DATASCI 347 Machine Learning

Lecture 10: Regularization

Ruoxuan Xiong

Suggested reading: ISL Chapter 6

Lecture plan

- Comparison between ridge regression and lasso

- Elastic net

- Lab session

Ridge regression

- Ridge regression minimizes

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{i,j}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

  - $X_{i,j}$: j-th predictor of i-th observation
  - $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$: $\|\beta\|_2$ is called the ℓ_2 norm of $\beta \in \mathbb{R}^p$
  - β_0: mean of Y_i
  - Shrinkage penalty λ does not apply to β_0

Least absolute shrinkage and selection operator (Lasso)

- Lasso minimizes

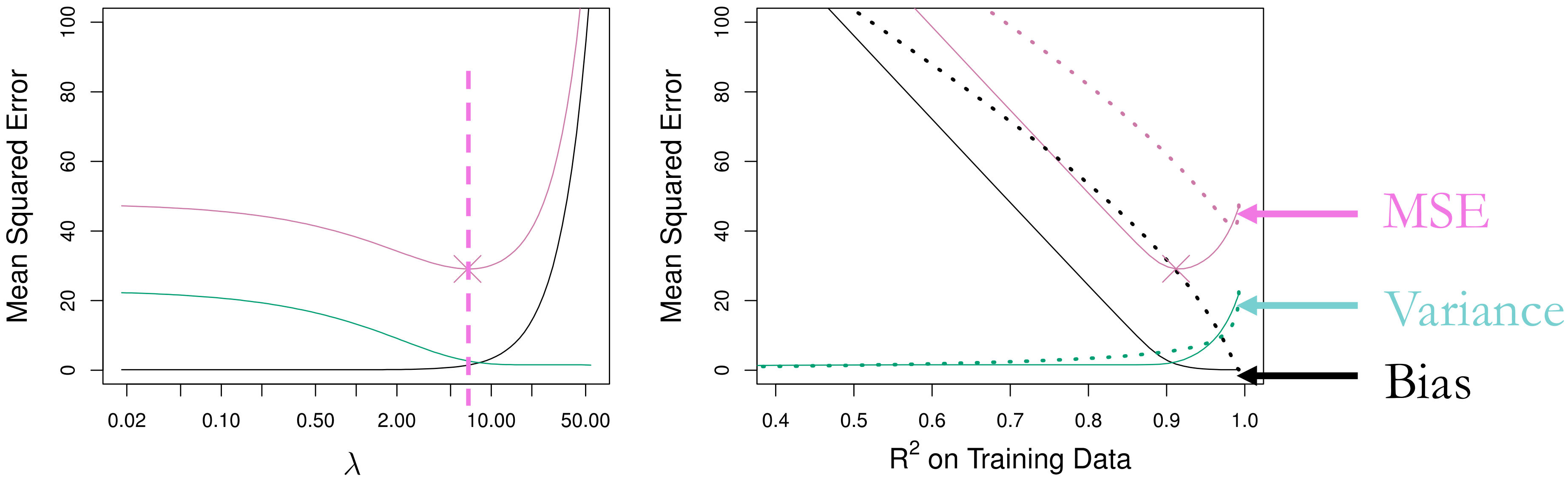
$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{i,j}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

  - $X_{i,j}$: j-th predictor of i-th observation
  - $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$: $\|\beta\|_1$ is called the ℓ_1 norm of $\beta \in \mathbb{R}^p$
  - β_0: mean of Y_i
  - Shrinkage penalty λ does not apply to β_0

Lasso vs. Ridge regularization

- **Simulation I:** Only 2 coefficients are non-zero
  - Simulated data: 45 predictors, 2 out of $\beta_1, \ldots, \beta_{45}$ are nonzero

