QTM 347 Machine Learning

Lecture 11: Lasso and elastic net

Ruoxuan Xiong

Suggested reading: ISL Chapter 6

EMORY

Lecture plan

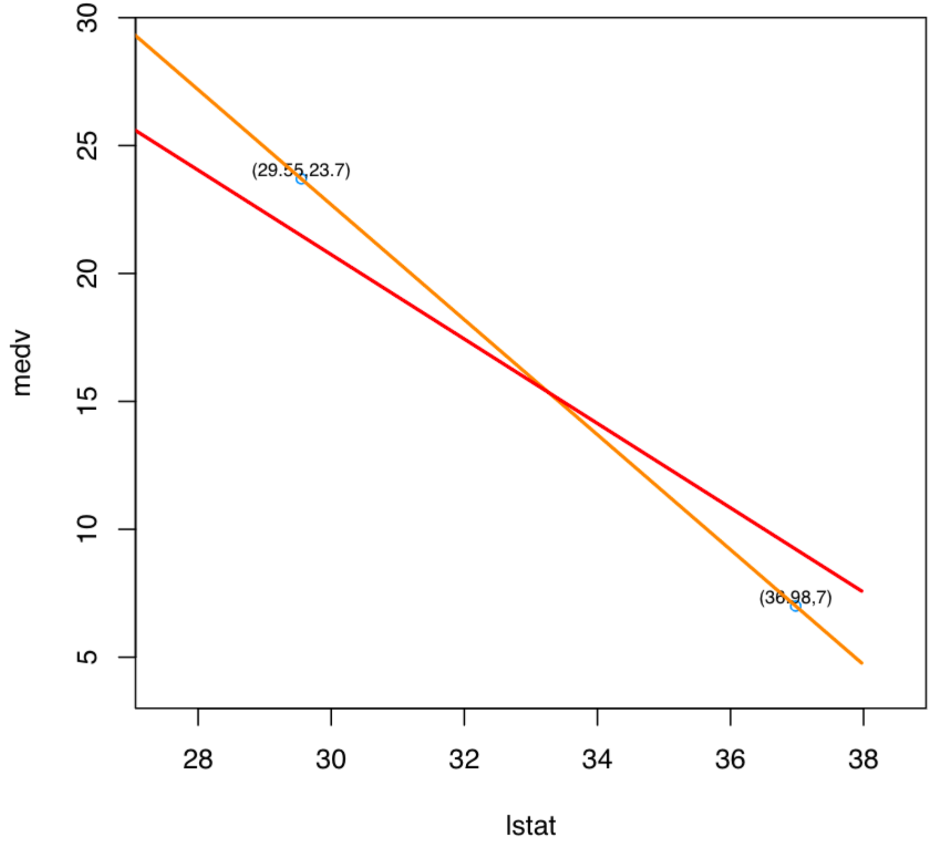
- Lasso
- Elastic net

Ridge regression

- Linear regression minimizes residual sum of squares
  - $RSS = \sum_{i=1}^{n}(medv_i - \beta_0 - lstat \cdot \beta_1)^2$

- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$

  - $\lambda \geq 0$: tuning hyper-parameter



EMORY

Ridge regression for more than one predictor

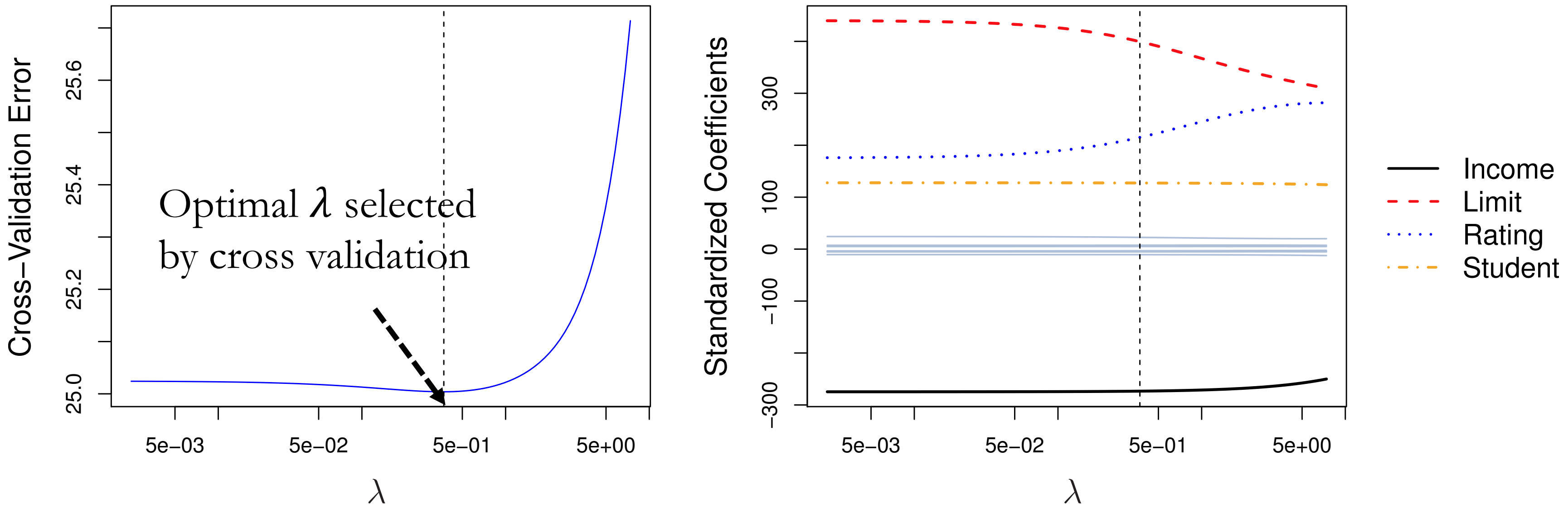
- Ridge regression minimizes

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

  - $X_{i,j}$: 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

Example: Credit card data set (ridge regression)

