DATASCI 347 Machine Learning

Lecture 9: Regularization

Ruoxuan Xiong

Suggested reading: ISL Chapter 6

Lecture plan

- Ridge regression
- Lasso

Motivation

$$Y = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + \varepsilon$$

- The number of predictors $p > n$
- We have more parameters than observations
- How can we estimate $\beta_1, \beta_2, \cdots, \beta_p$?

Example

- Predict Boston house price

- Suppose we only have one observation ($n = 1$)

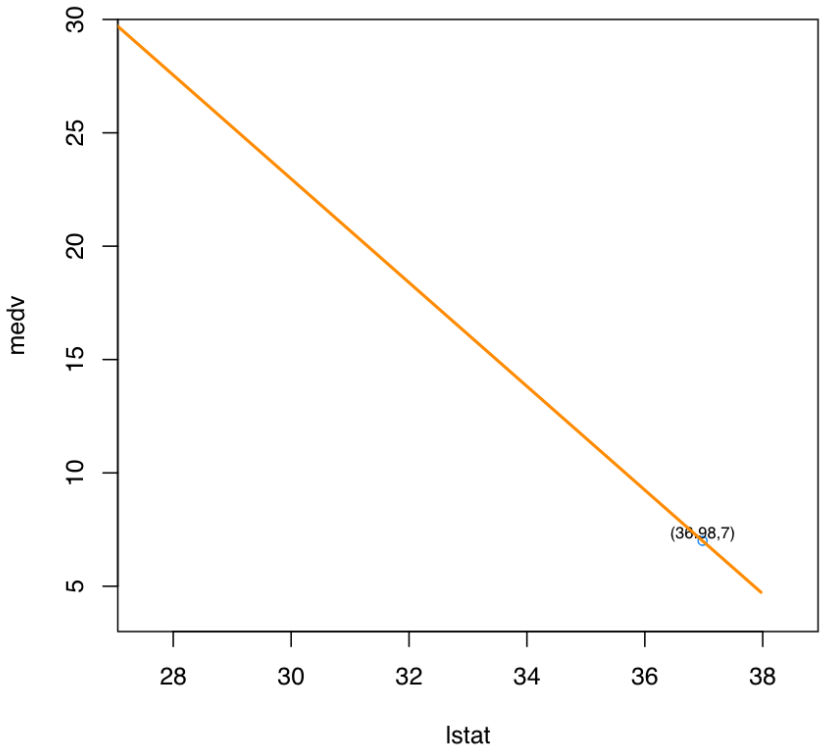
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
45.7461	0	18.1	0	0.693	4.519	100	1.6582	24	666	20.2	36.98	7

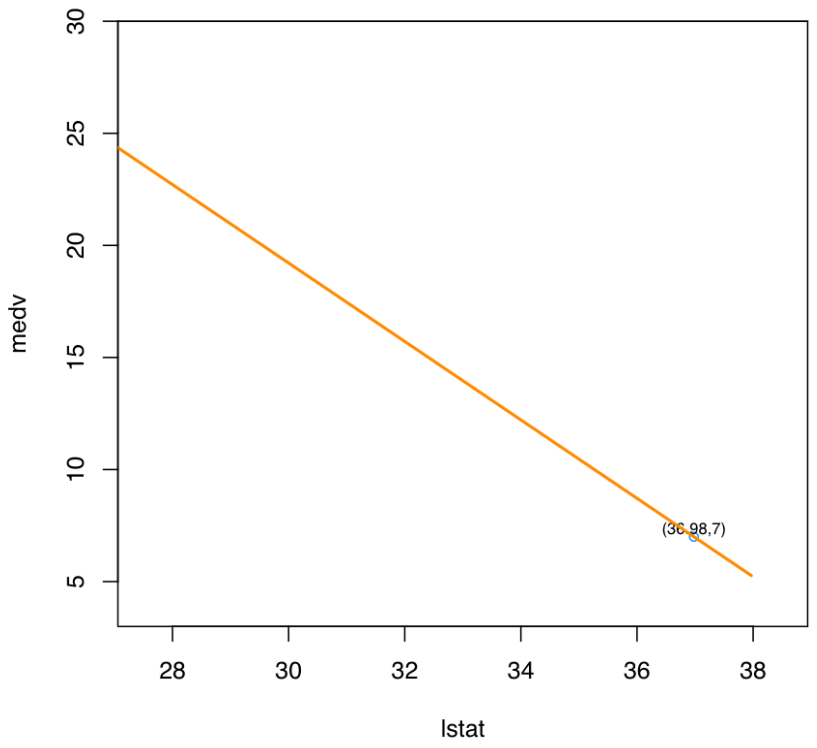
- But we want to estimate the coefficients in the linear model ($p = 2$)