Solid lines (—): Lasso
Dash lines (⋯): Ridge

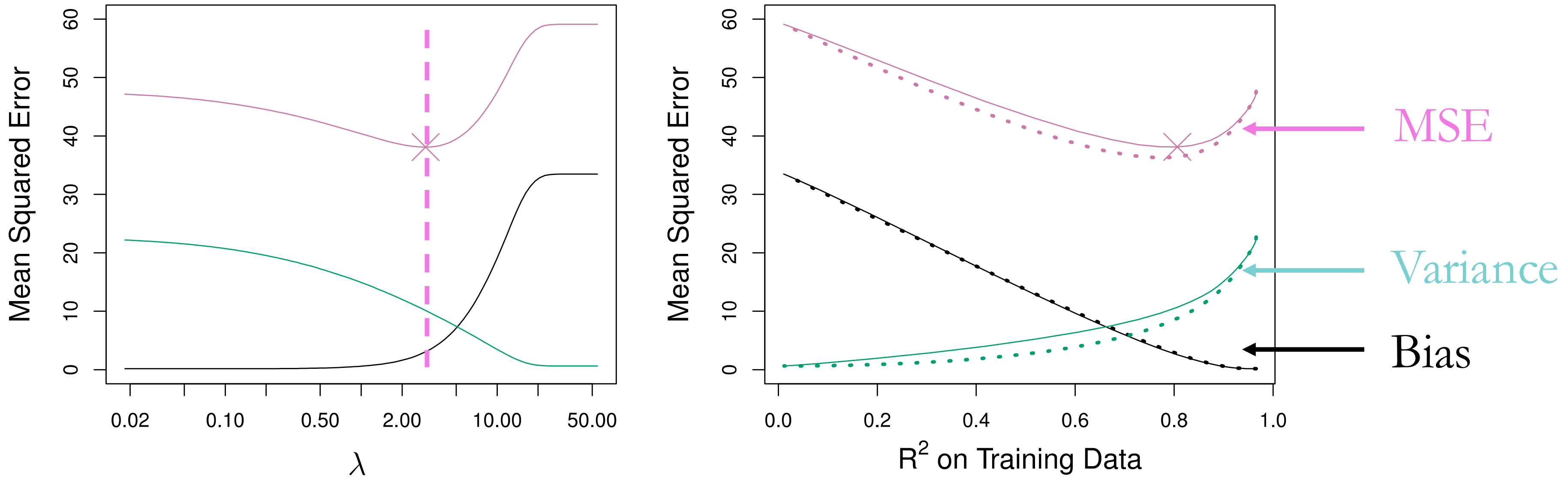


- The **bias**, variance, and MSE are all lower for the lasso

Lasso vs. Ridge regularization

- **Simulation II:** Most of the coefficients are non-zero
  - Simulated data: 45 predictors $\beta_1, \ldots, \beta_{45}$ are nonzero

Solid lines (—): Lasso
Dash lines (⋯): Ridge



- The variance of ridge regression is smaller
- The **bias** is about the same for both
- Hence the MSE of ridge regression is smaller

Lasso vs. Ridge regularization

- **Takeaways:** Neither ridge nor the lasso universally dominates
  - Lasso performs better if **a small number of predictors with large coefficients**
  - Ridge performs better if **many predictors with similar coefficients**
  - Select which one by **cross-validation** ☺