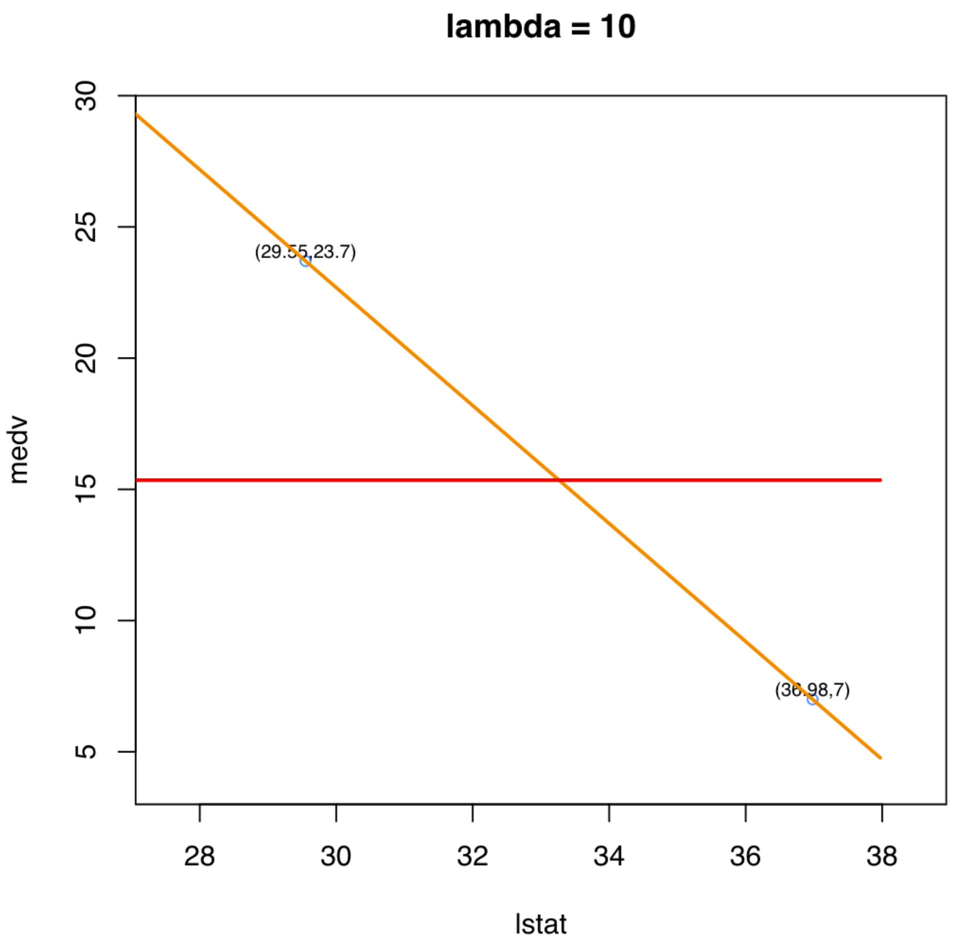
- Cross validation to choose the optimal λ



Lasso

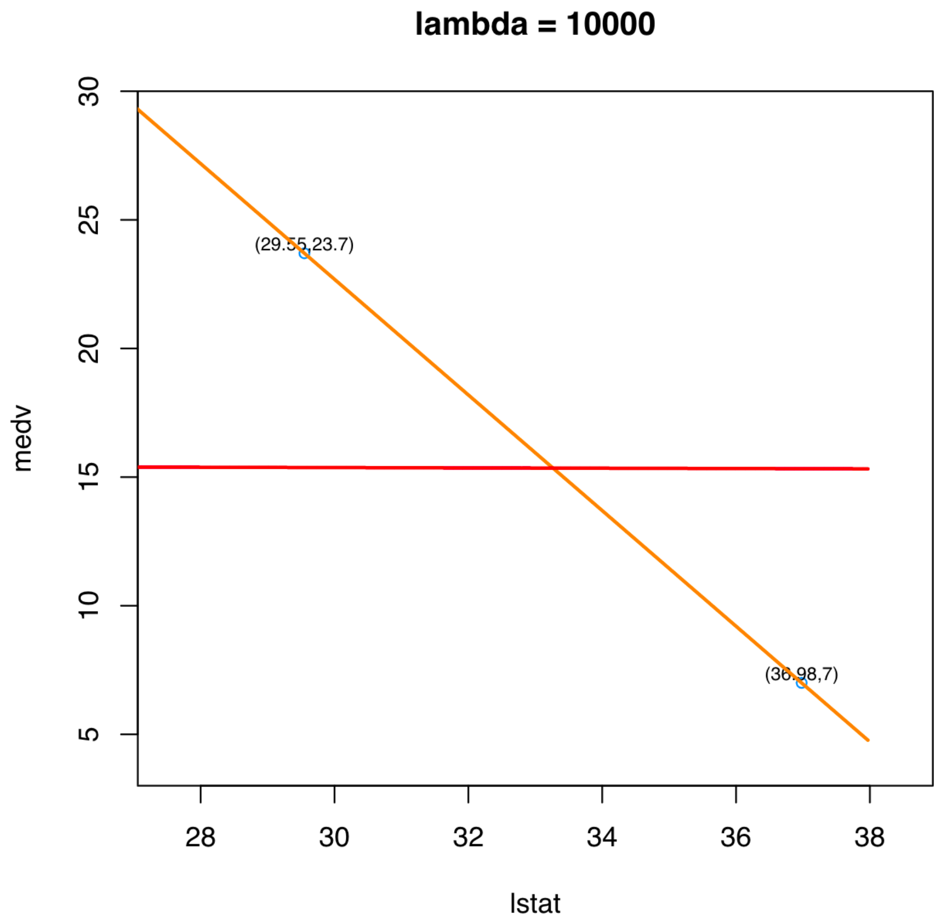
- Lasso: least absolute shrinkage and selection operator

- Lasso minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot |\beta_1|$
  - $\lambda \geq 0$: tuning hyper-parameter



