Supervised Machine Learning and Learning Theory

Lecture 7: Regularization

September 27, 2024

Warm up questions

- What is the purpose of the bootstrap procedure? And how does it work?
- When do we expect to perform subset selection?
- Could you write down the logistic regression model for classifying handwritten digits?

Example

- Credit card dataset: Predict whether customers default on their credit card debt
- Predictors (11 predictors in total)
    - Income: Income in $1,000's
    - Limit: Credit limit
    - Rating: Credit rating
    - Cards: Number of credit cards
    - Age: Age in years
    - Education: Number of years of education
    - Gender: A factor with levels Male and Female
    - Student: A factor with levels No and Yes indicating the individual was a student
    - Married: A factor with levels No and Yes indicating whether the individual was married
    - Ethnicity: A factor with levels African American, Asian, and Caucasian indicating the individual's ethnicity
    - Balance: Average credit card balance in $

Stepwise selection methods

- Forward stepwise selection
  - Start with a model with no predictors
  - Add predictors to the model one-at-a-time
- Backward stepwise selection
  - Start with a model with p predictors
  - Remove the least useful predictor one-at-a-time

Forward stepwise selection

- Fit at most $1 + p + (p-1) + \cdots + 1 = 1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$ models in total
- Much fewer than $\binom{p}{k}$ (best subset selection)

Forward selection vs. best subset

- Forward stepwise selection may fail to select the best k-variable subset

# 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.*

Lecture plan

- Ridge regression

Motivation

Linear model: $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + \varepsilon$

- Suppose the number of predictors $p > n$ (e.g., this happens a lot in bioinformatics, such as gene expressions): we have more parameters than observations

- How can we estimate β?

Example

- Predict Boston house prices: Suppose we only have one observation ($n = 1$)

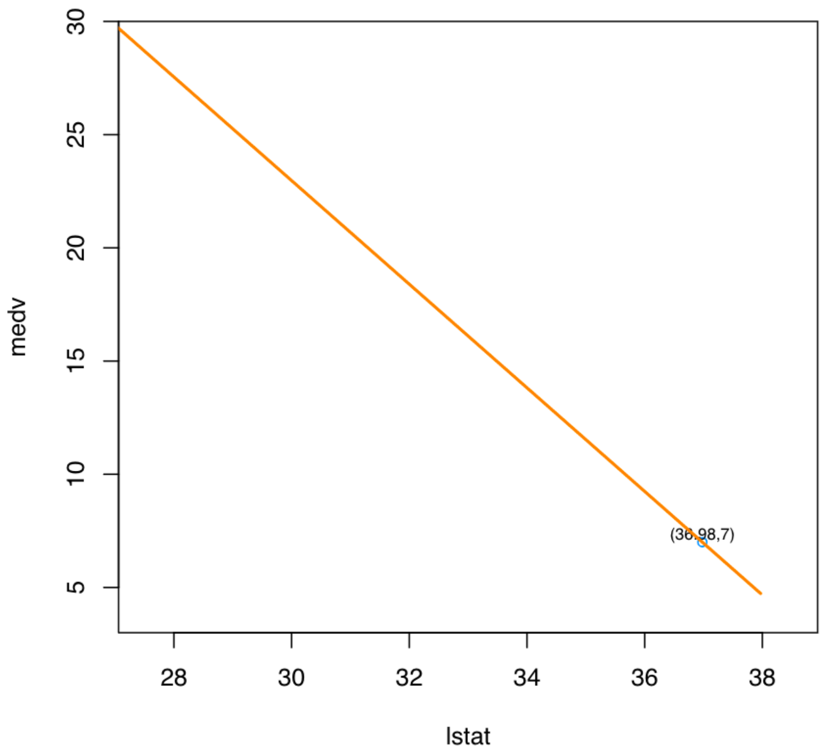
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
45.7461	0	18.1	0	0.693	4.519	100	1.6582	24	666	20.2	36.98	7

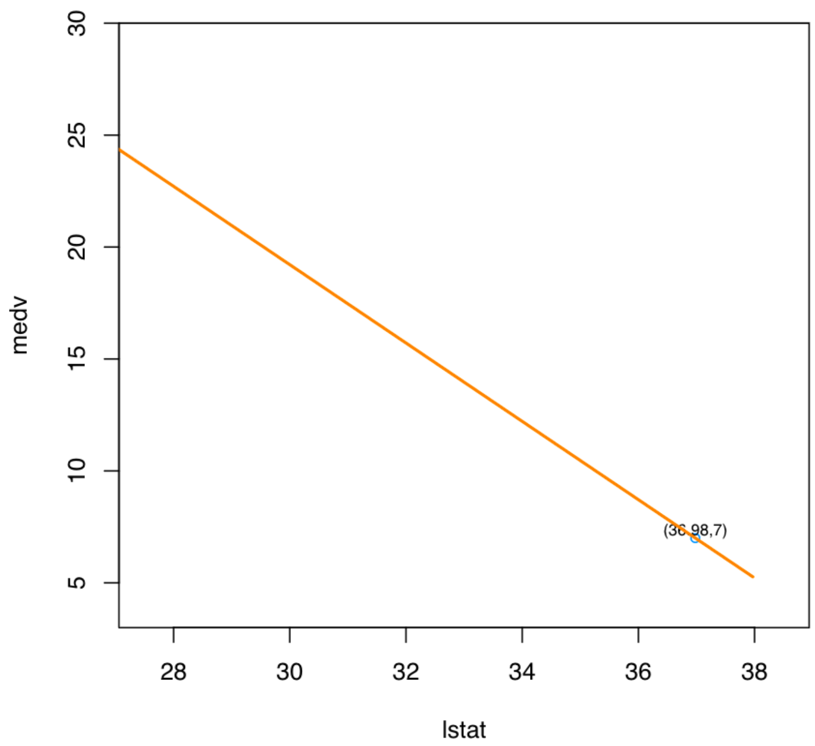
- Suppose we want to estimate the coefficients in simple linear regression:

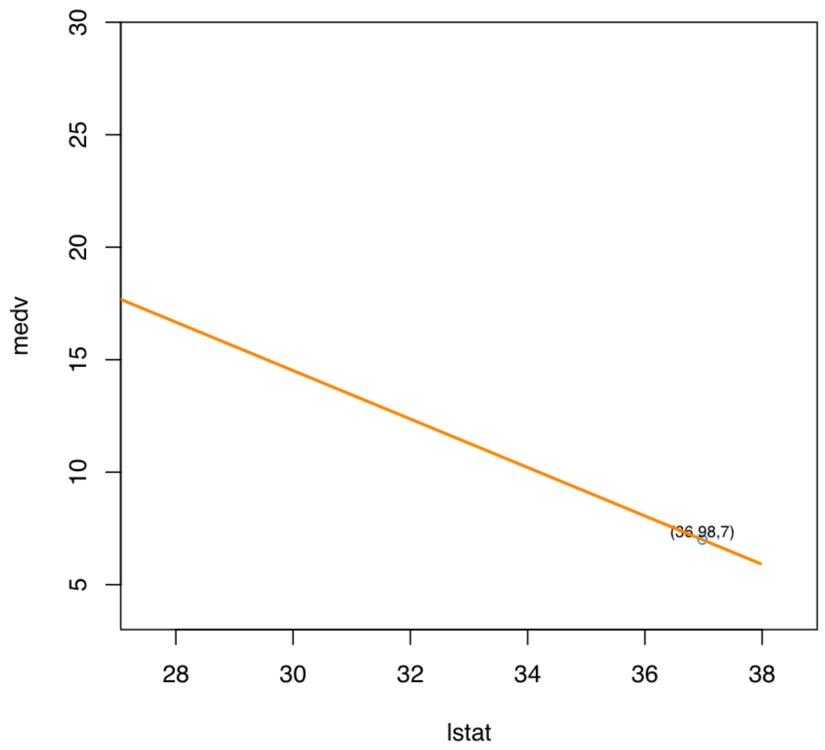
$$medv = \beta_0 + lstat \cdot \beta_1 + \varepsilon$$

- How can we use one observation to estimate β_0, β_1?

Which β_0 and β_1 should we choose?







All of these are valid solutions!

If we have one more observation…

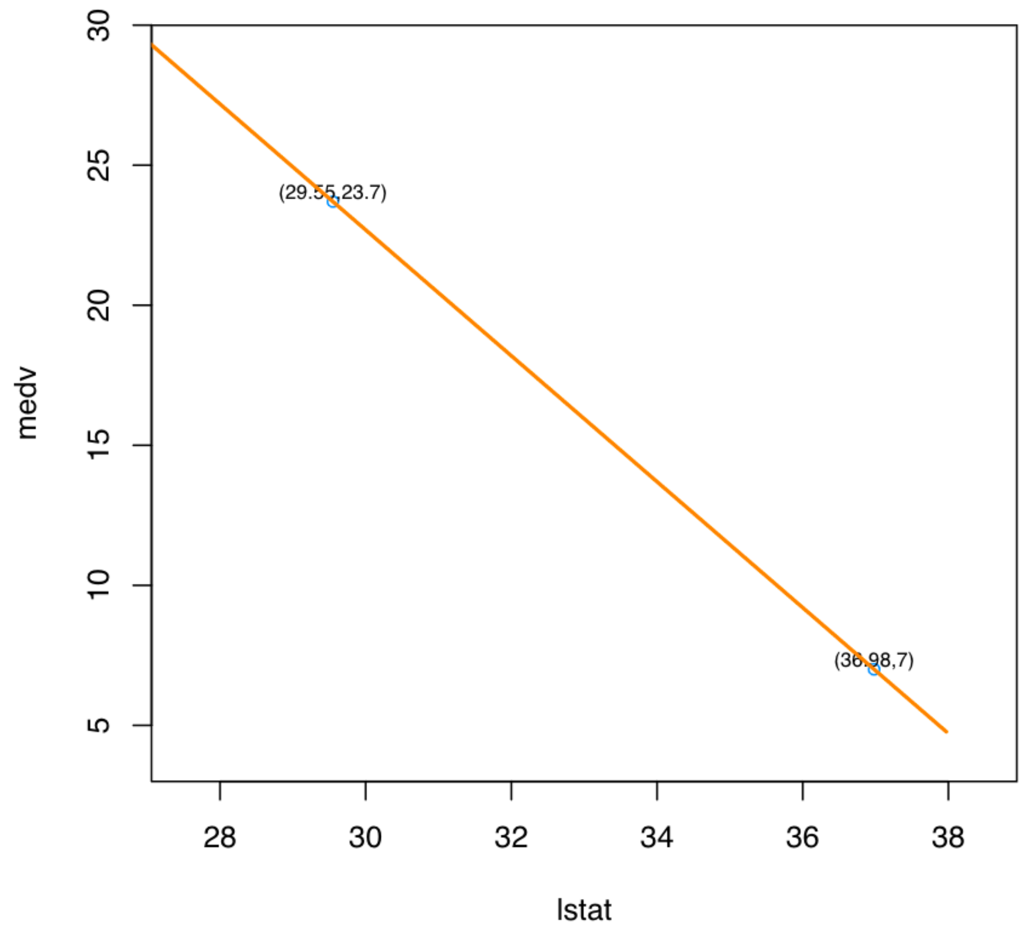
- Suppose we only have two observations ($n = 2$)