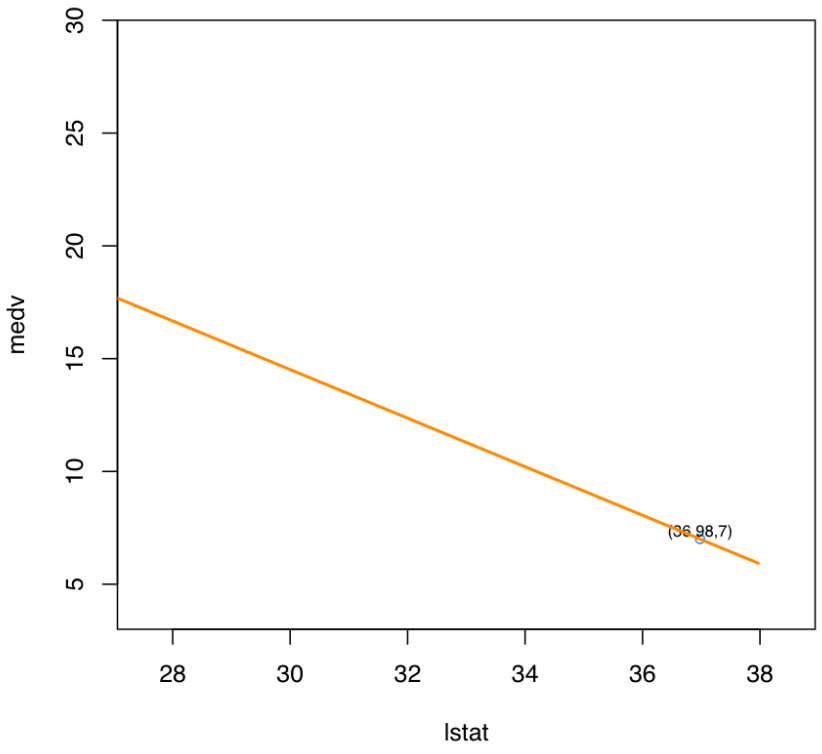
$$medv = \beta_0 + lstat \cdot \beta_1 + \varepsilon$$

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

Which β_0 and β_1 should we choose?







If we have one more observation…

- Predict Boston house price

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

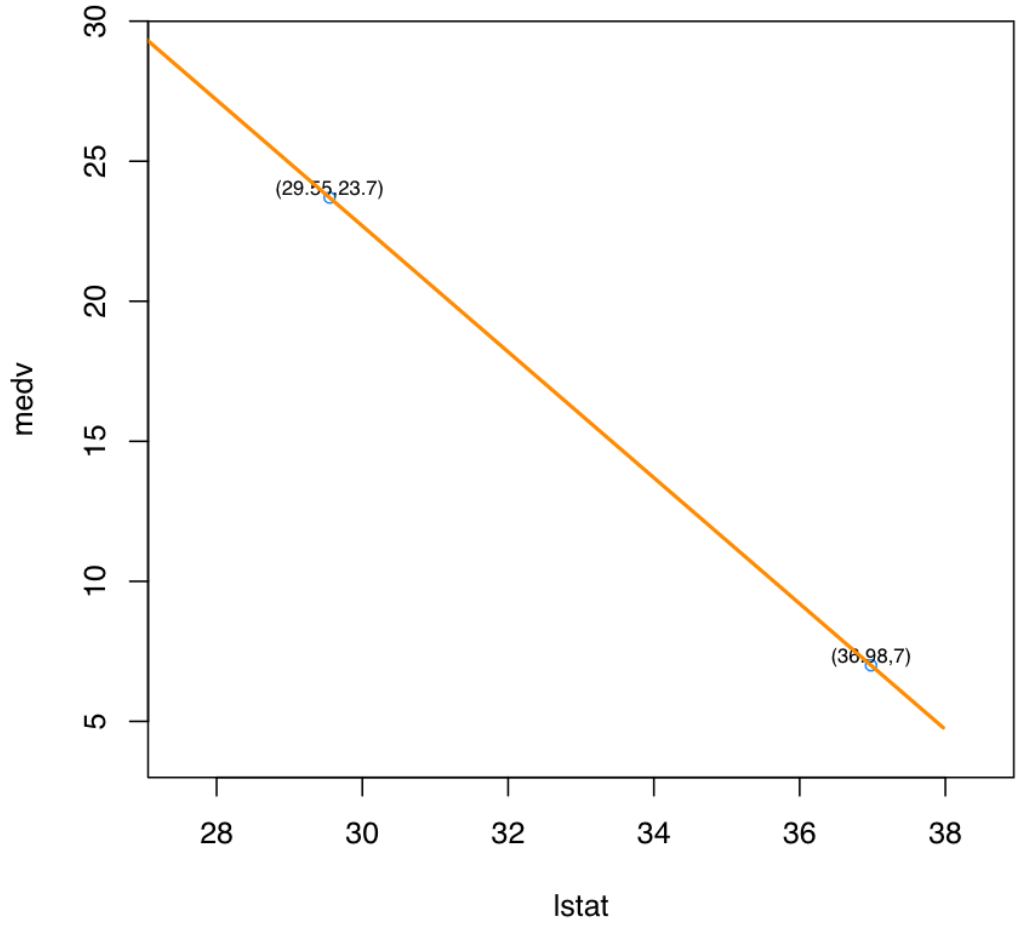
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
0.28955	0	10.59	0	0.489	5.412	9.8	3.5875	4	277	18.6	29.55	23.7
45.74610	0	18.10	0	0.693	4.519	100.0	1.6582	24	666	20.2	36.98	7.0

- Let us consider a simpler linear model ($p = 2$)