Motivation

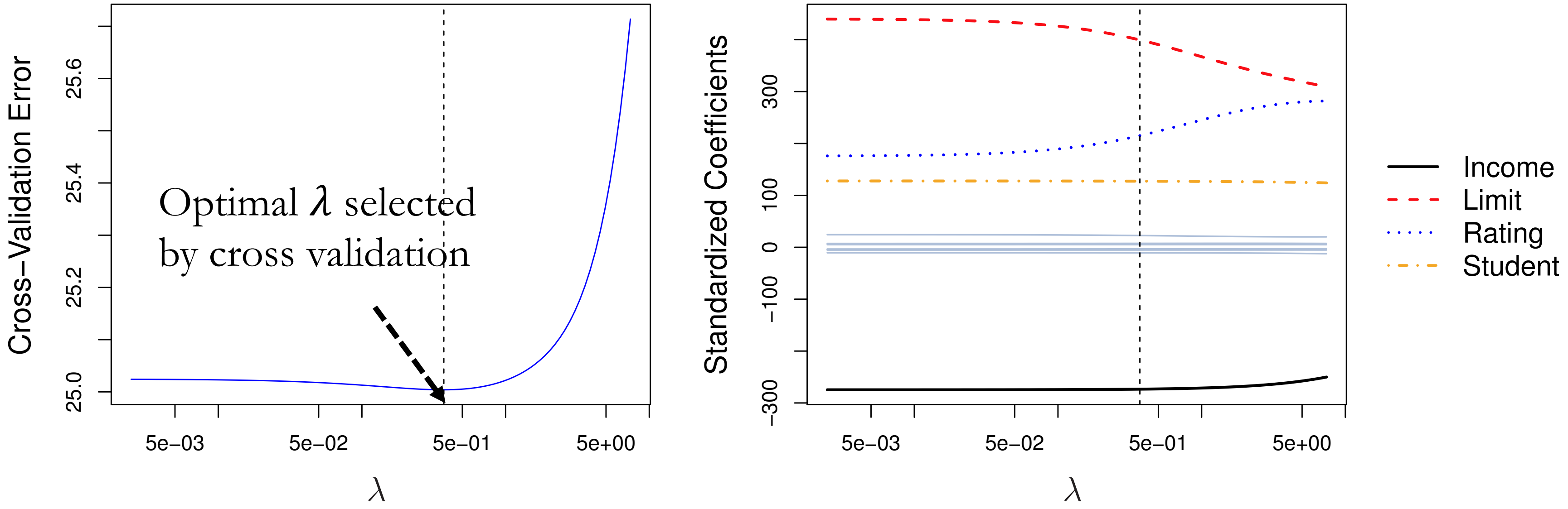
- Ridge regression shrinks coefficients to approximately zero, but not exactly zero
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - When $\lambda = 10{,}000$, $\hat{\beta}_1^R = -0.0062$



What if we want to exclude useless variables?

- In the credit data set, the standardized ridge coefficients for variables other than income, limit, rating, and student are nonzero

- What if we want to perform variable selection?



Lasso

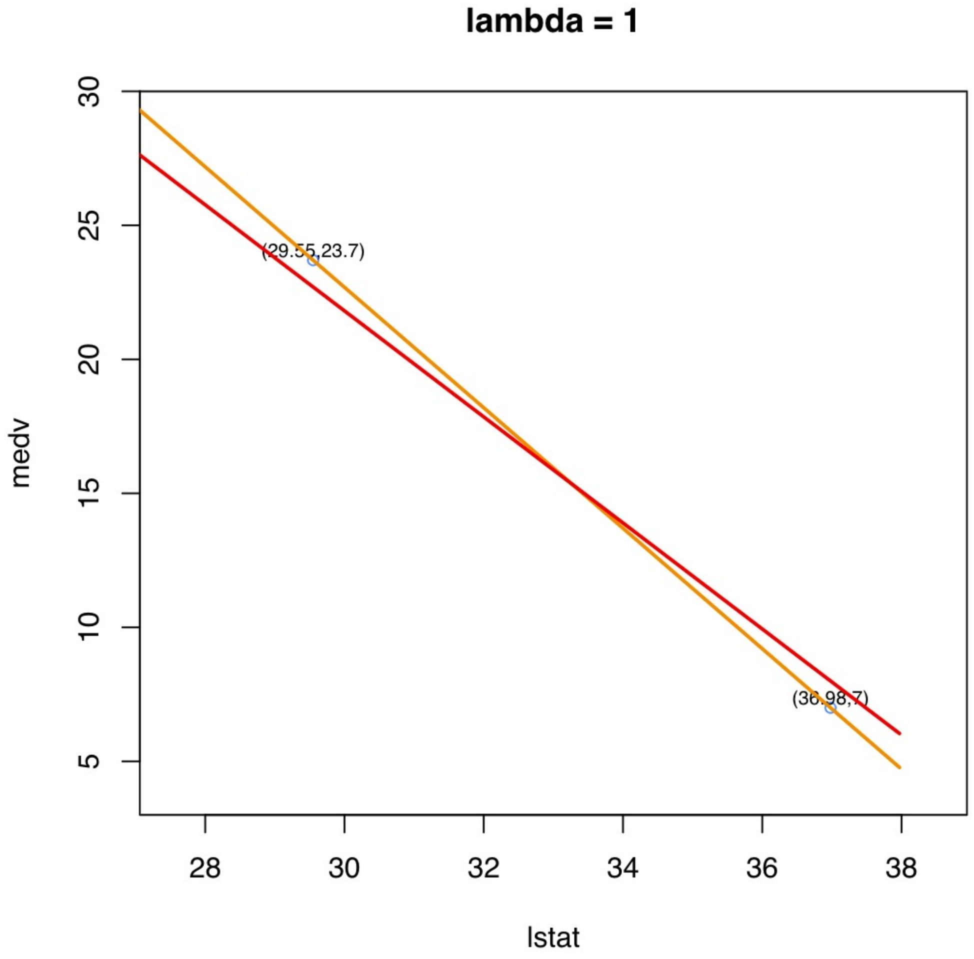
- Lasso: least absolute shrinkage and selection operator

- Lasso minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot |\beta_1|$

  - $\lambda \geq 0$: tuning hyper-parameter

Role of λ in Lasso

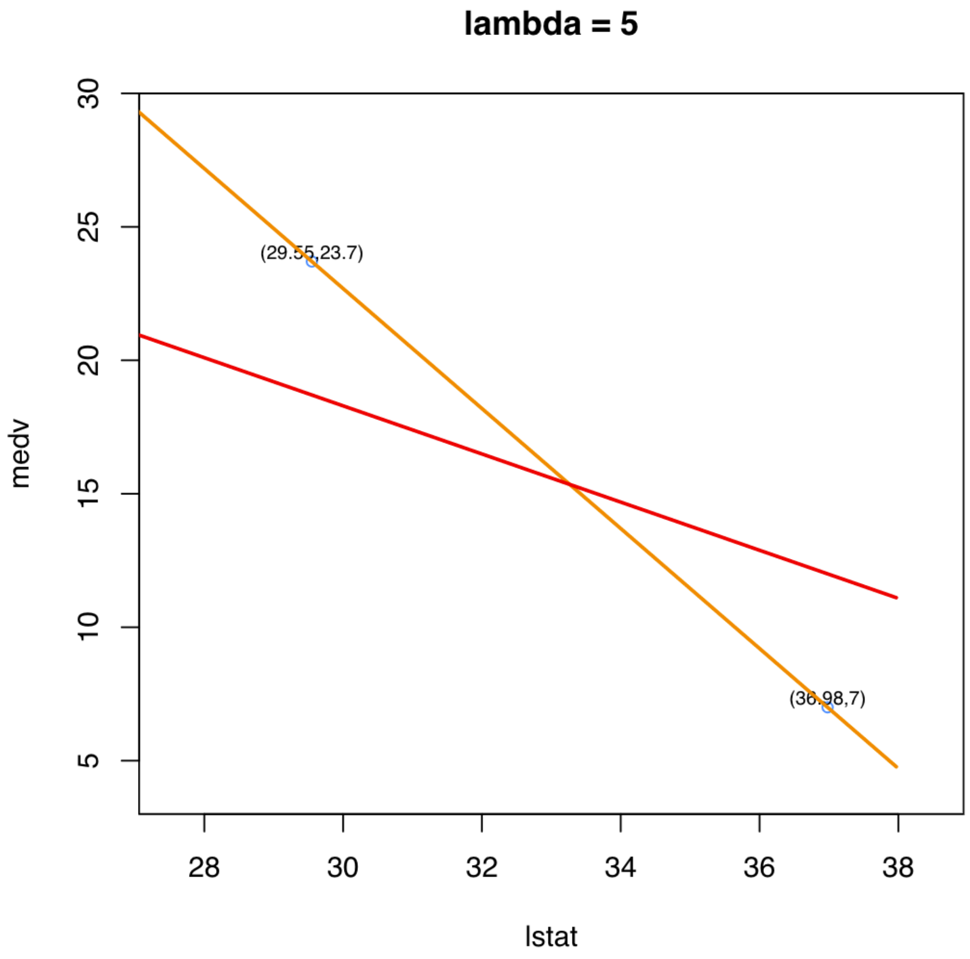
- Lasso minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot |\beta_1|$
  - $\lambda = 1 : \hat{\beta}_1^L = -1.978$