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
0.28955	0	10.59	0	0.489	5.412	9.8	3.5875	4	277	18.6	29.55	23.7
45.74610	0	18.10	0	0.693	4.519	100.0	1.6582	24	666	20.2	36.98	7.0

- Let us consider the same model: $medv = \beta_0 + lstat \cdot \beta_1 + \varepsilon$

- We can estimate β_0 and β_1 with two data points (solving a linear system)

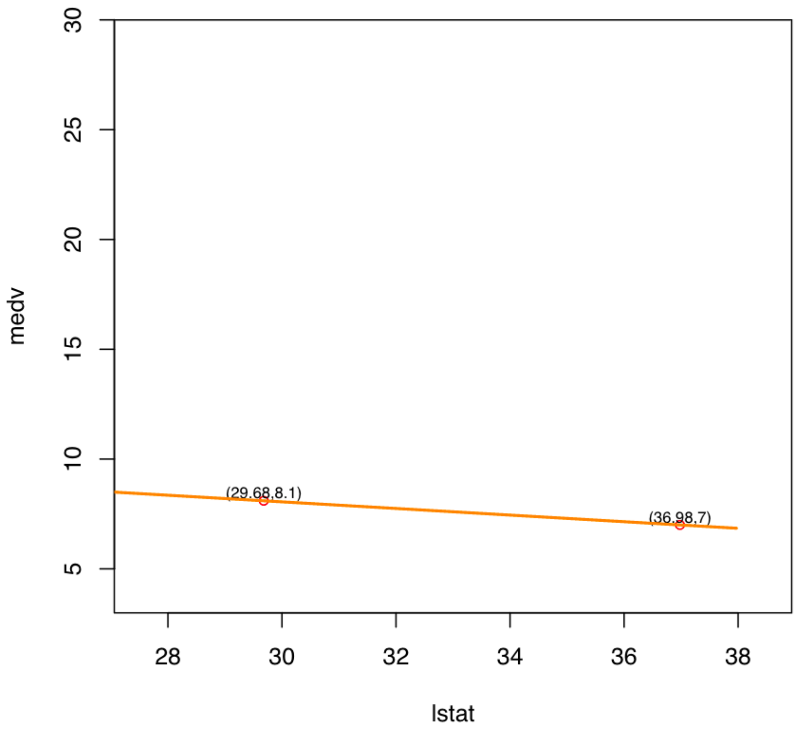
Example



- Problem: The fitted curve is sensitive to the $medv$ of these two observations

Example

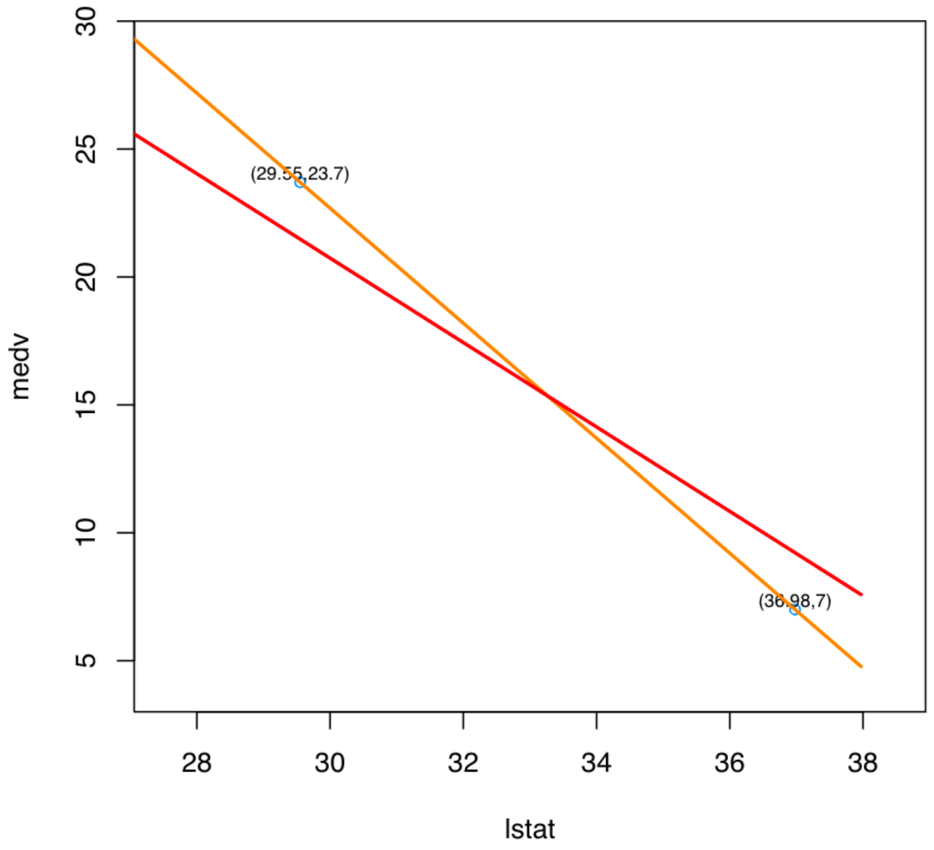
- If one of the two observations changes, we can get a very different fitted curve



- This is an example of overfitting…
- Question: can you think of other examples of overfitting?

Ridge regression

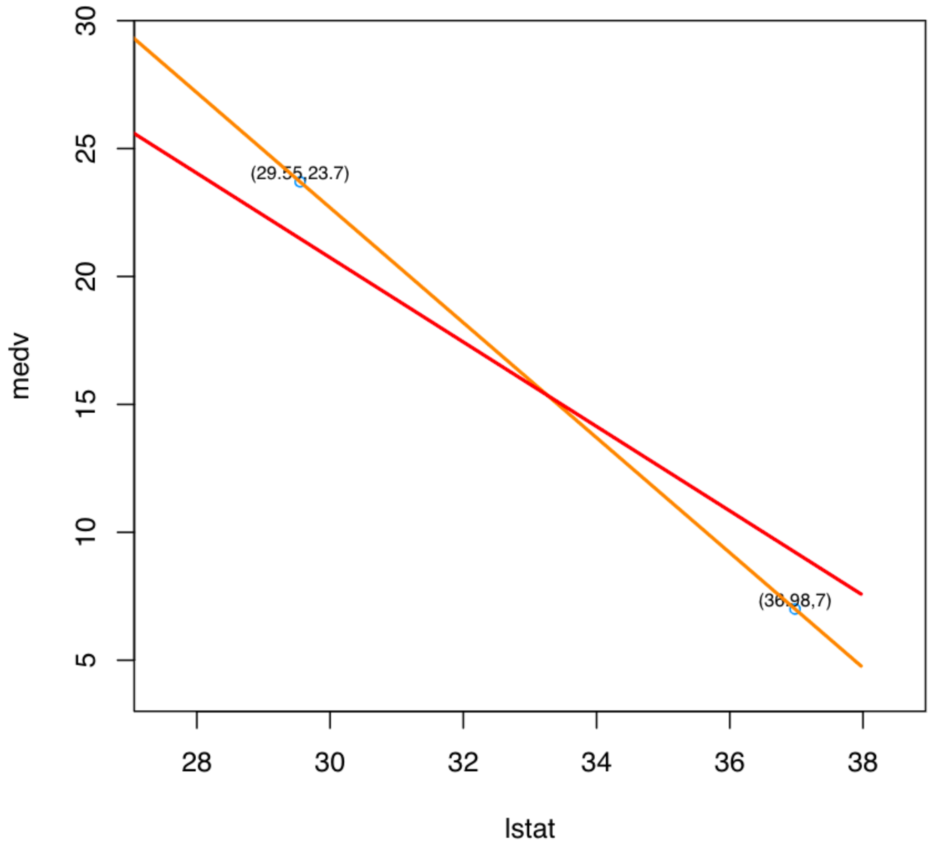
- Find **a new line** that **does not fit** the **training data perfectly**



- Introduce **a small amount of bias** into the fit to data

Ridge regression

- This can be achieved with Ridge regression: by adding **a small amount of bias,** we reduce variance (i.e., the fitted lines are less sensitive to changes with the input)