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

- We can estimate β_0 and β_1

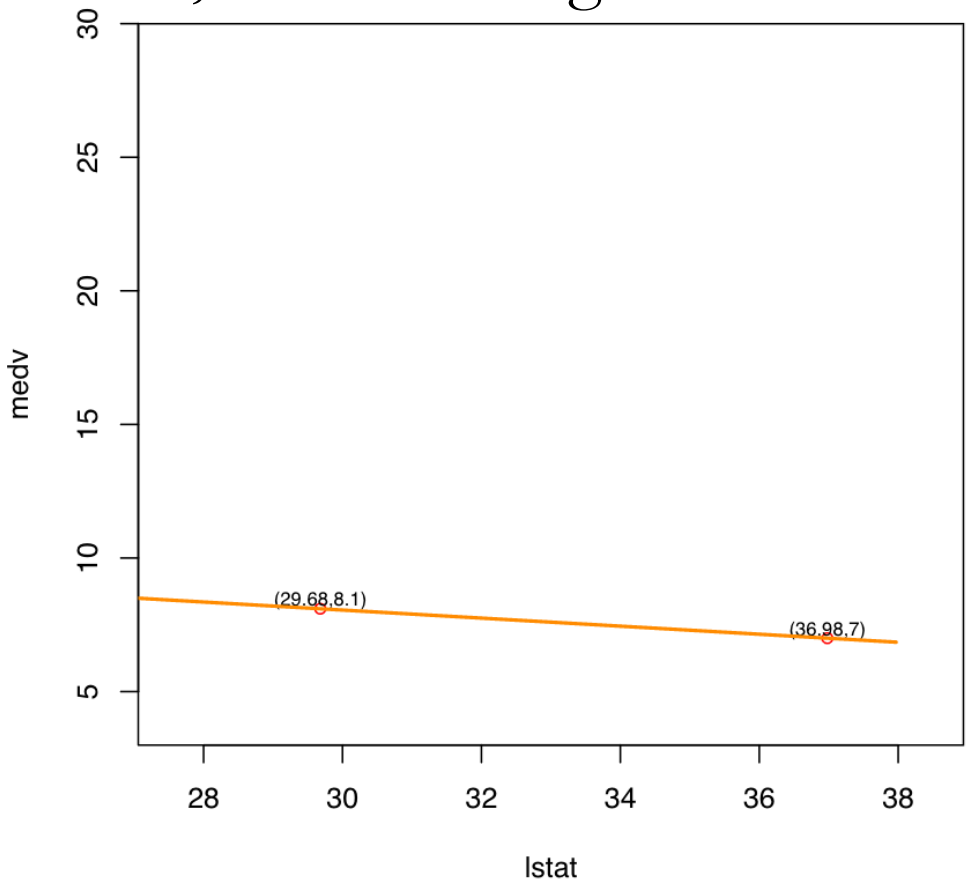
Example



- We still have a problem: The fitted curve is very sensitive to the *medv* of these two observations

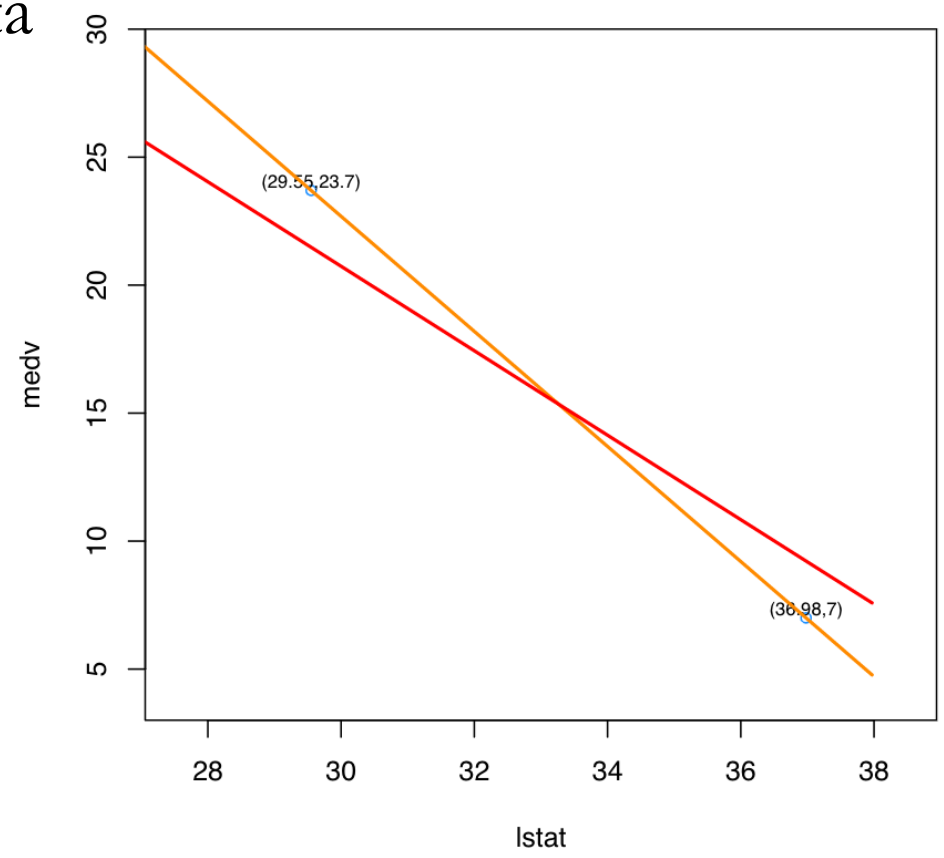
Example

- If one of the two observations changes, we can a very different fitted curve
- The linear model overfits, and has a high variance…