Lecture plan

- Comparison between ridge regression and lasso
- Elastic net
- Lab session

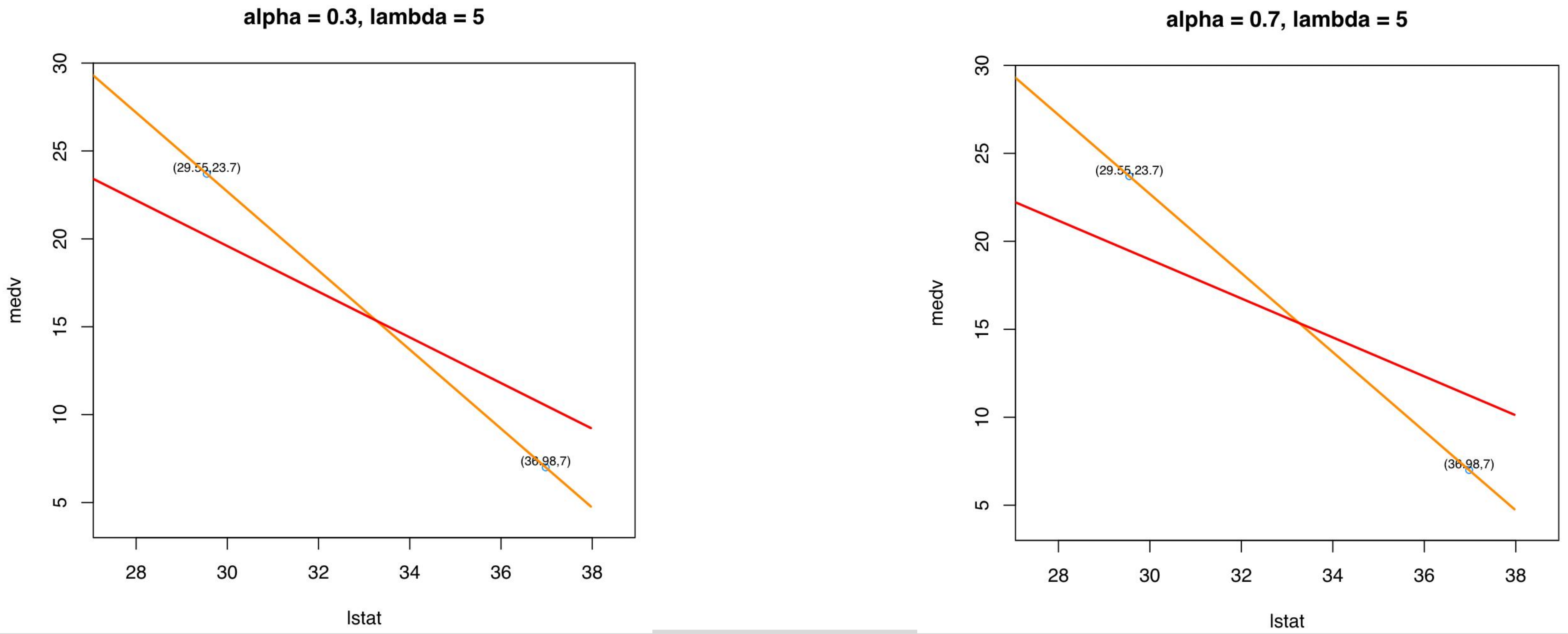
EMORY

Elastic net

- Elastic net combines lasso and ridge penalty, and minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot (1-\alpha) \cdot \frac{\beta_1^2}{2} + \lambda \cdot \alpha \cdot |\beta_1|$
  - $\lambda \geq 0$: tuning hyper-parameter
  - $\alpha \in [0,1]$: tuning hyper-parameter
    - $\alpha = 0$: ridge
    - $\alpha = 1$: lasso