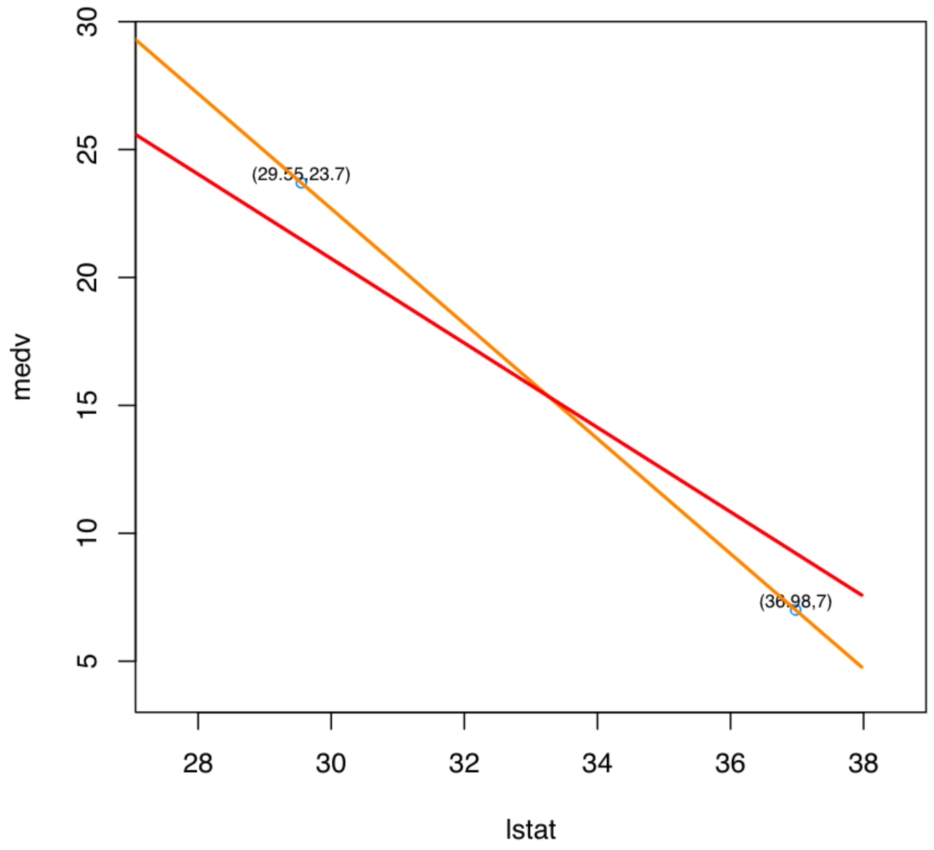


Fitting ridge regression

- Linear regression minimizes

$$MSE = \sum_{i=1}^{n} (medv_i - \beta_0 - lstat \cdot \beta_1)^2$$

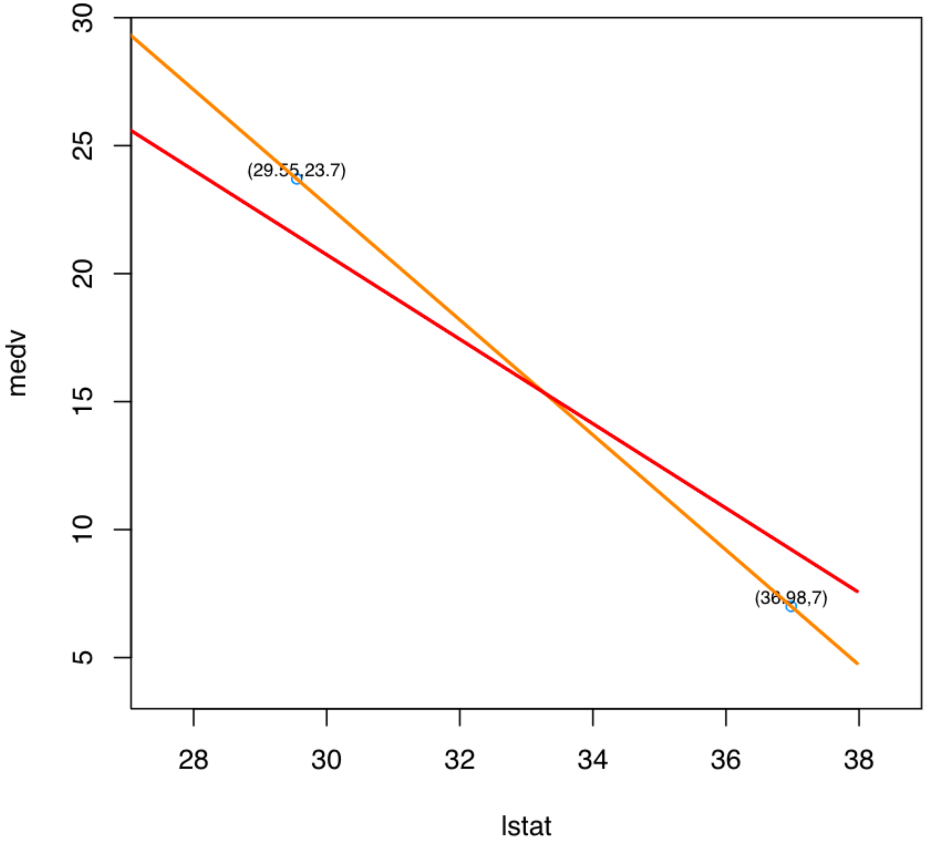
- Ridge regression minimizes
  - $\sum_{i=1}^{n} (medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - $\lambda \geq 0$: tuning hyper-parameter



Example

- Suppose $\lambda = 10$
- Linear regression fit: $\widehat{medv} = 90.118 - 2.248 \cdot lstat$

  - $\hat{\beta}_1 = -2.248$
  - $\sum_{i=1}^{n}\left(medv_i - \hat{\beta}_0 - lstat_i \cdot \hat{\beta}_1\right)^2 + \lambda \cdot \hat{\beta}_1^2$
  $= 0 + 10 \cdot 2.248^2 = 50.535$
  - Perfectly fitting the data incurs high loss

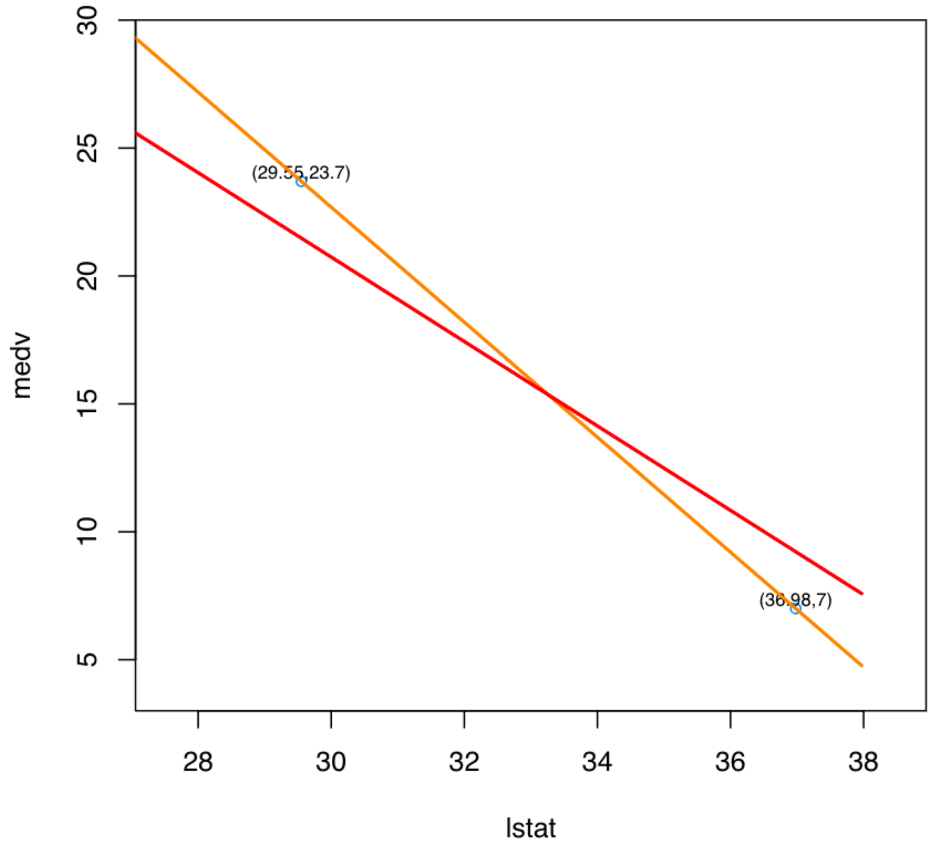


Ridge regression

- Suppose $\lambda = 10$
- Ridge regression fit: $\widehat{medv} = 70.234 - 1.650 \cdot lstat$

  - $\hat{\beta}_1^R = -1.650$

  - $\sum_{i=1}^{n}\left(medv_i - \hat{\beta}_0 - lstat_i \cdot \hat{\beta}_1^R\right)^2 + \lambda \cdot \left(\hat{\beta}_1^R\right)^2$
    $= 4.931 + 4.931 + 10 \cdot 1.650^2 = 37.084$
    < 50.535



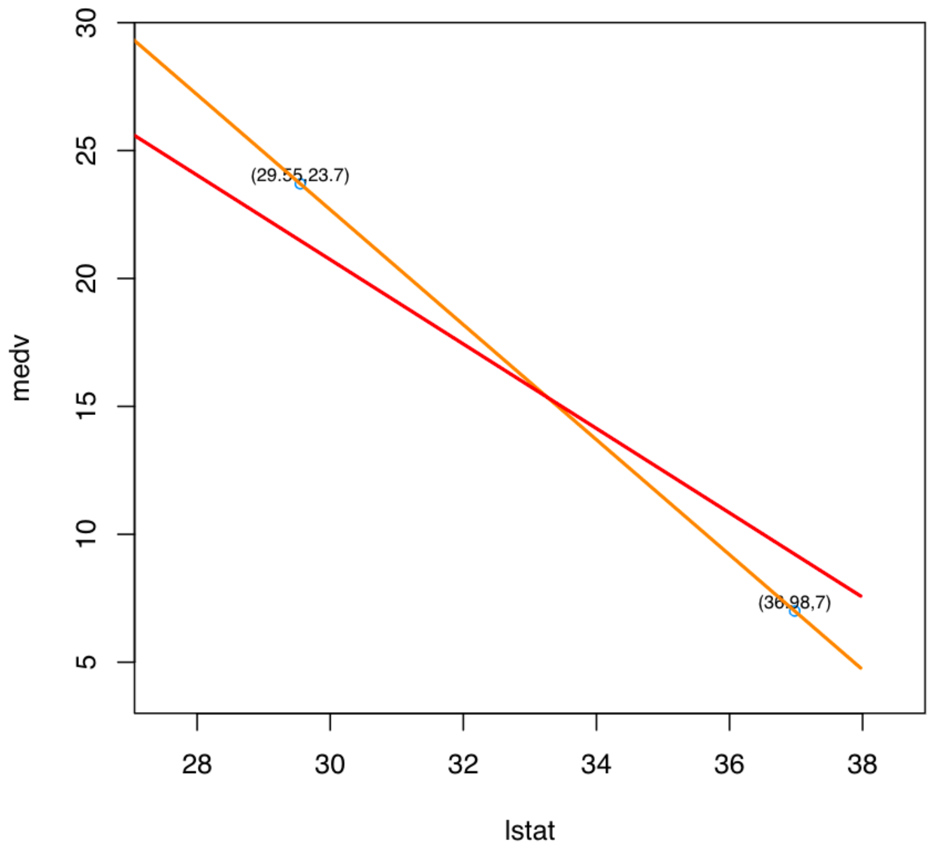
Ridge regression is less sensitive to $lstat$