Role of λ in Lasso

- Lasso minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot |\beta_1|$

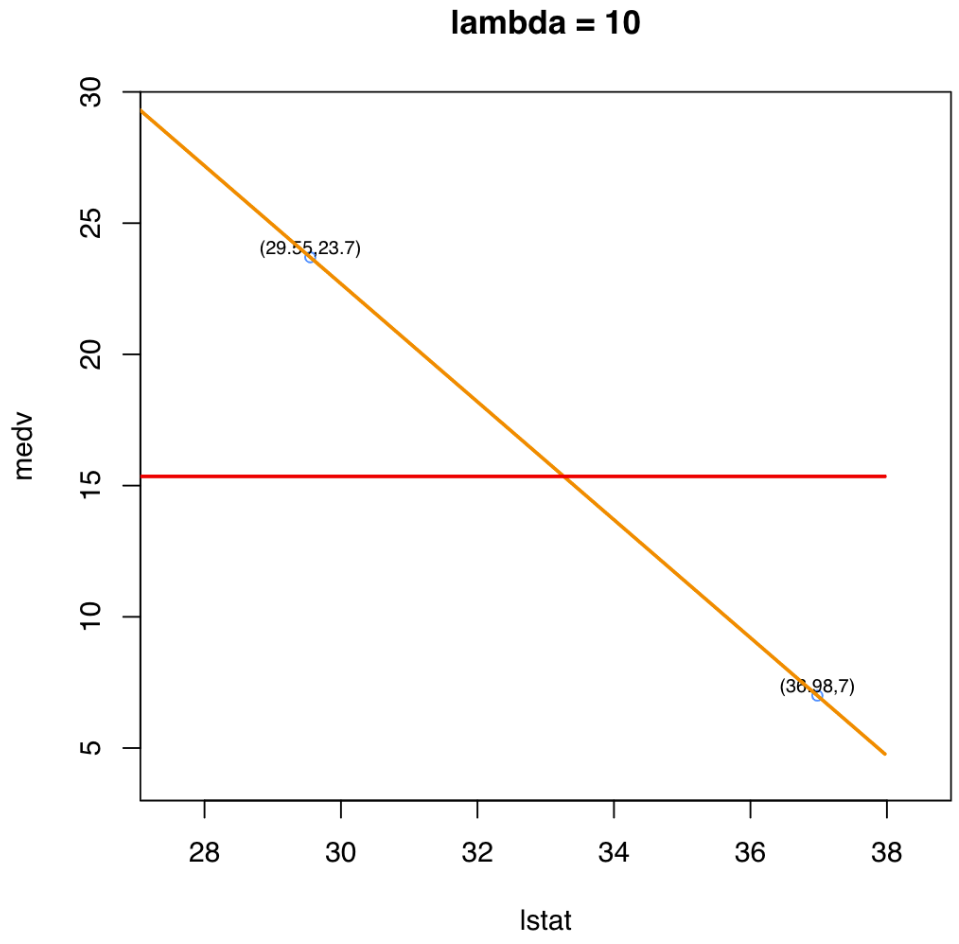
  - $\lambda = 5 : \hat{\beta}_1^L = -0.902$