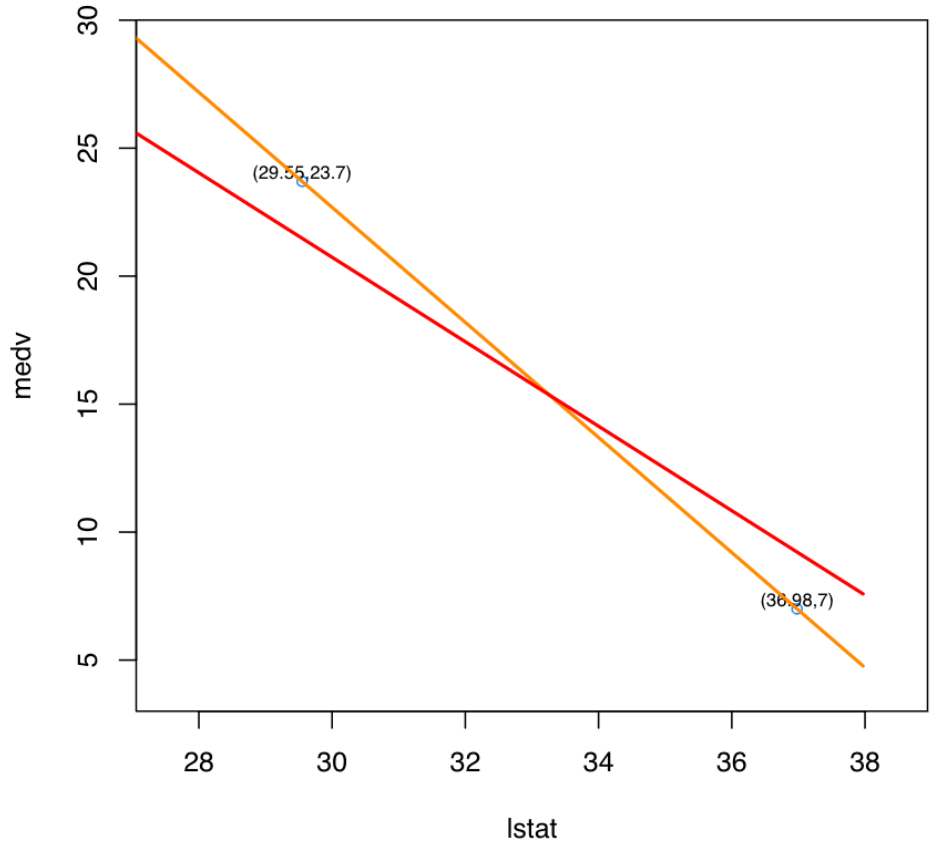
Ridge regression

- Find **a new line** that **does not fit** the **training data** as well
- In other words, we introduce **a small amount of bias** into how the new line is fit to the data



Ridge regression

- We introduce **a small amount of bias** into how the new line is fit to the data
- But in turn for that small amount of bias, we get a **significant drop in variance**