- Linear regression fit: $\widehat{medv} = 90.118 - 2.248 \cdot lstat$

- One unit change in $lstat$ results in $-\ 2.248$ units change in $medv$

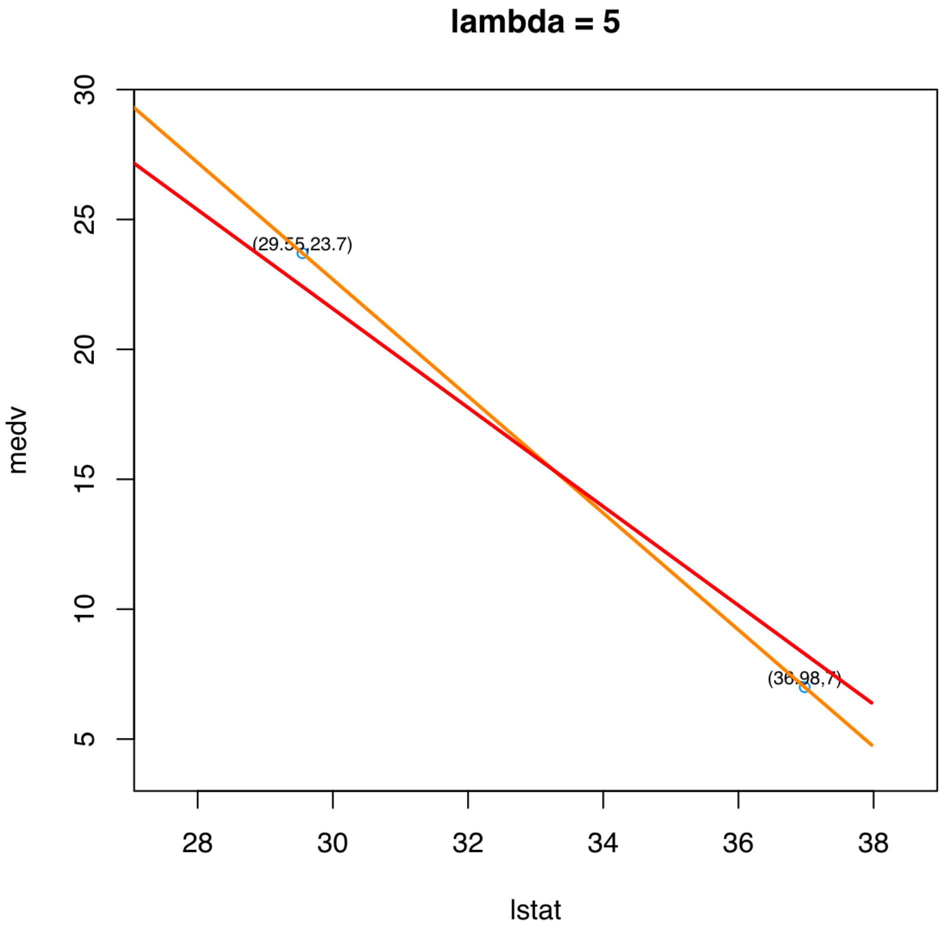
- Ridge regression fit: $\widehat{medv} = 70.234 - 1.650 \cdot lstat$