Role of λ in Lasso

- Lasso minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot |\beta_1|$

- $\lambda = 10 : \hat{\beta}_1^L = 0$



Lasso for more than one predictor

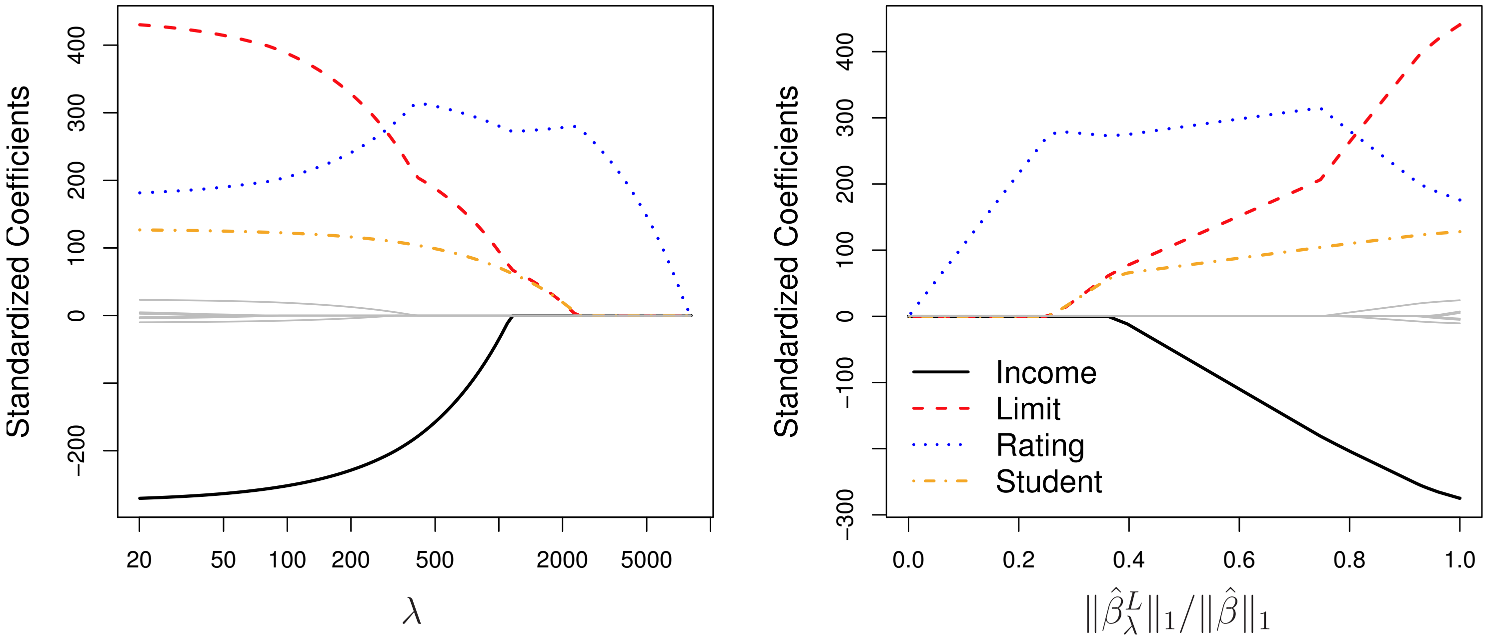
- Lasso minimizes

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

  - $X_{i,j}$: j-th predictor of i-th observation
  - $\|\beta\|_1 = \sum_{j=1}^{p}\left|\beta_j\right|$: $\|\beta\|_1$ 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

Example: Credit card data set (lasso)

- Predict default or not; **11** predictors
  - $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

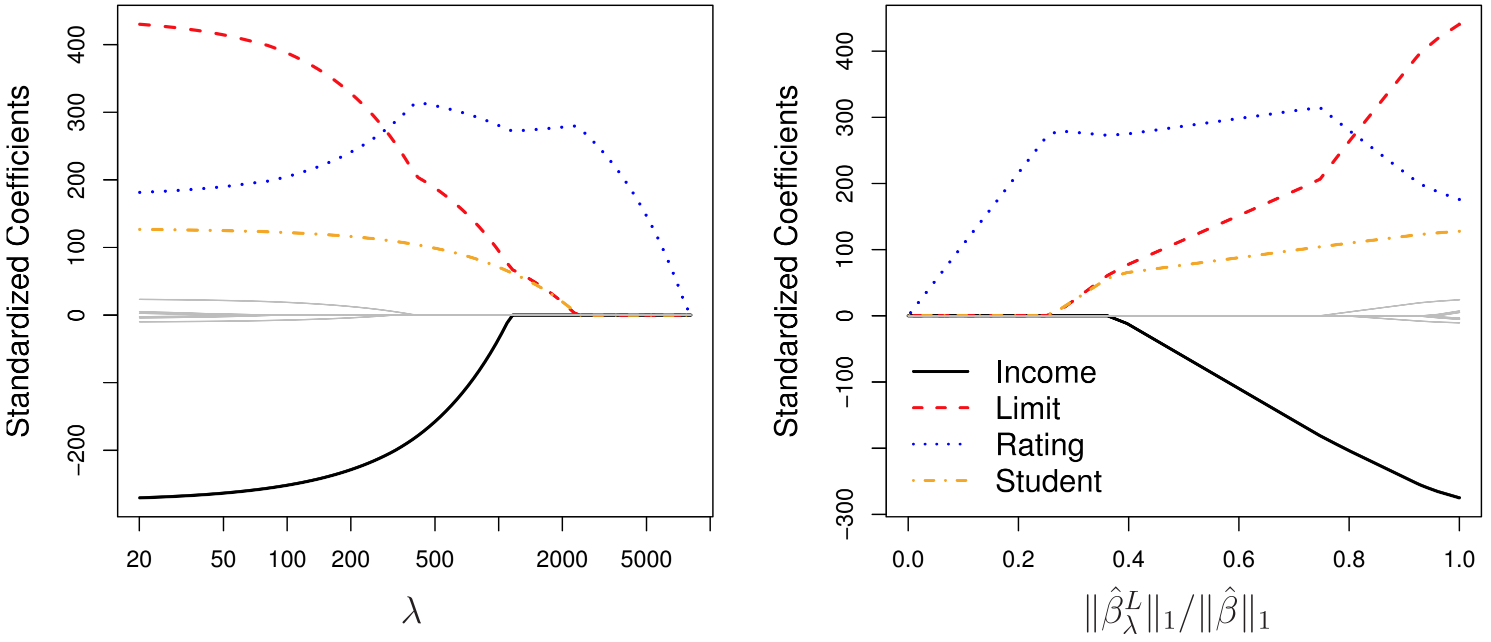


Shrinkage ratio

- **Shrinkage ratios**: coefficients shrink to zero at varying rates

Example: Credit card data set (lasso)

- Predict default or not; **11** predictors



- **Variable selection**: As λ increases, lasso selects less variables
  - {"empty"} → {rating} → {limit, rating, student} → {income, limit, rating, student}
- **Lasso path**: Different coefficient values by varying λ

