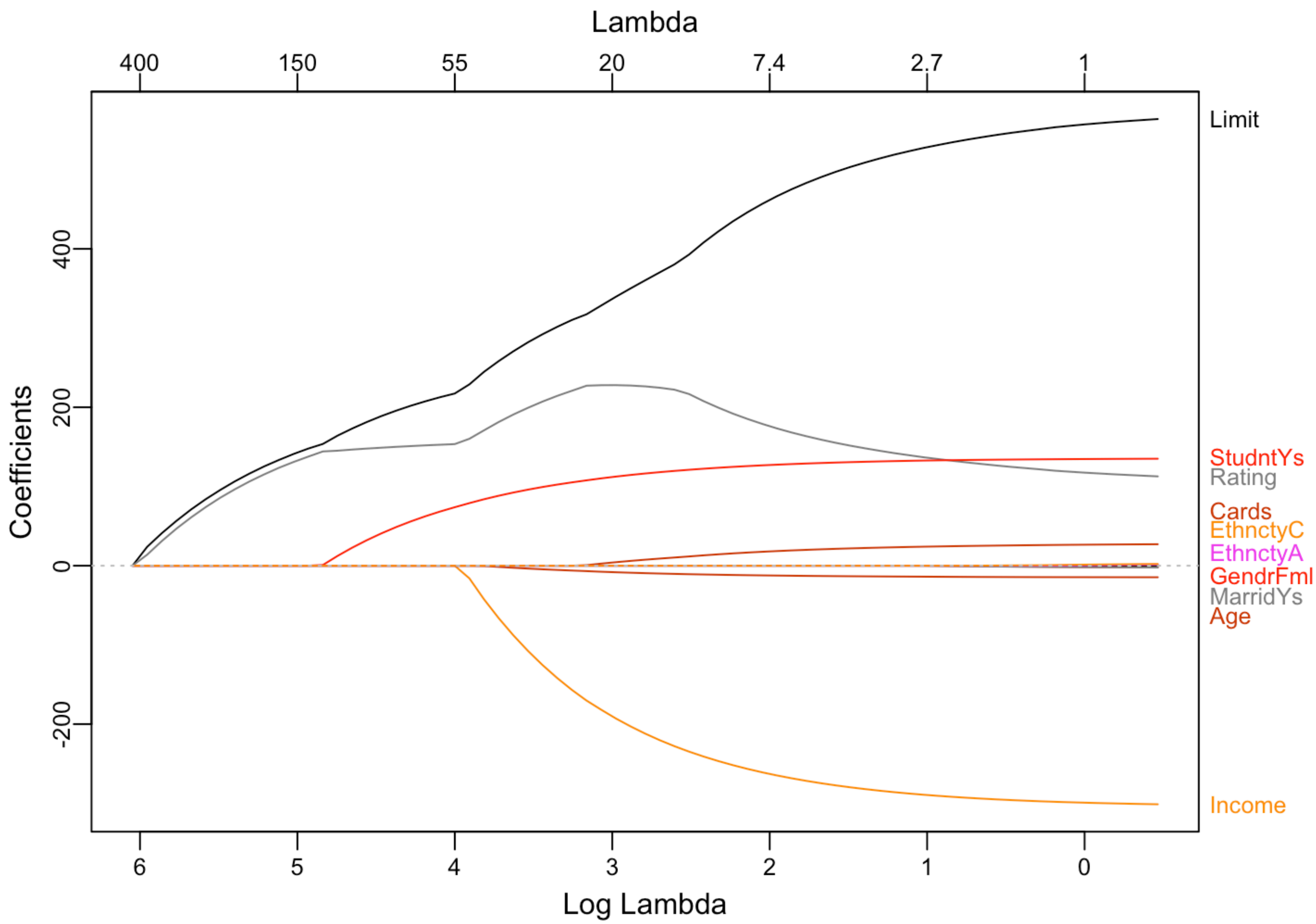


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses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

Lasso encourages sparse models



Another fomulation of ridge and lasso

- The Lasso's sparsity is better interpreted by an alternative formulation of Lasso and ridge regression
- λ and s has some corresponding relationships.