- One unit change in $lstat$ results in $-\ 1.650$ units change in $medv$



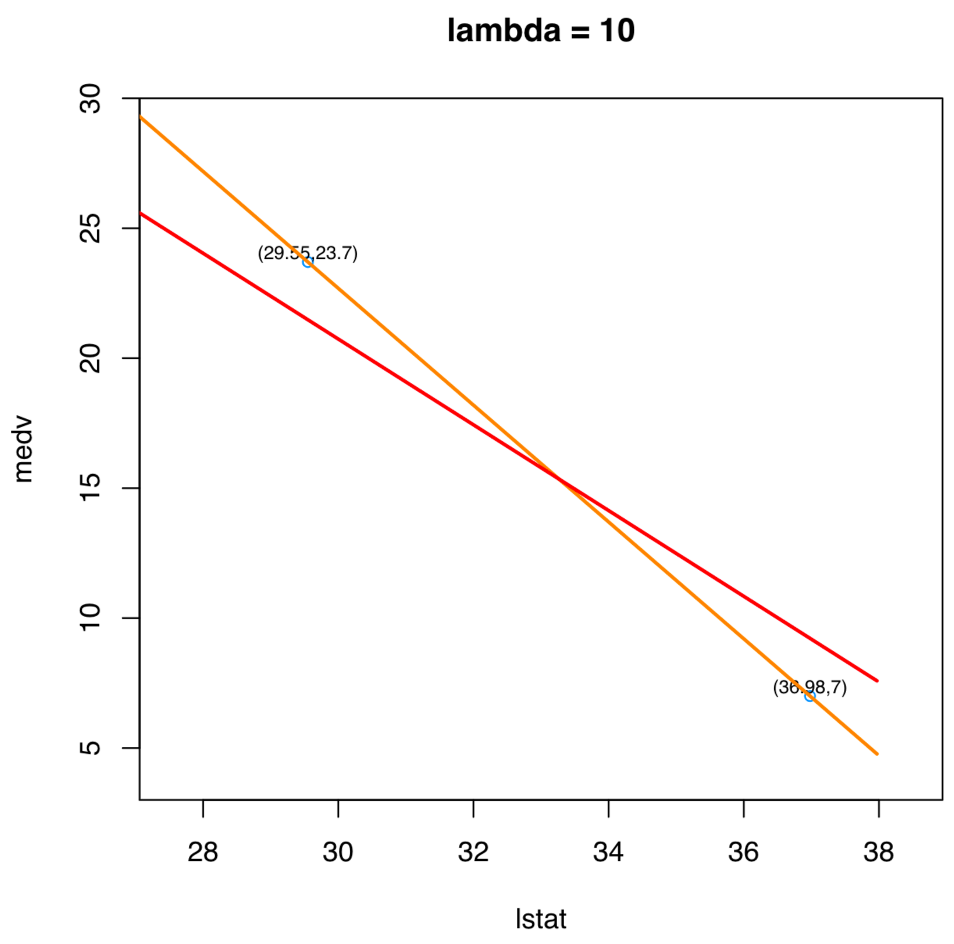
Role of λ in ridge regression

- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - $\lambda = 5$



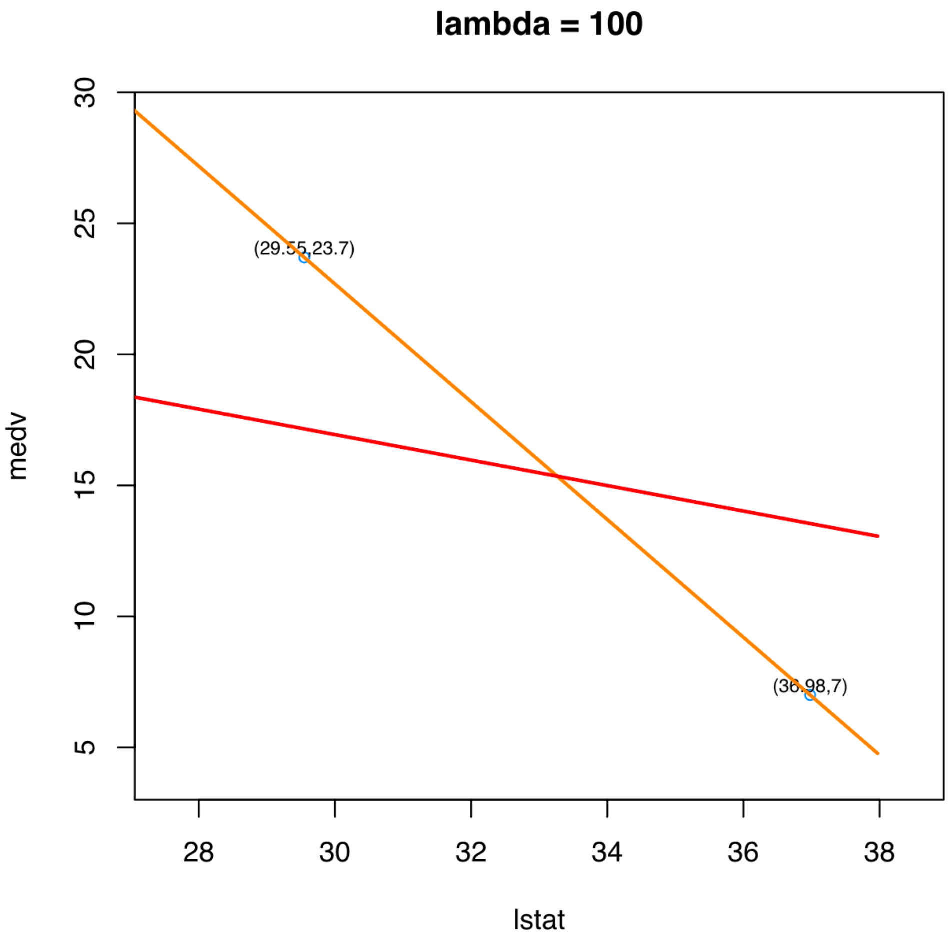
Role of λ in ridge regression

- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - $\lambda = 10$



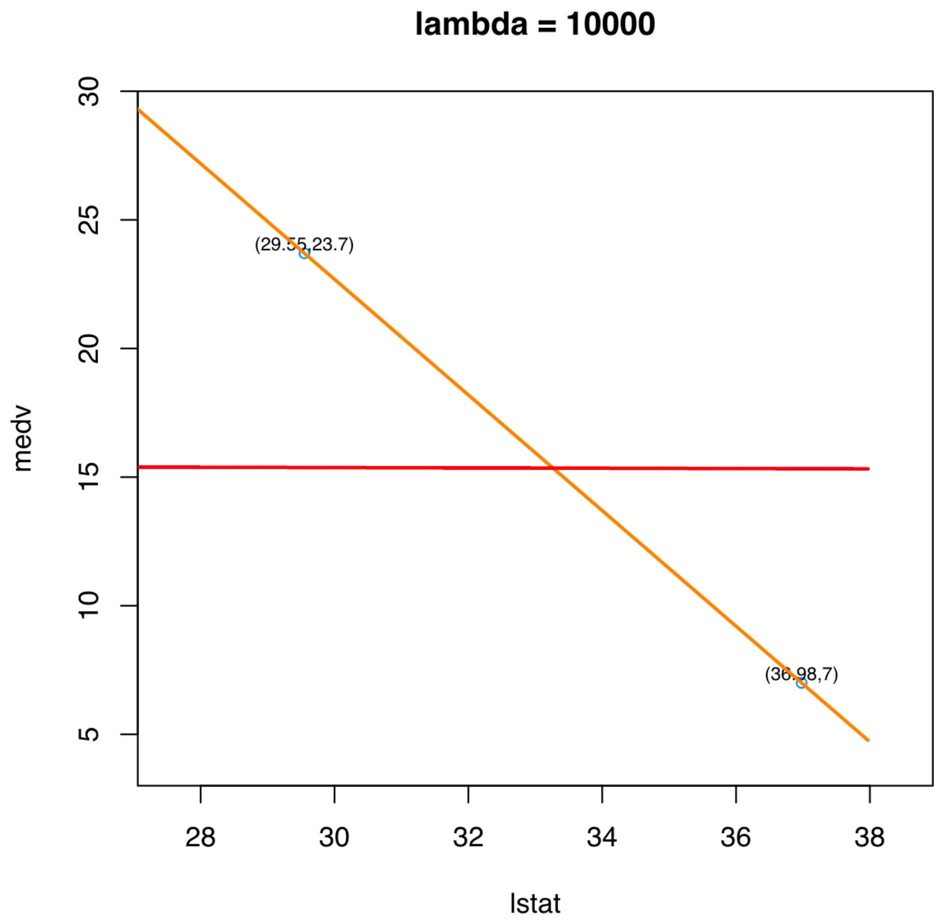
Role of λ in ridge regression

- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - $\lambda = 100$



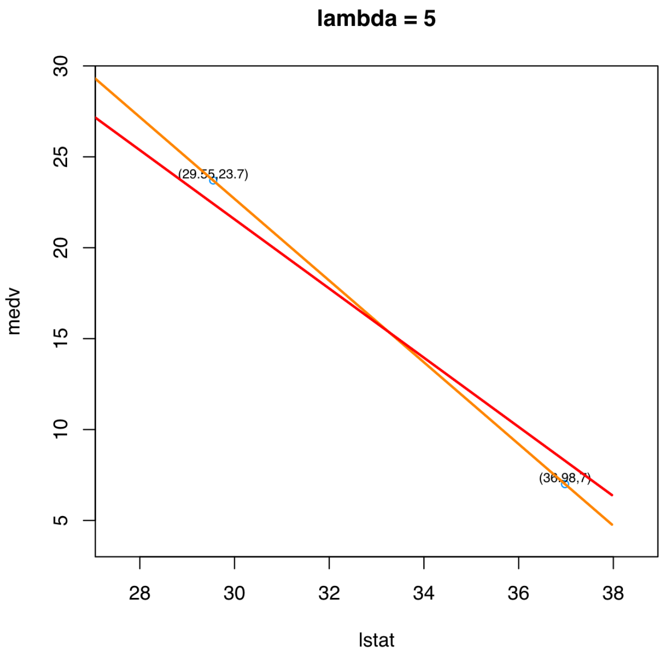
Role of λ in ridge regression

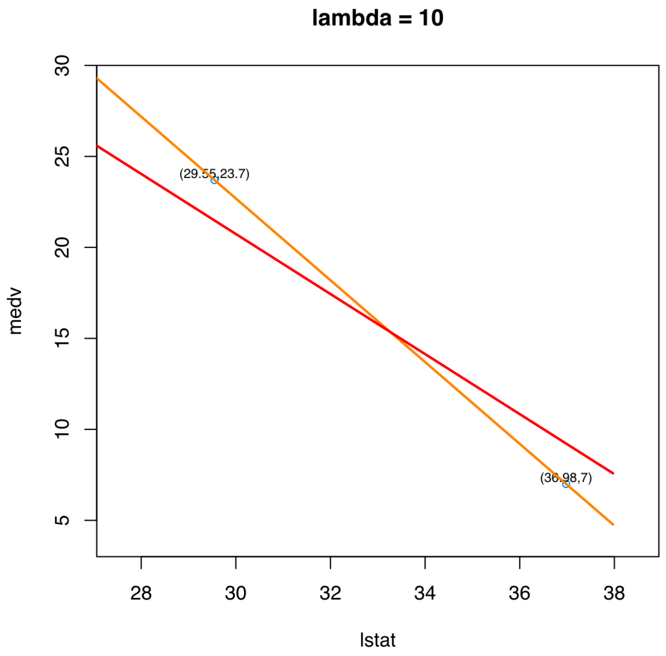
- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - $\lambda = 10{,}000$

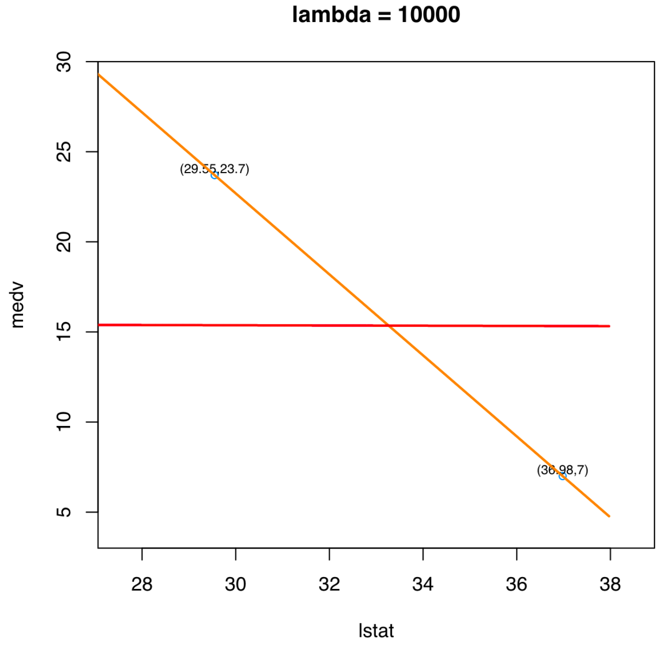


Predictive line is less sensitive to $\Delta lstat$ as λ increases

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







Choose λ by cross-validation

How to choose the optimal λ?

1. Select a grid of λ values

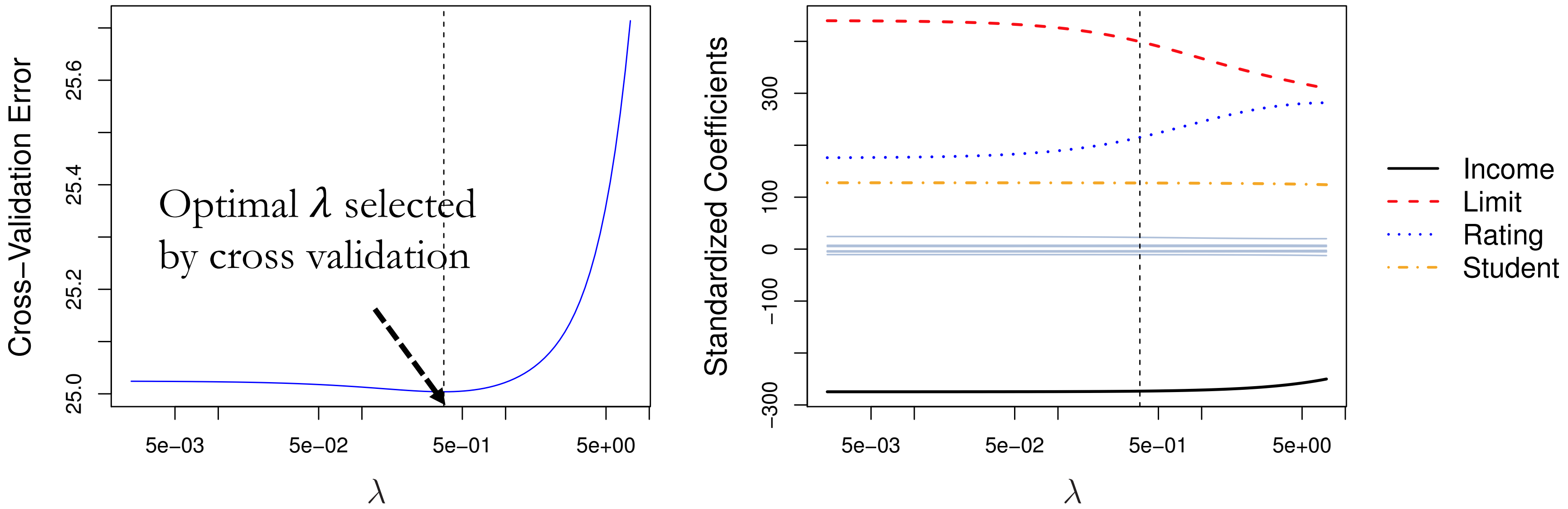
2. Compute the cross-validation error for each λ value

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

4. Refit the model using all observations and selected λ

Example: Credit card dataset (ridge regression)

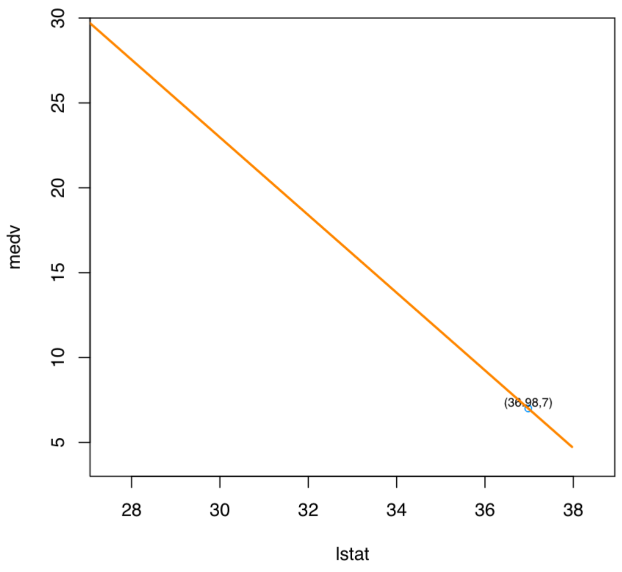
- Cross validation to choose the optimal λ

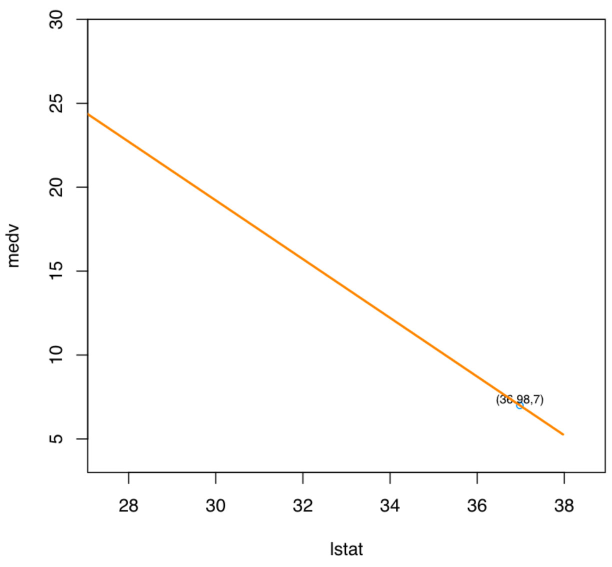


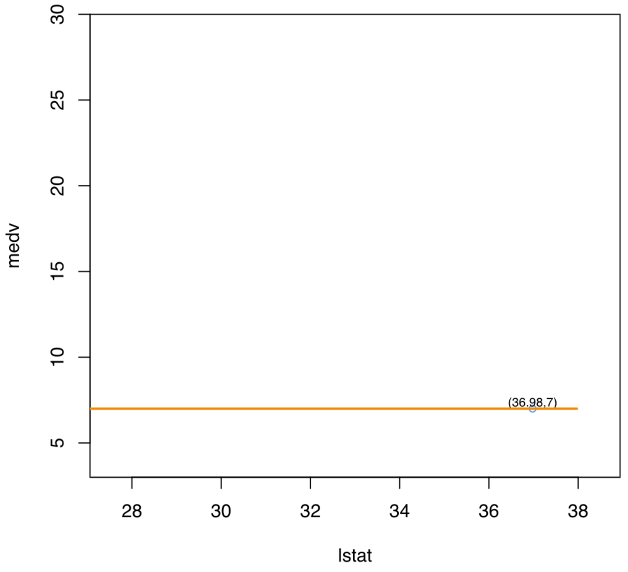
Quiz: Which line is the ridge regression fit?