Choose λ by cross-validation

- The procedure is the same for ridge and lasso

1. Choose a grid of λ values

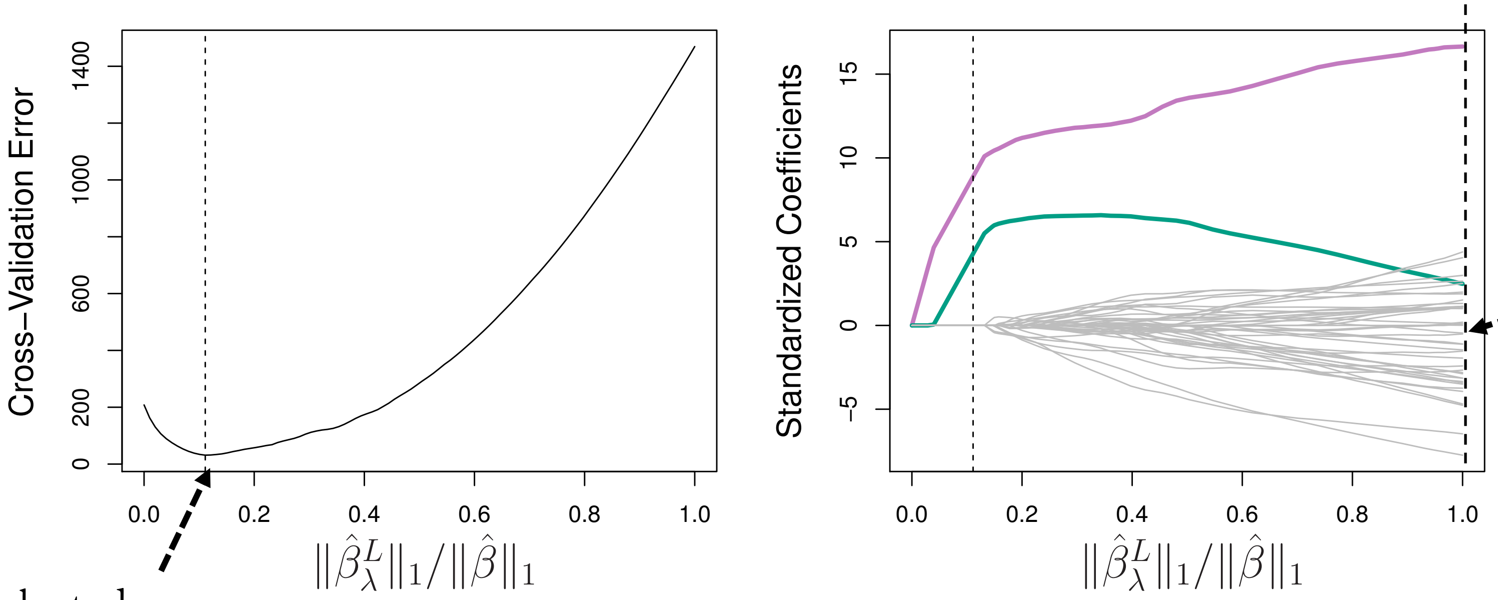
2. Compute the cross-validation error for each λ value

3. Select the λ with the smallest cross-validation error

4. Refit the model using all observations and selected λ

Example

- **Simulation I:** Only 2 coefficients are non-zero
  - Simulated data: 45 predictors, 2 out of $\beta_1, \ldots, \beta_{45}$ are nonzero
  - **10-fold CV to select the lasso regularization parameter**