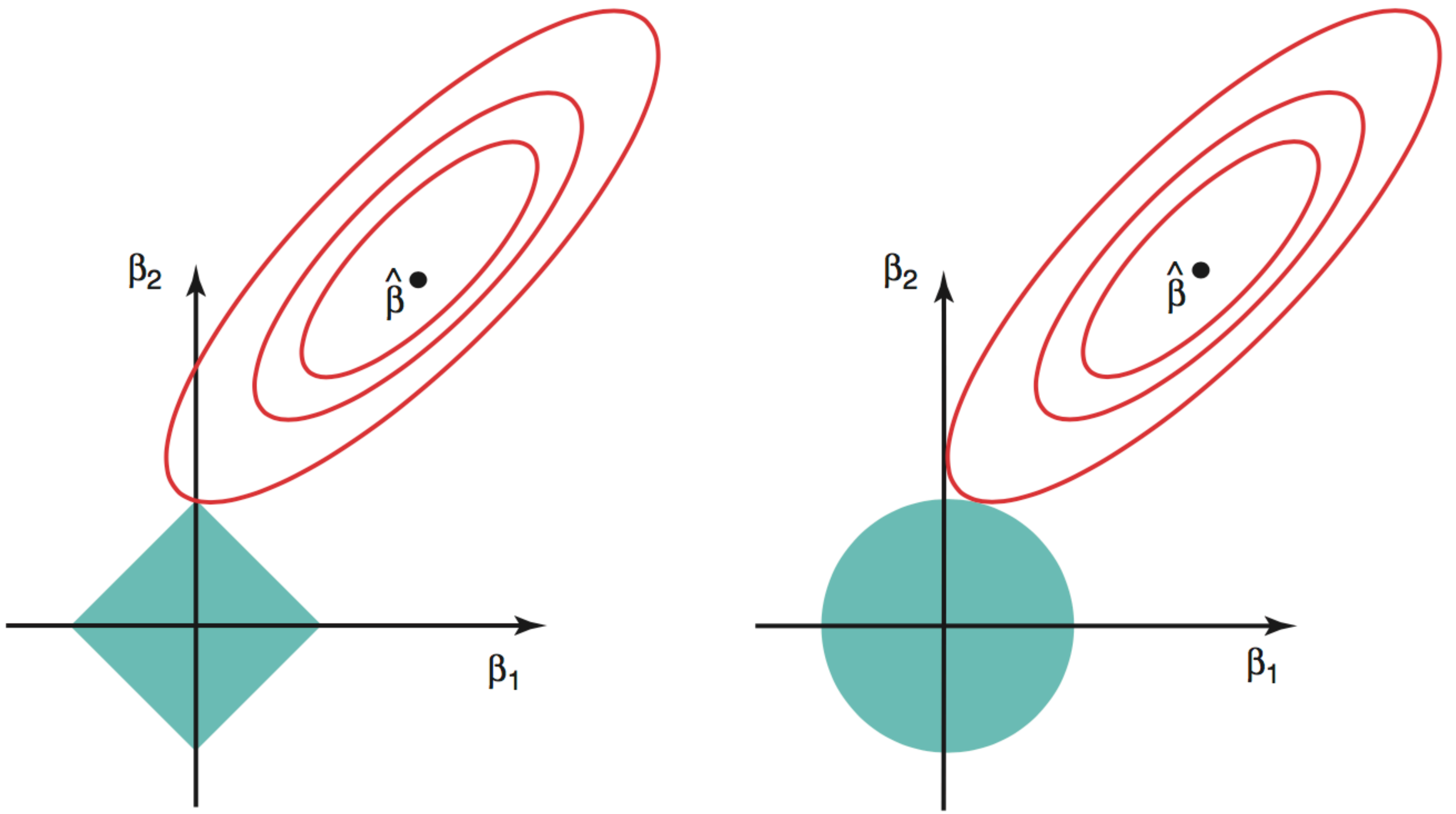


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.*

- How are they related to best subset selection?