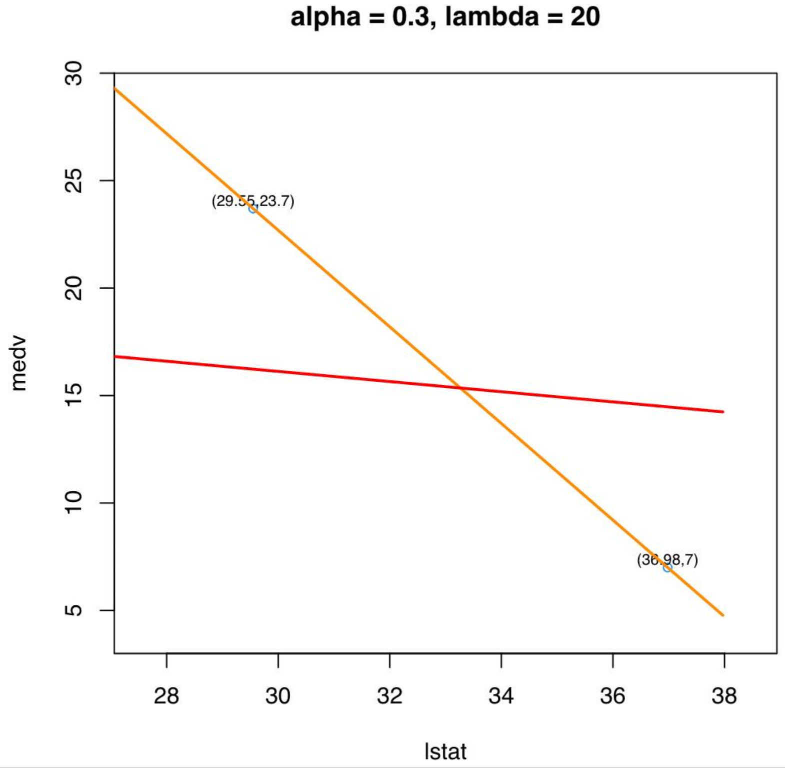
Role of α and λ in elastic net

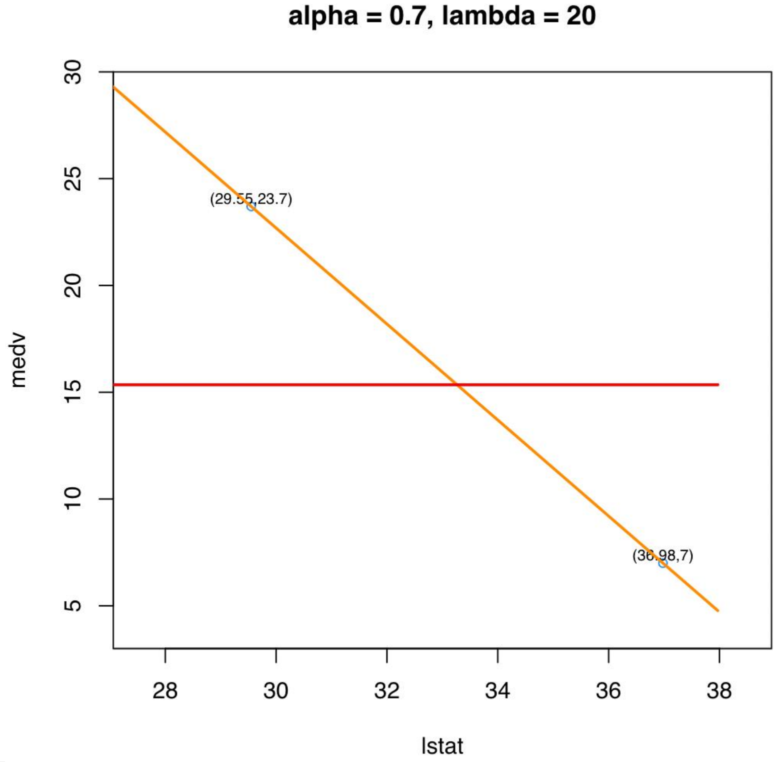
- Elastic net combines lasso and ridge penalty, and minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot (1-\alpha) \cdot \frac{\beta_1^2}{2} + \lambda \cdot \alpha \cdot |\beta_1|$
  - $\alpha = 0.3, \lambda = 5$: $\hat{\beta}_1^E = -1.299$; $\alpha = 0.7, \lambda = 5$: $\hat{\beta}_1^E = -1.107$