- One observation ($n = 1$)

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
45.7461	0	18.1	0	0.693	4.519	100	1.6582	24	666	20.2	36.98	7





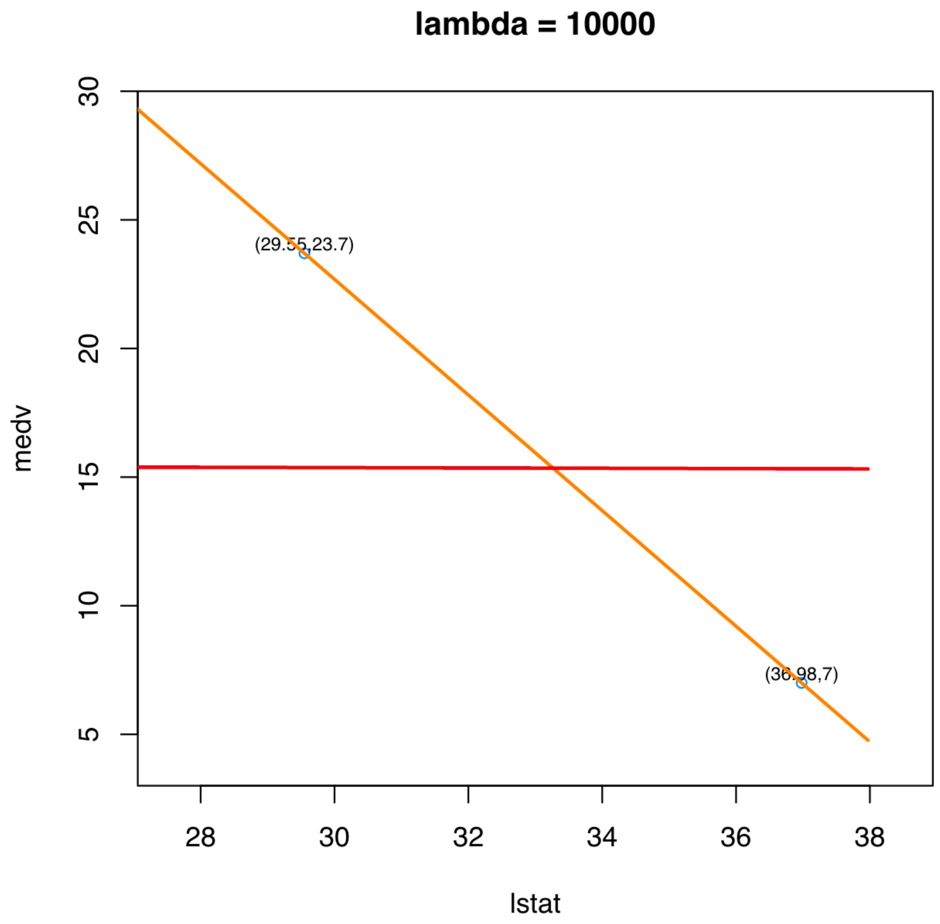


Lecture plan

- LASSO

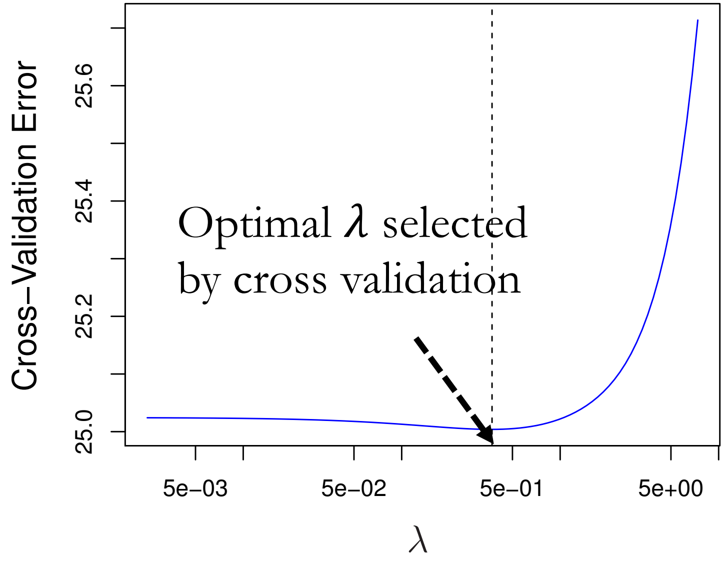
Motivation

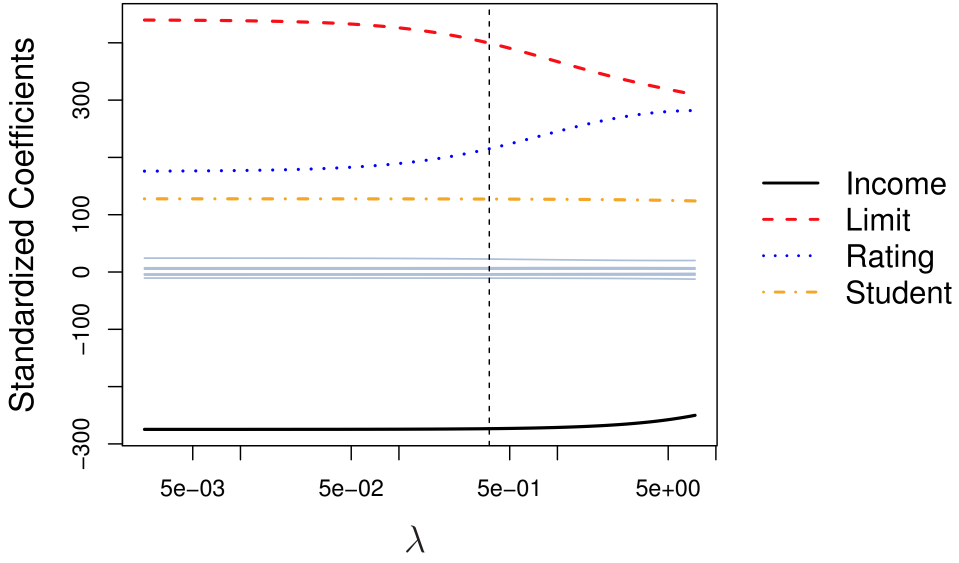
- Ridge regression shrinks coefficients to approximately zero, but not exactly zero
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - When $\lambda = 10{,}000$, $\hat{\beta}_1^R = -0.0062$



What if we want set them as zero?