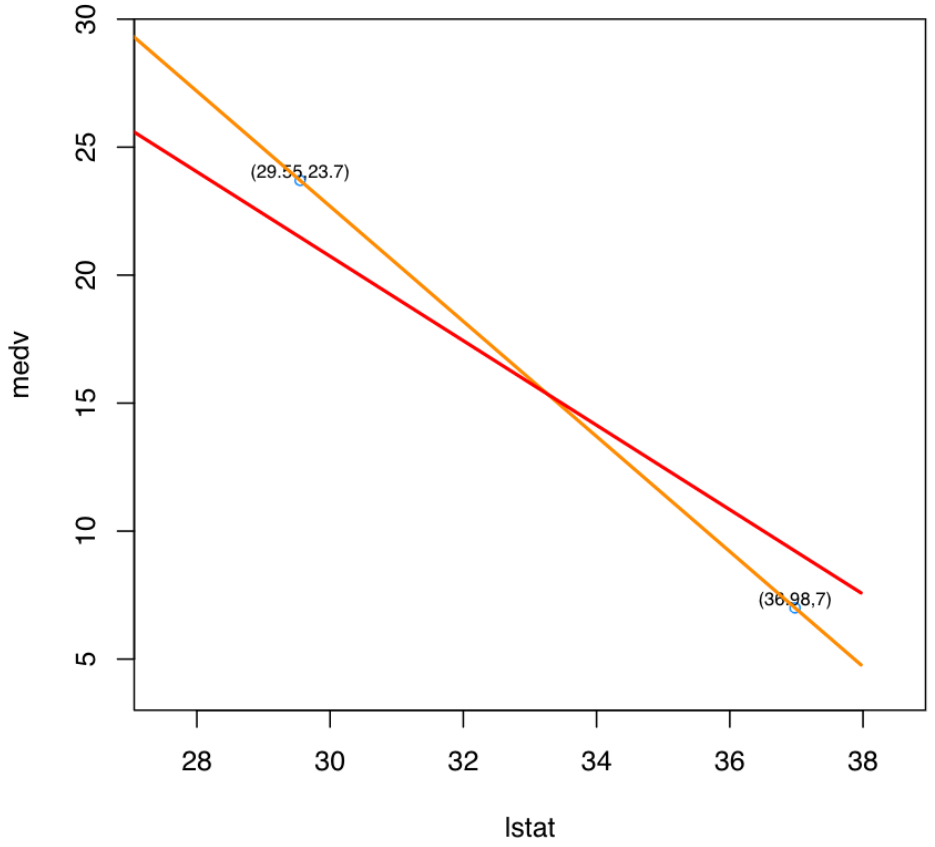


Fitting ridge regression

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

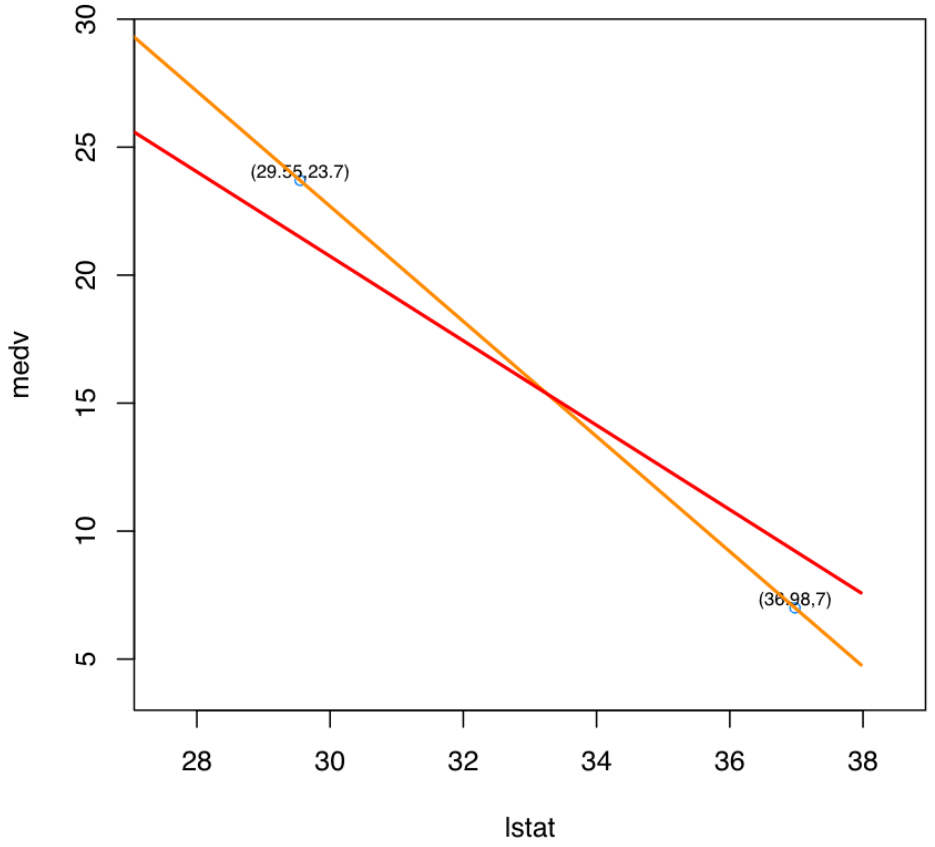
- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$

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



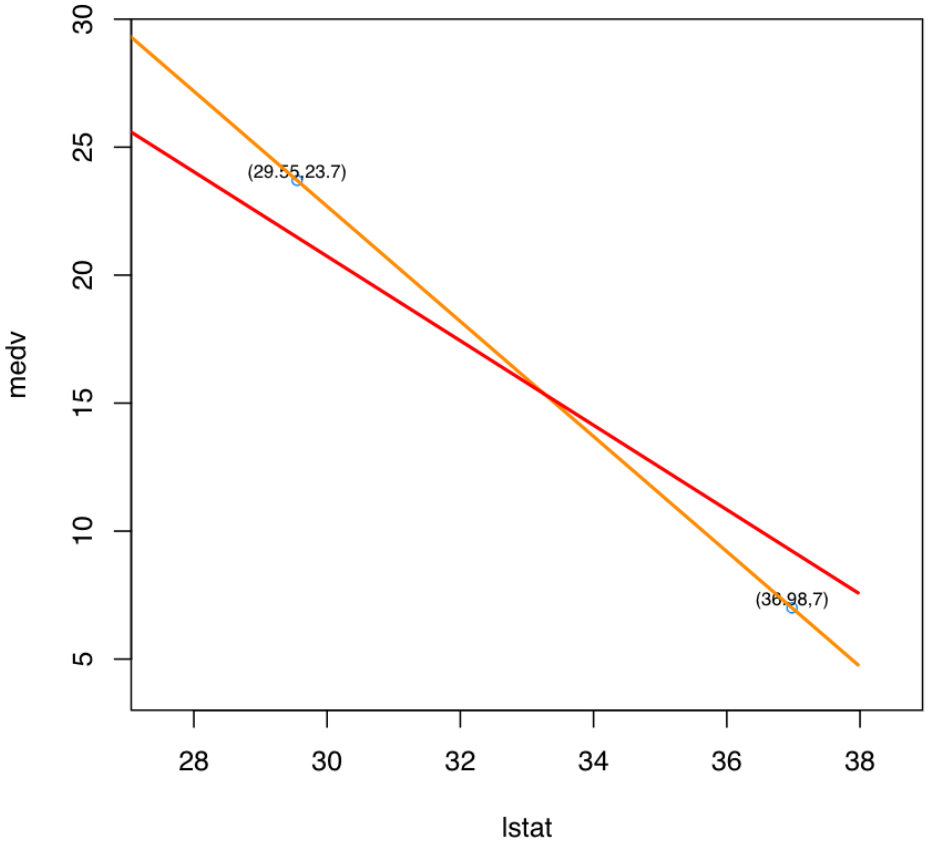
Objective value of least squares solution