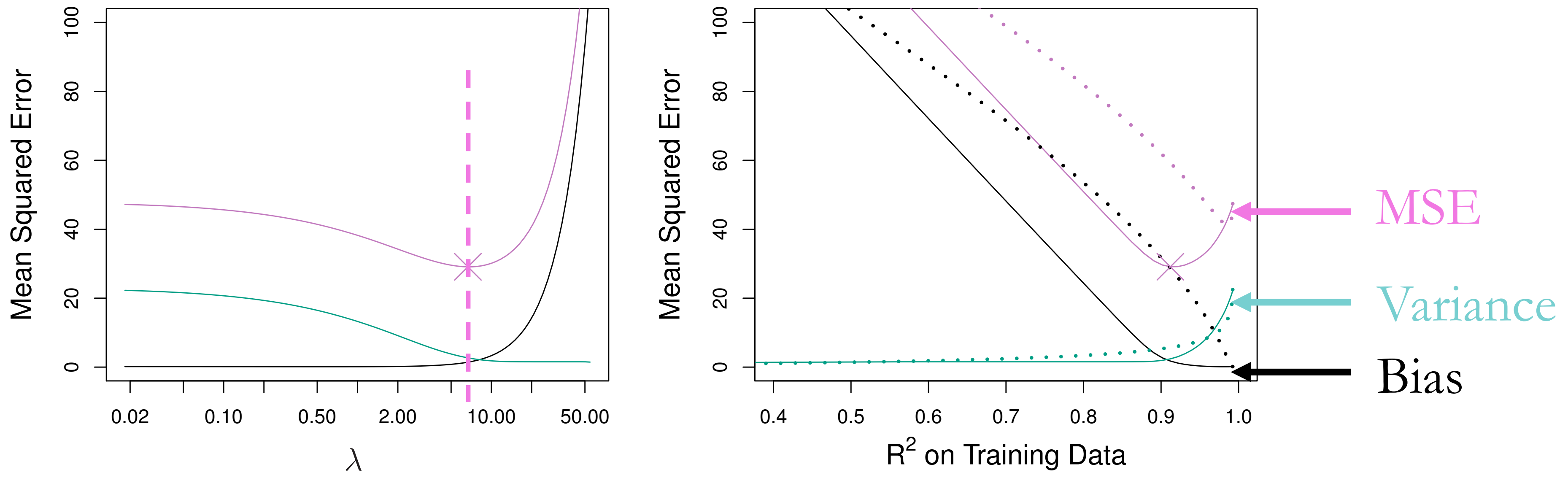
Linear regression solution

Optimal λ selected by cross-validation

Lasso vs. Ridge regularization

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

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

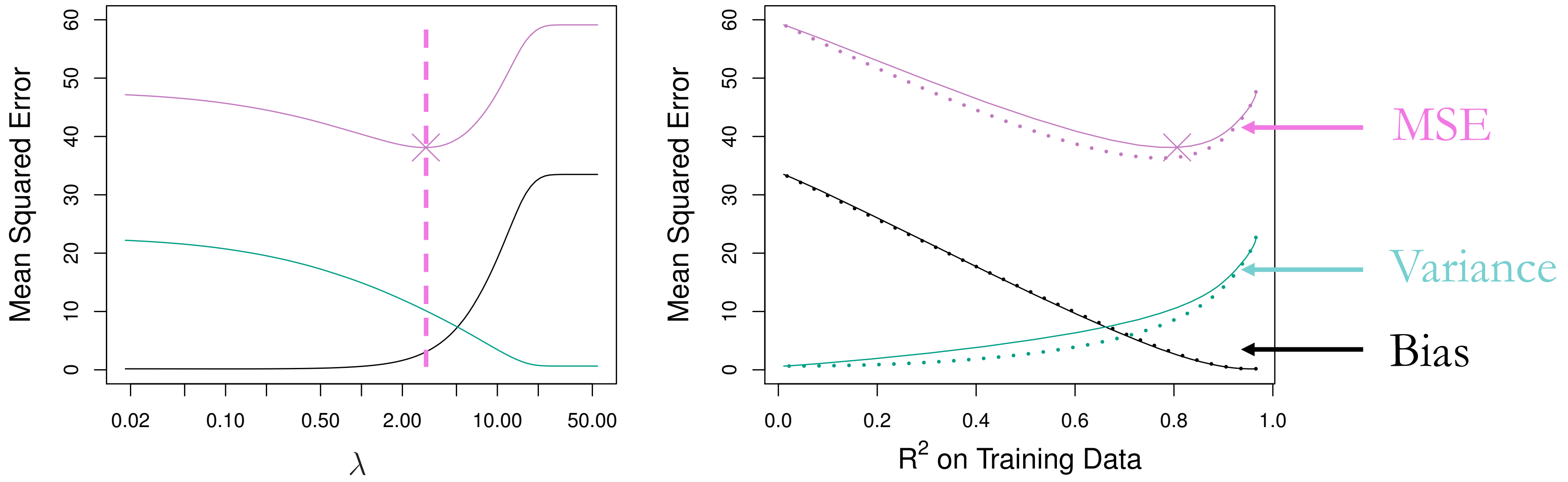


- The **bias**, variance, and MSE are all lower for the lasso

Lasso vs. Ridge regularization

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

- Lasso

- Elastic net

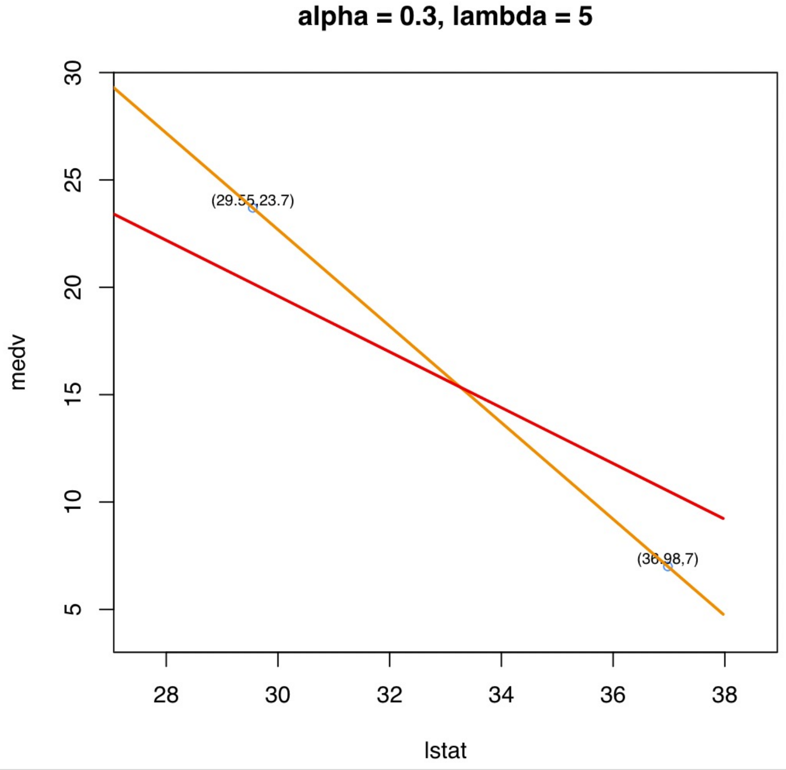
EMORY

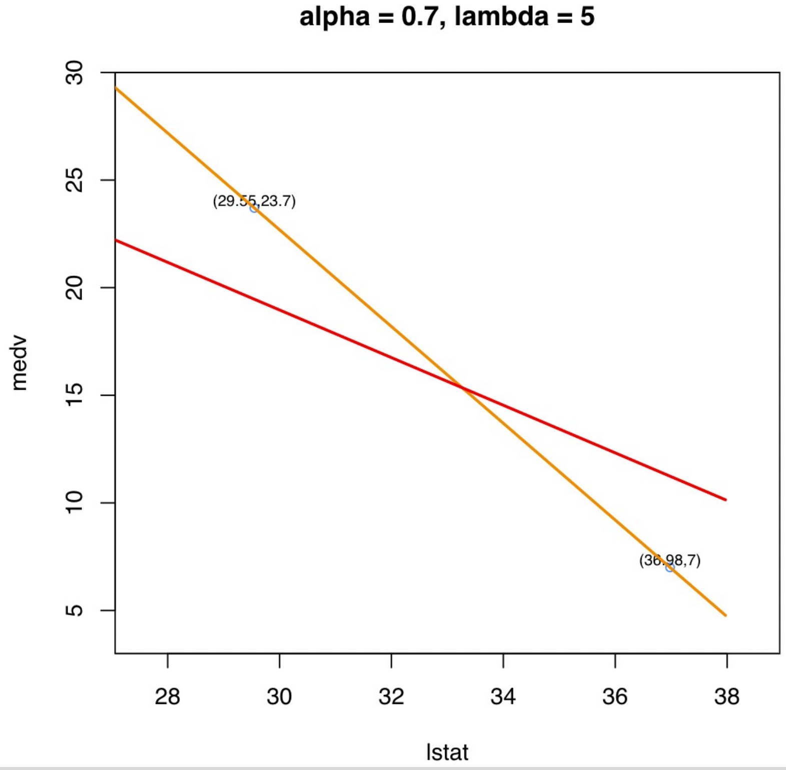
Elastic net

- 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

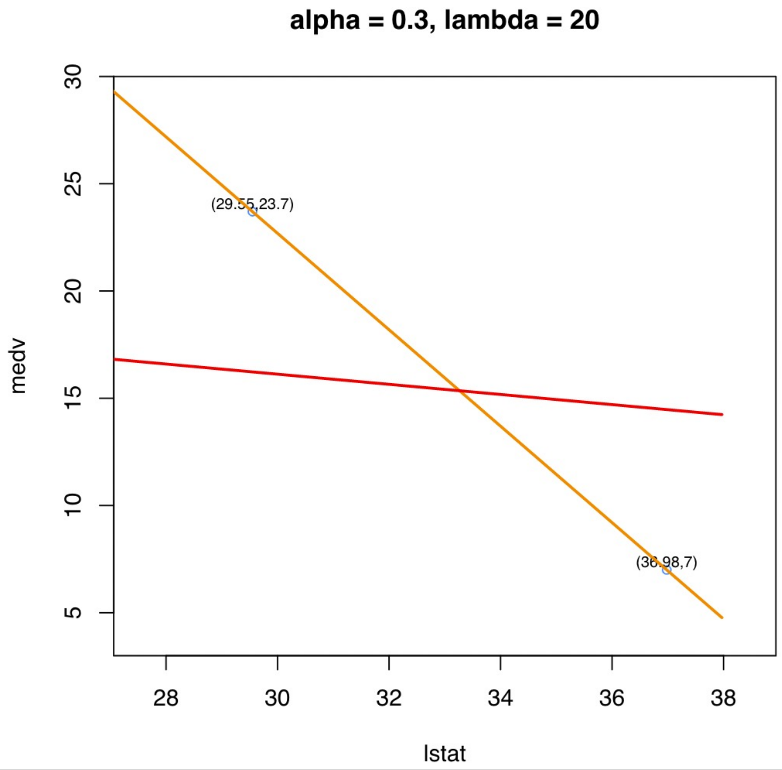
- Elastic net combines lasso and ridge penalty, and minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot (1-\alpha) \cdot \frac{\beta_1^2}{2} + \lambda \cdot \alpha \cdot |\beta_1|$
  - $\alpha = 0.3, \lambda = 5$: $\hat{\beta}_1^E = -1.299$; $\alpha = 0.7, \lambda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