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





- **What if we want to perform variable selection?**

One predictor

- **LASSO: Least Absolute Shrinkage and Selection Operator**

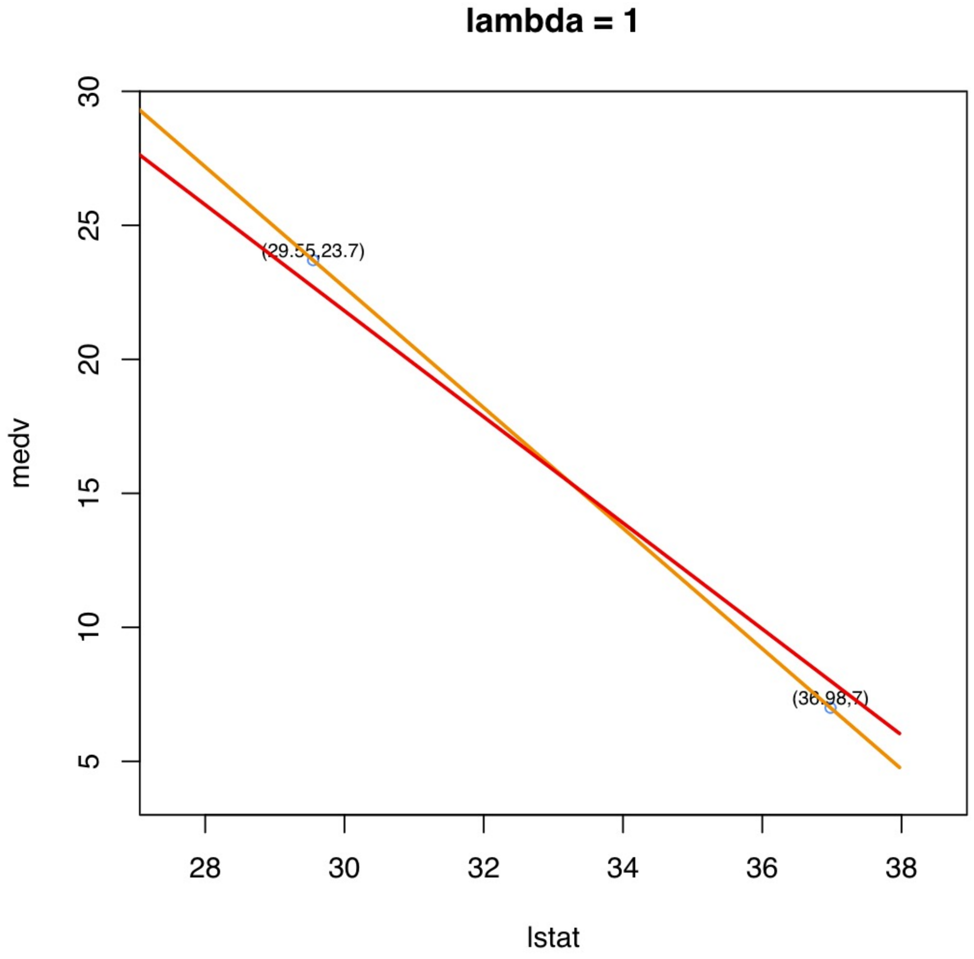
- Lasso minimizes

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

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

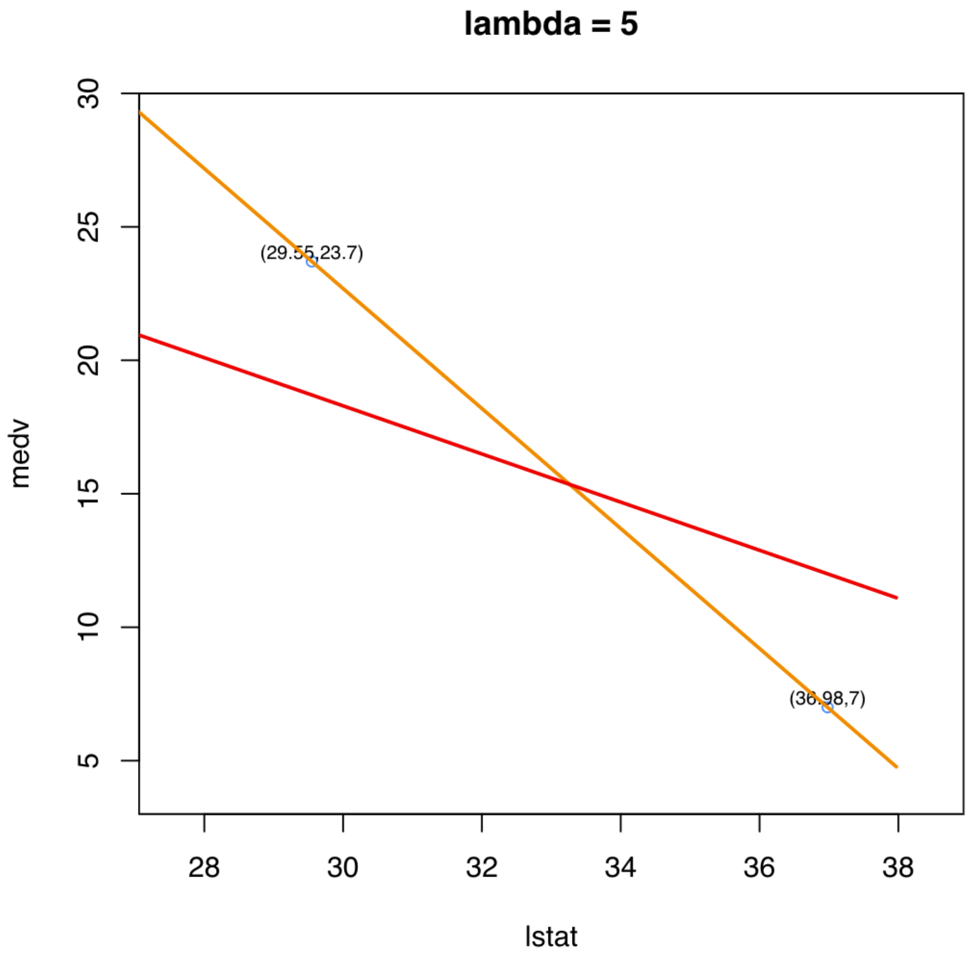
Role of λ in Lasso

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



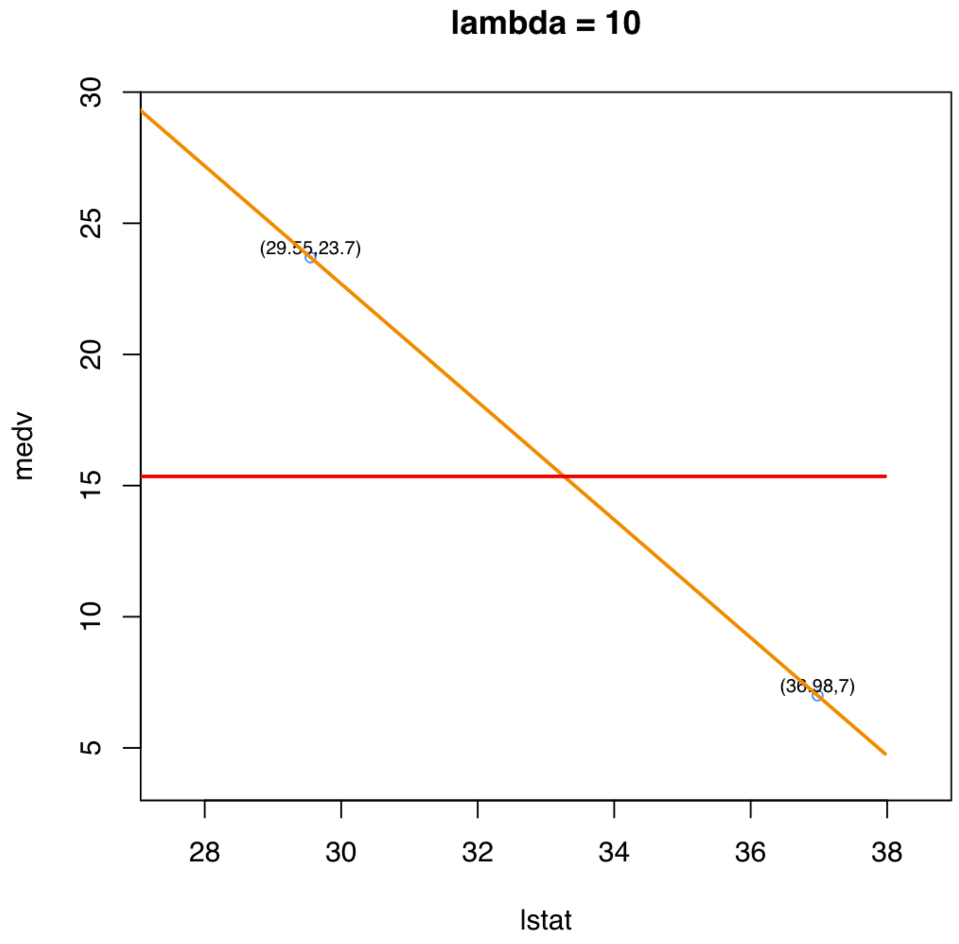
Role of λ in Lasso

- 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
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot |\beta_1|$
  - $\lambda = 10 : \hat{\beta}_1^L = 0$