A simple special case for ridge regression and the Lasso

- Consider a simple case with $n = p$, and X is a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements
- Assume that we perform regression without an intercept
- Under these assumptions, the approach amounts to finding $\beta_1, \cdots, \beta_p$ that minimize

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

- The least squares solution is given by

$$\hat{\beta}_j = y_j$$

- In this setting, amounts to finding $\beta_1, \cdots, \beta_p$ to minimize

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

- The ridge regression estimates take the form

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

A simple special case for ridge regression and the Lasso (cont')

- The amounts to finding coefficients to minimize

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

- The lasso estimate takes the form

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

or

$$\hat{\beta}_j^L = sign(y_j)[abs(y_j) - \lambda/2]_+$$

A simple special case for ridge regression and the Lasso (cont’)

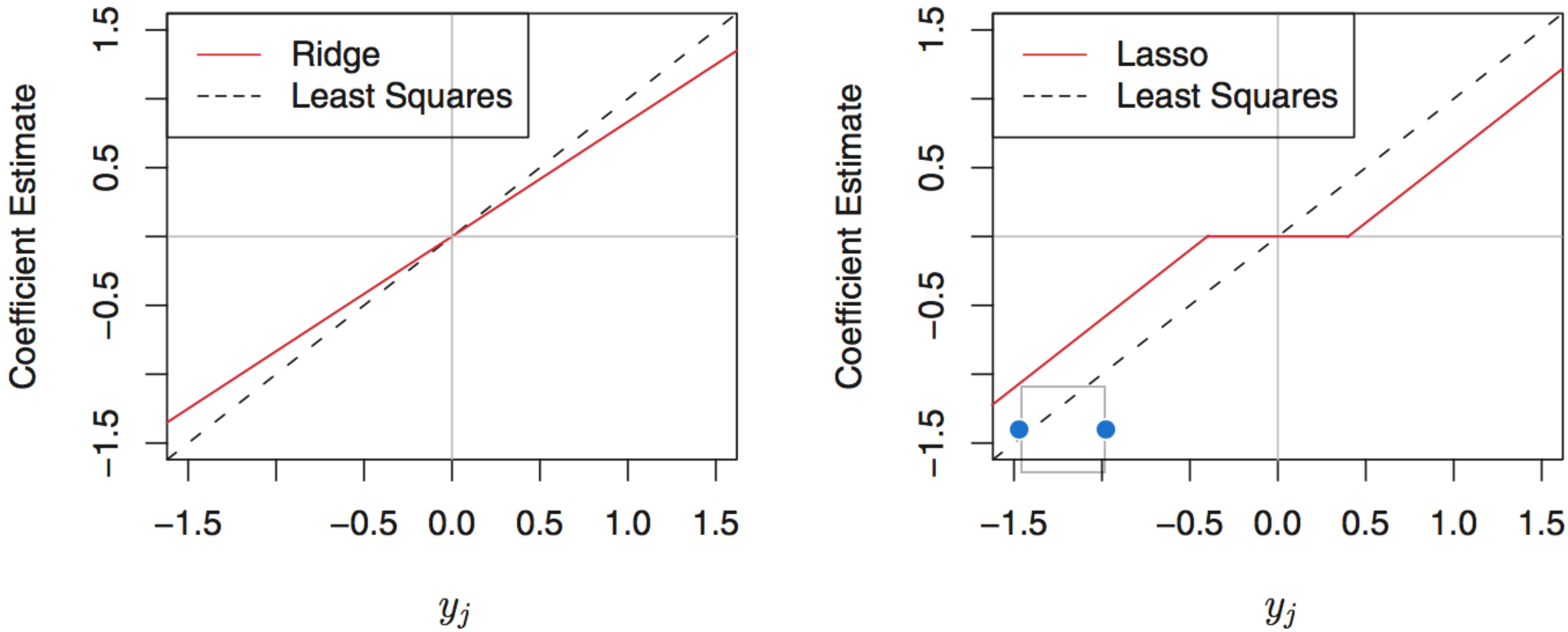


FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: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

High-dimensional setting

- High-dimensional settings: the scenarios where the number of predictors p is bigger than the sample size n
- A situation common in modern biology and medical sciences, but probably less so in business
- Including more variables into the regression, we potentially might find some useful features, but this benefit needs to be weighted against including many noise features.
- Example: Suppose there are in total 1000 features, 10 features are useful for the outcome, and the remaining 990 are noise features. We fix sample size (say at $n = 50$). Please compare the following three scenarios
  - *1. Use all 10 useful features for prediction*
  - *2. Use 9 of the 10 useful features for prediction*
  - *3. Use all 10 useful features, plus 100 noise features for prediction*
- Here, 1 is clearly better than 3 (no decrease in bias, but increase in variance). It is hard to compare 1 and 2 only based on this abstract description (think about variance and bias again).

Next Class

- Regression and Classification Trees