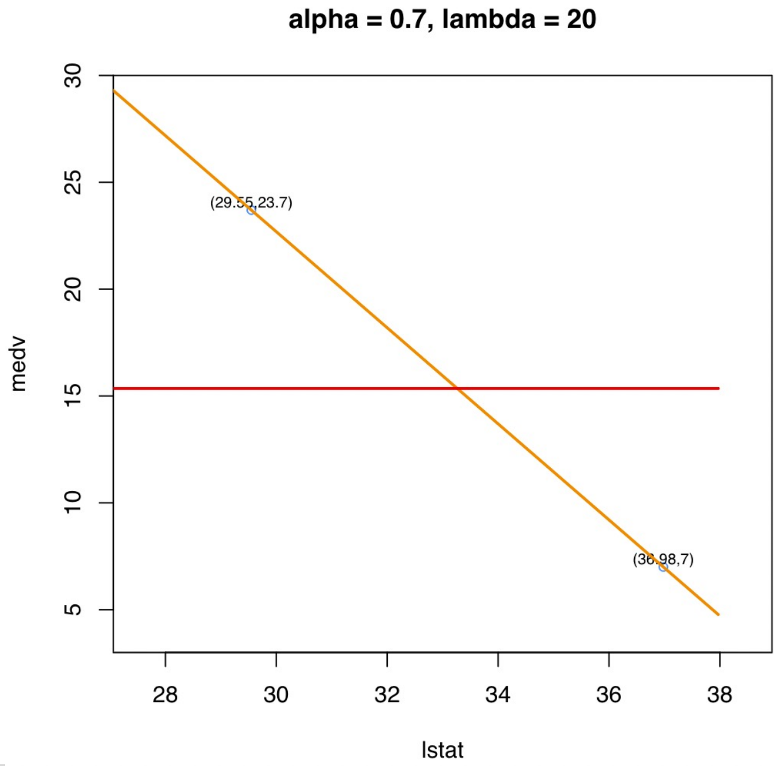




Role of α and λ in elastic net

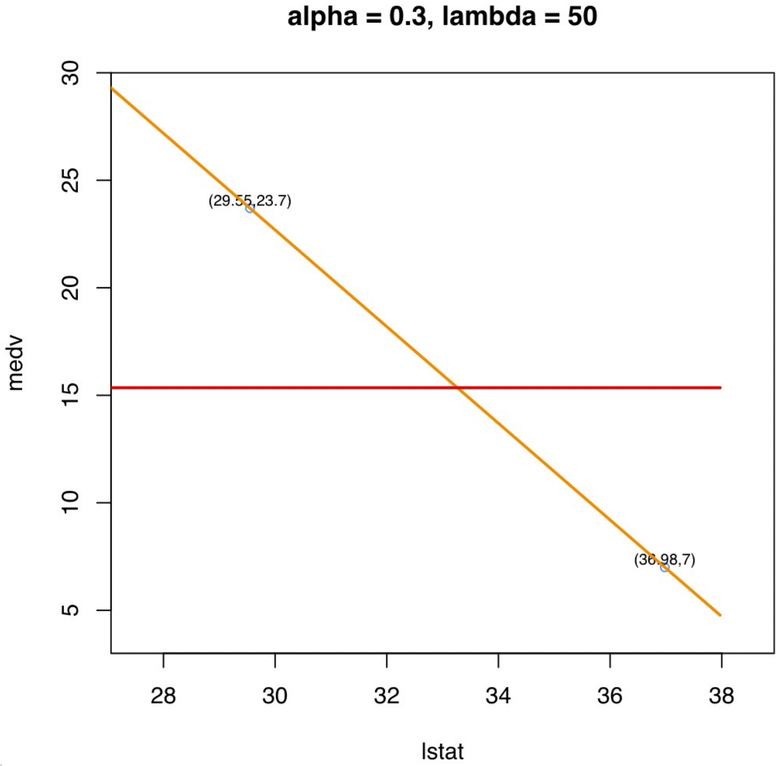
- Elastic net combines lasso and ridge penalty, and minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot (1-\alpha) \cdot \frac{\beta_1^2}{2} + \lambda \cdot \alpha \cdot |\beta_1|$
  - $\alpha = 0.3, \lambda = 20$: $\hat{\beta}_1^E = -0.236$; $\alpha = 0.7, \lambda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

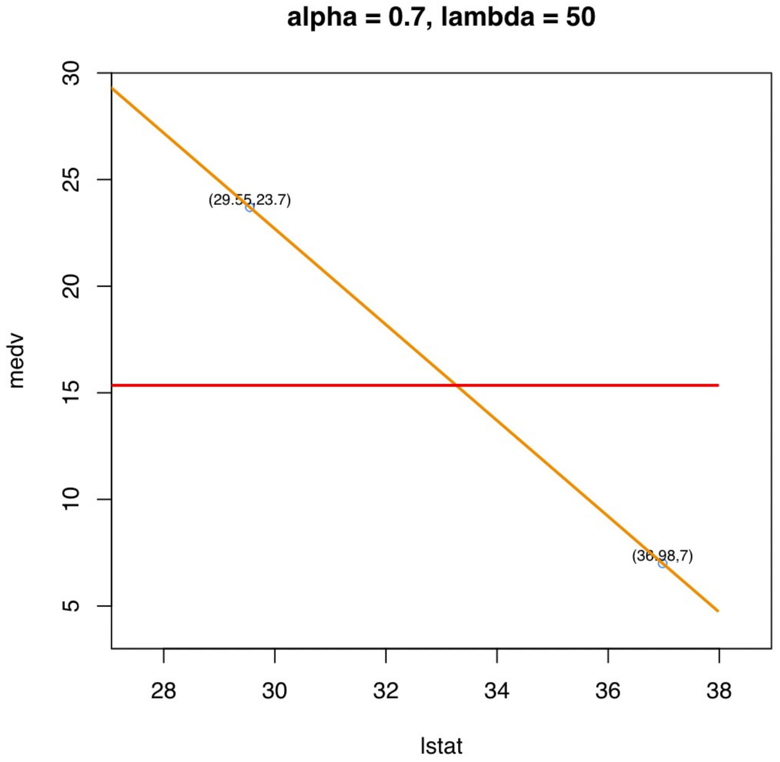




Role of α and λ in elastic net

- 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)