- Suppose $\lambda = 10$
- Linear regression fit: $\widehat{medv} = 90.118 - 2.248 \cdot lstat$
  - $\hat{\beta}_1 = -2.248$
  - $\sum_{i=1}^{n} \left(medv_i - \hat{\beta}_0 - lstat_i \cdot \hat{\beta}_1\right)^2 + \lambda \cdot \hat{\beta}_1^2$
    $= 0 + 10 \cdot 2.248^2 = 50.535$



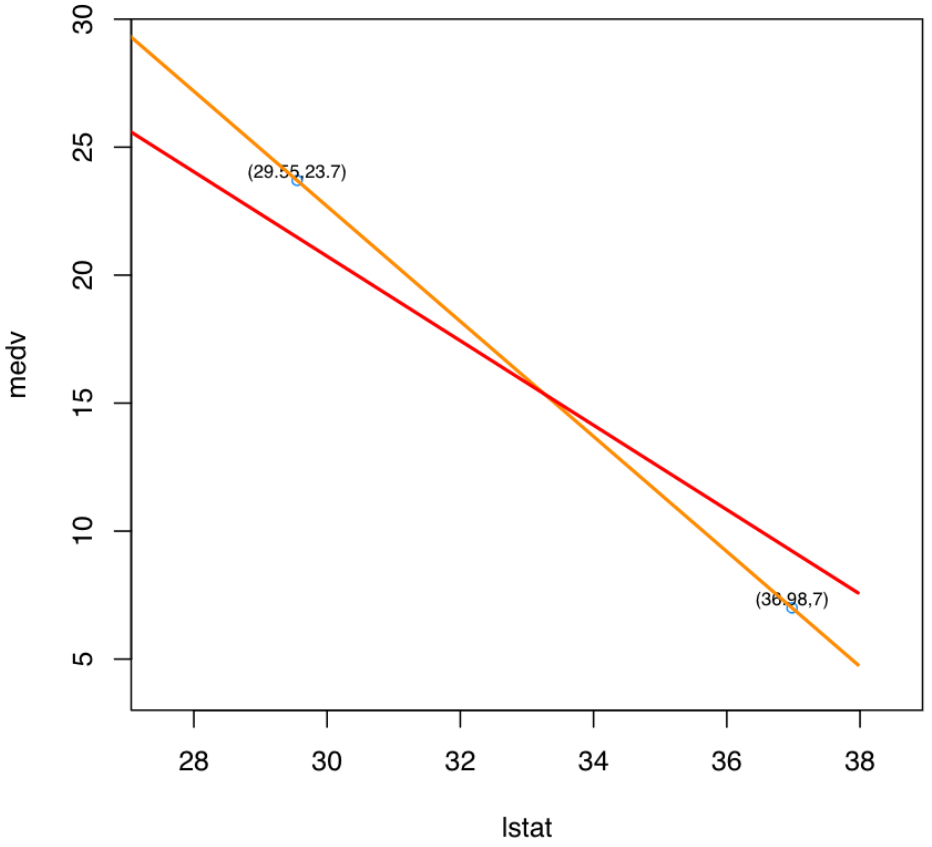
Objective value of ridge regression solution

- 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



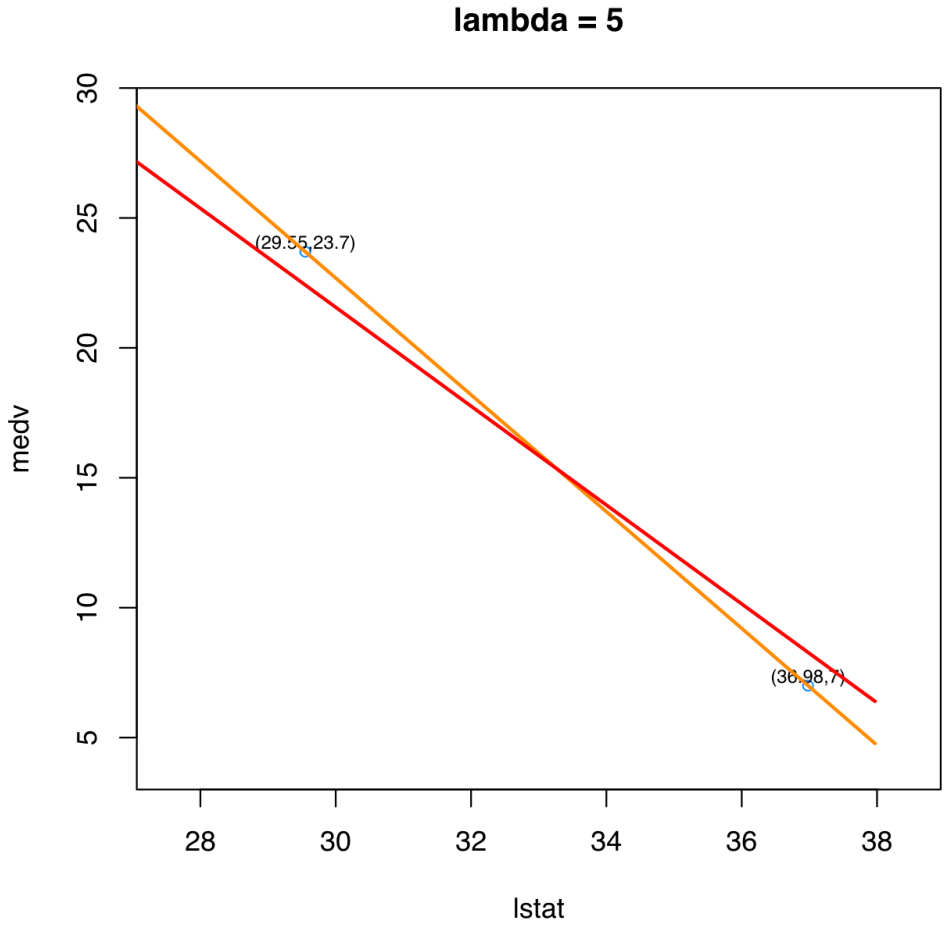
Ridge regression is less sensitive to $lstat$