Role of α and λ in elastic net

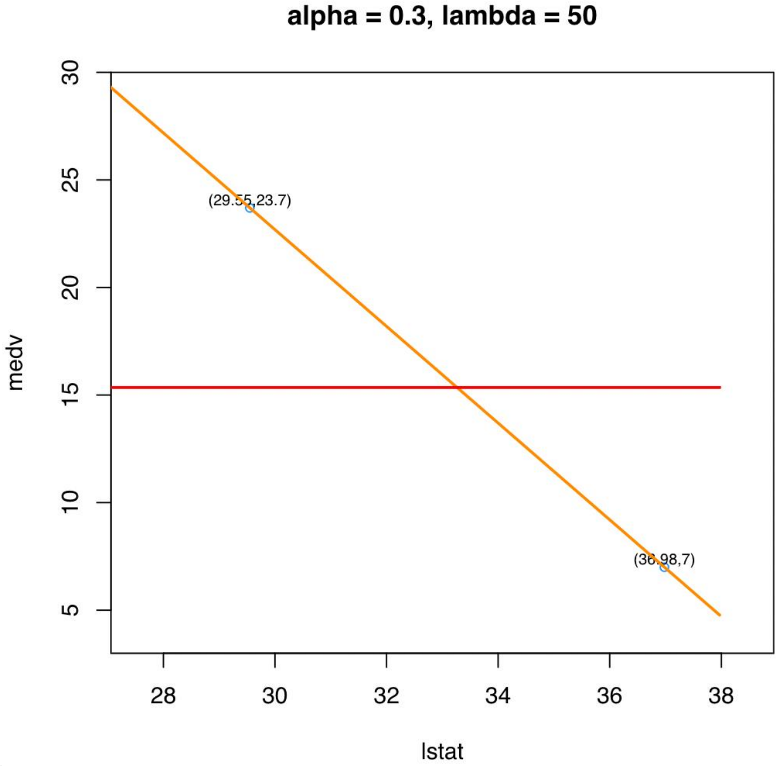
- Elastic net combines lasso and ridge penalty, and minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot (1-\alpha) \cdot \frac{\beta_1^2}{2} + \lambda \cdot \alpha \cdot |\beta_1|$
  - $\alpha = 0.3, \lambda = 20$: $\hat{\beta}_1^E = -0.236$; $\alpha = 0.7, \lambda = 20$: $\hat{\beta}_1^E = 0$

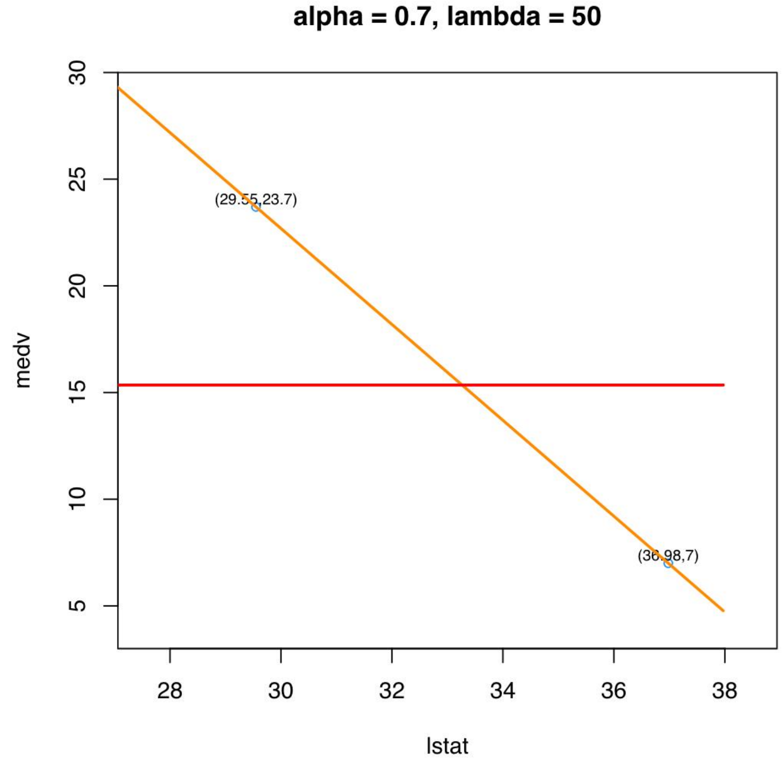




Role of α and λ in elastic net

- Elastic net combines lasso and ridge penalty, and minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot (1-\alpha) \cdot \frac{\beta_1^2}{2} + \lambda \cdot \alpha \cdot |\beta_1|$
  - $\alpha = 0.3, \lambda = 50$: $\hat{\beta}_1^E = 0$; $\alpha = 0.7, \lambda = 50$: $\hat{\beta}_1^E = 0$





Choose α and λ by cross-validation

- The procedure is the same for ridge and lasso

1. Choose a grid of α values and a grid of λ values
2. Compute the cross-validation error for each (α, λ) value
3. Select the (α, λ) with the smallest cross-validation error
4. Refit the model using all observations and selected (α, λ)