Multiple predictors

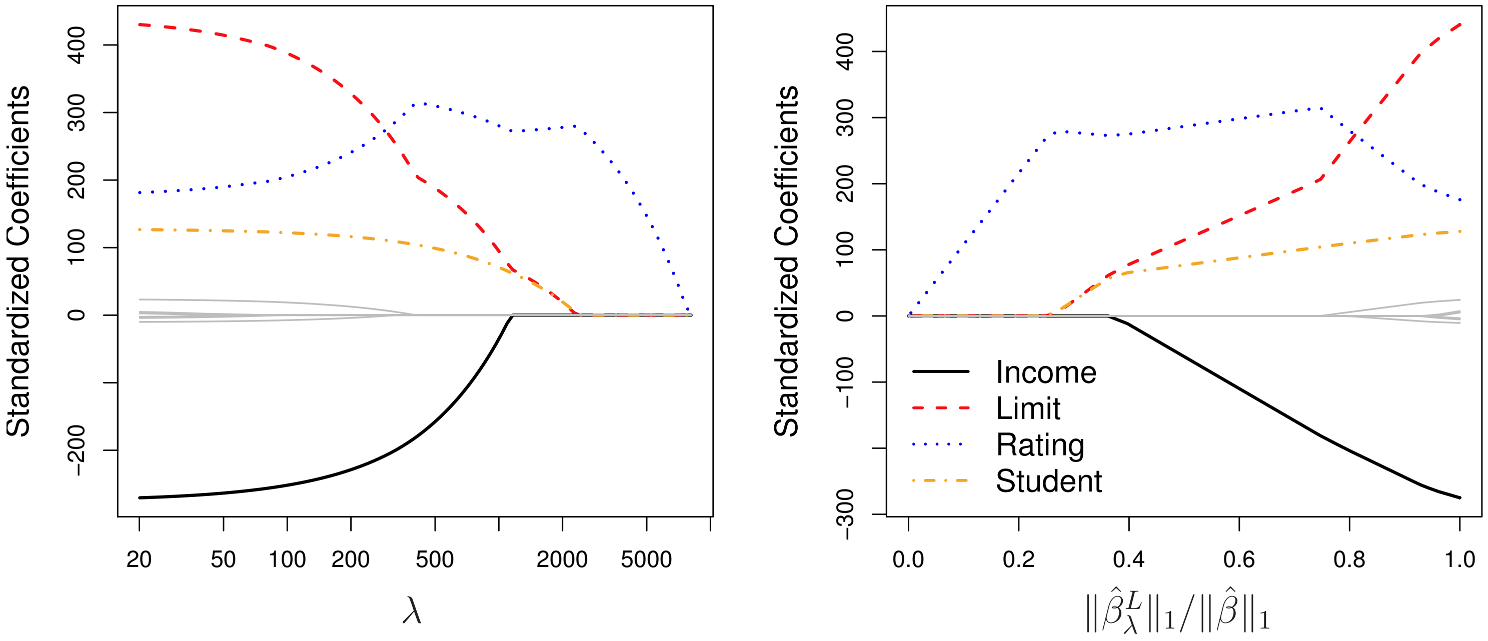
- LASSO minimizes

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

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

Example: Applying LASSO to the credit card dataset

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

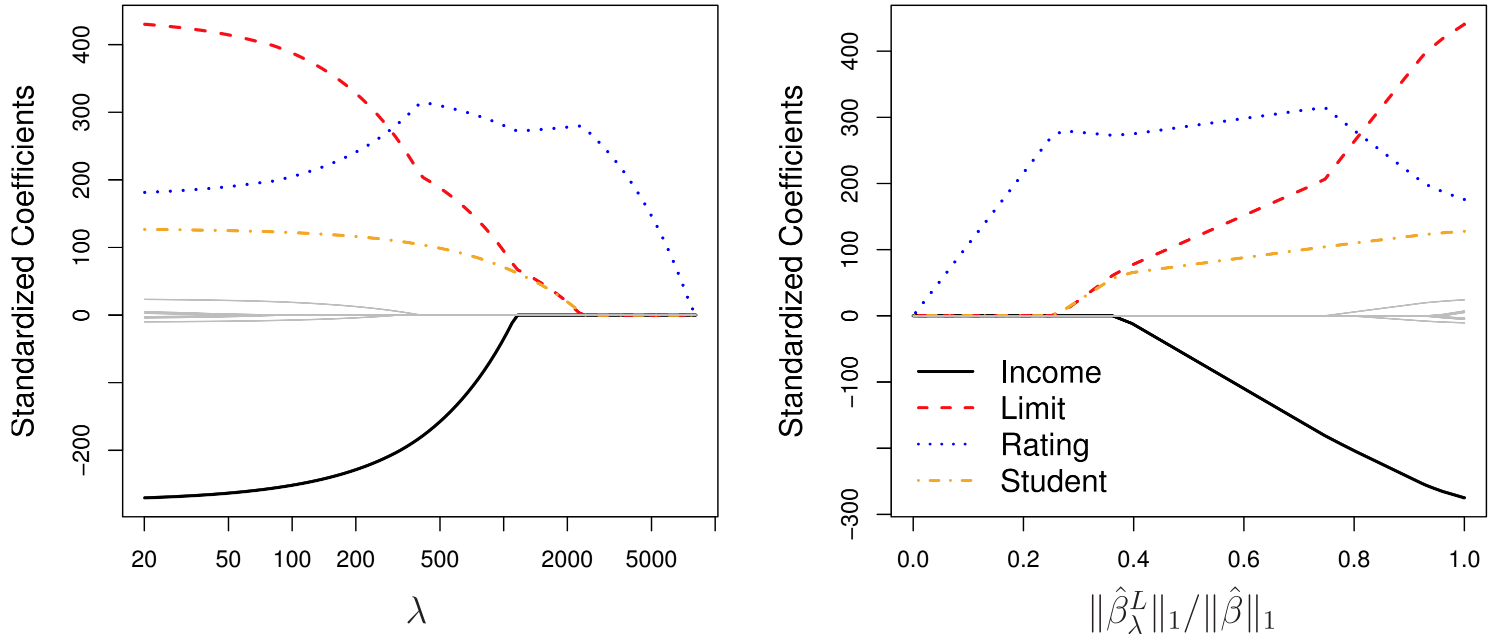


Shrinkage ratio

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

Example: Credit card dataset (LASSO)

- Predict default or not with **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

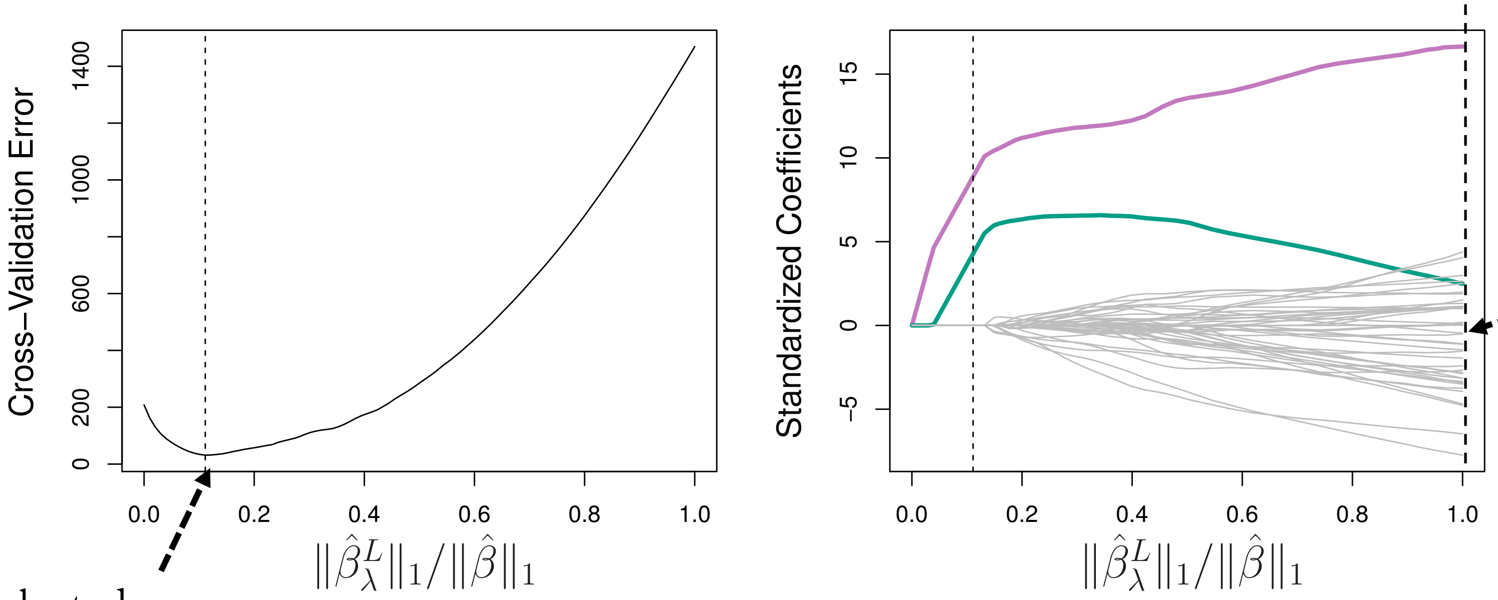
Choose λ by cross-validation

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 $\beta_1, \dots, \beta_{45}$ are nonzero
  - 10-fold CV to select the LASSO regularization parameter



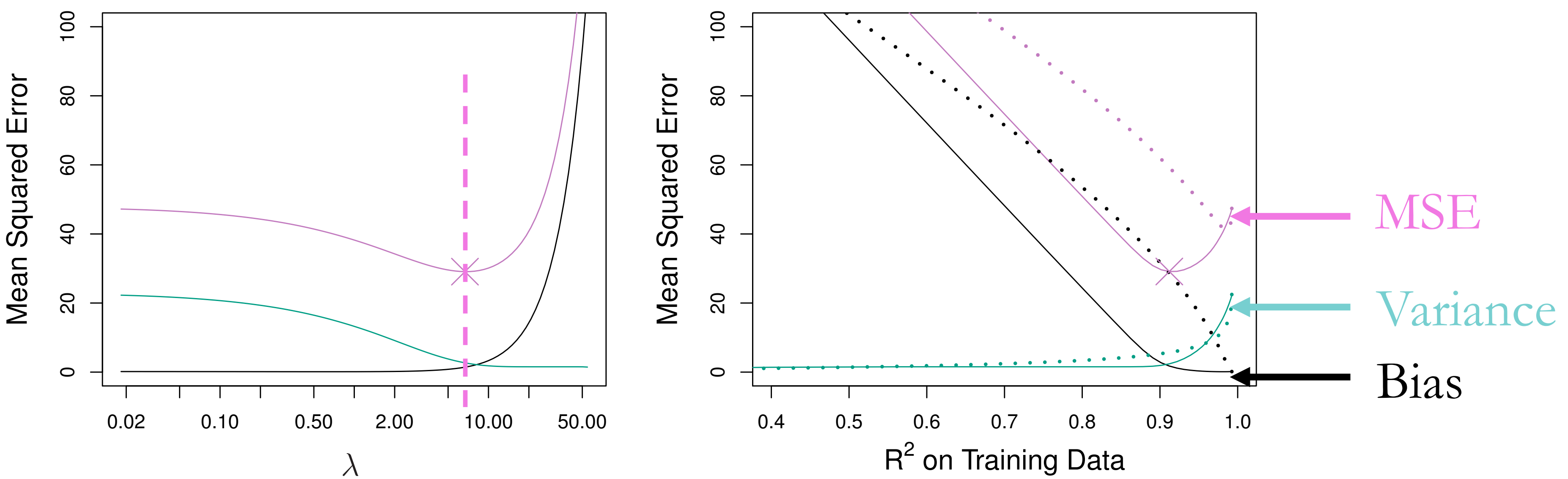
Linear regression solution

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 $\beta_1, \ldots, \beta_{45}$ are nonzero

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

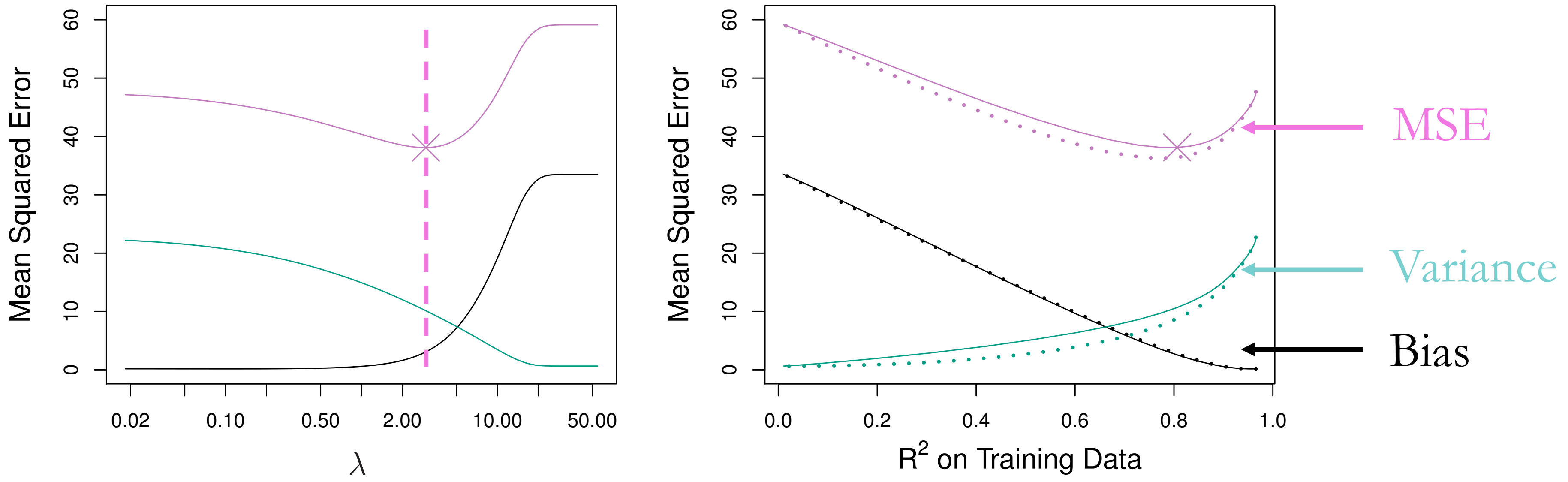


- The **bias**, variance, and MSE are all lower for the LASSO

LASSO vs. Ridge regularization

- **Simulation II:** Most of the coefficients are nonzero
  - 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

Summary

- Lasso performs better if **a small number of predictors with large coefficients**

- Ridge performs better if **many predictors with similar coefficients**