- 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 $medv$



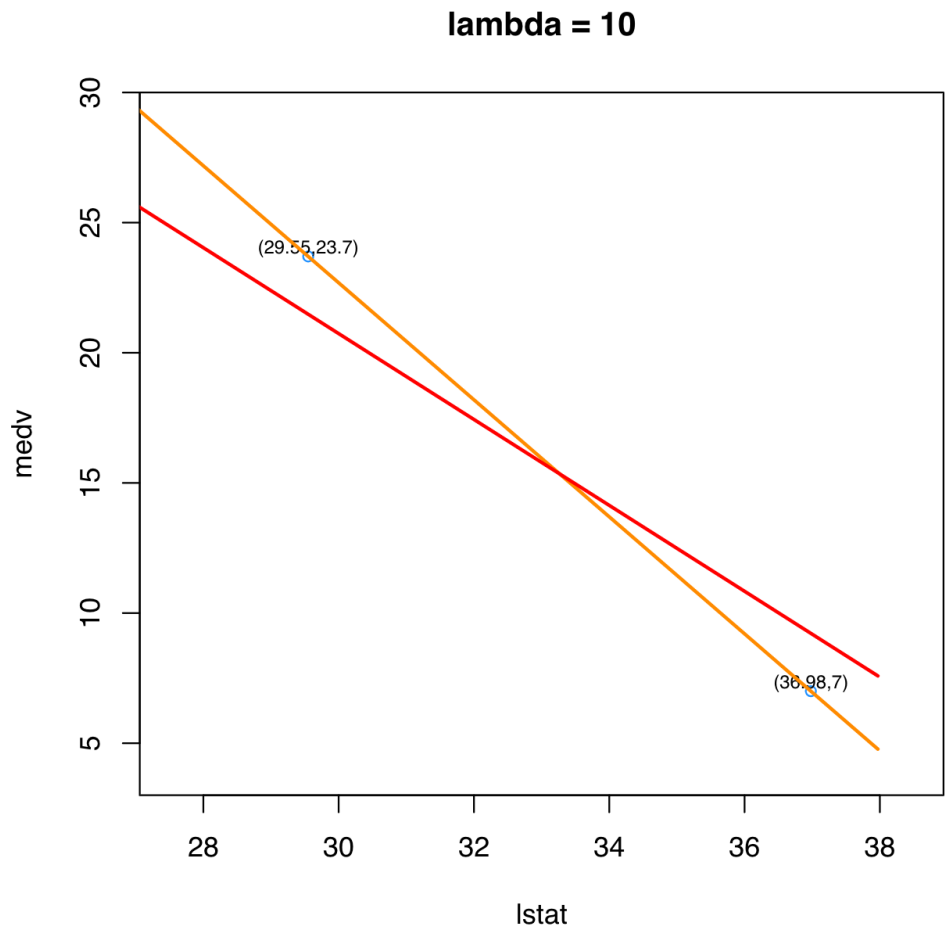
Role of λ in ridge regression

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



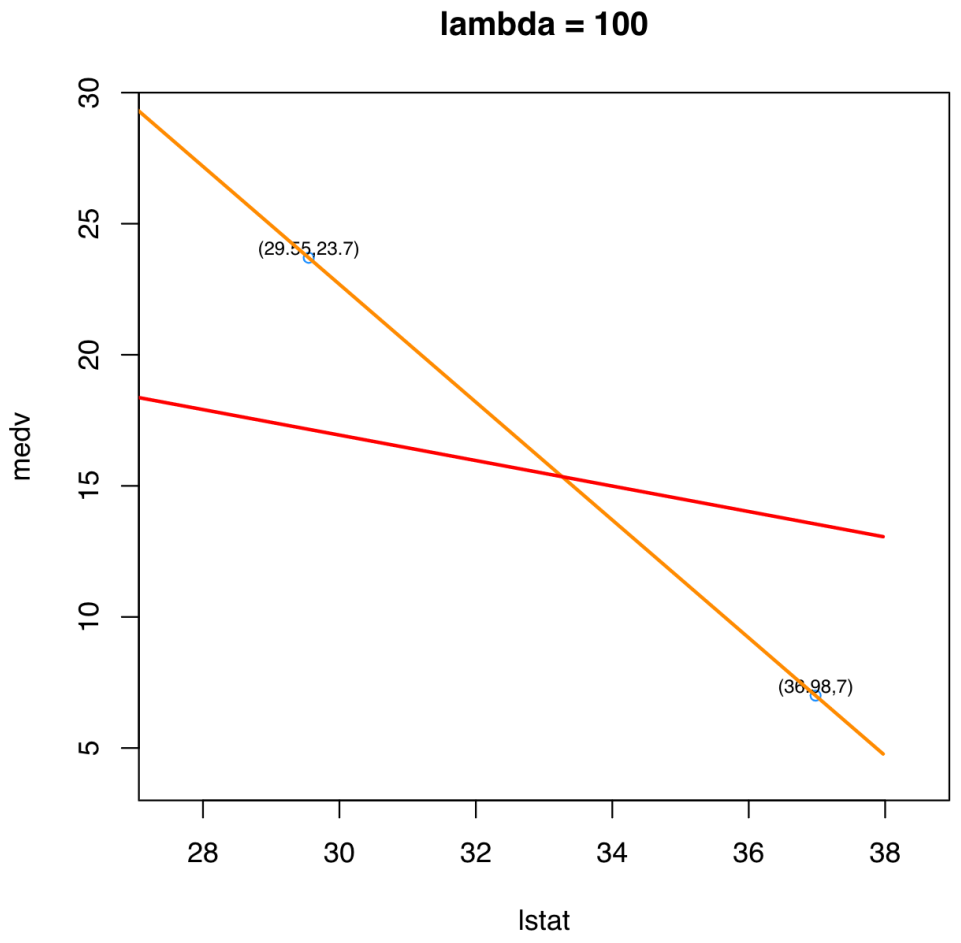
Role of λ in ridge regression

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



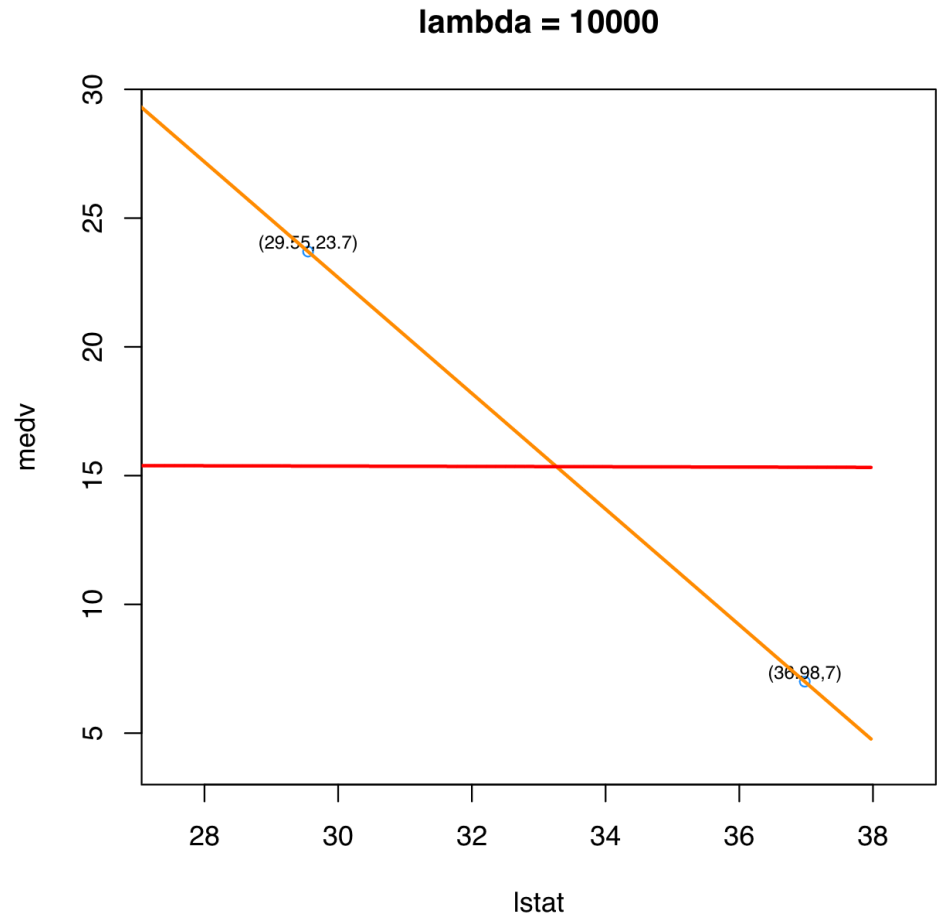
Role of λ in ridge regression

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



Role of λ in ridge regression

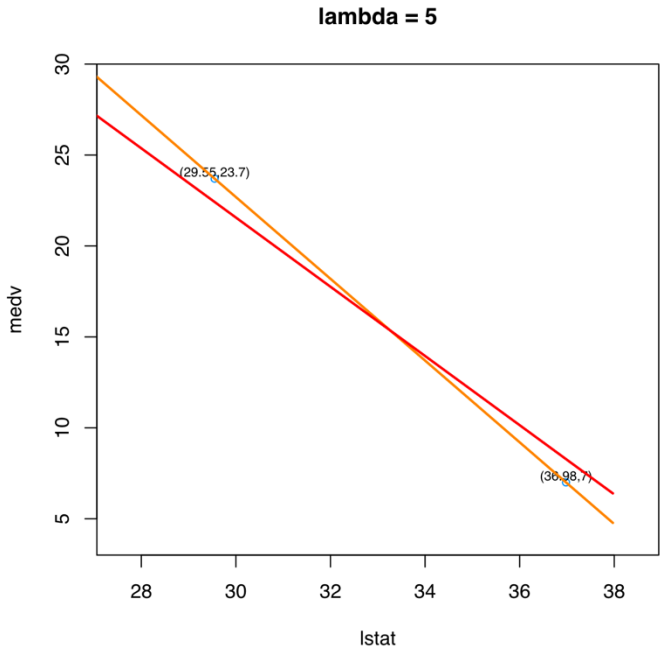
- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$

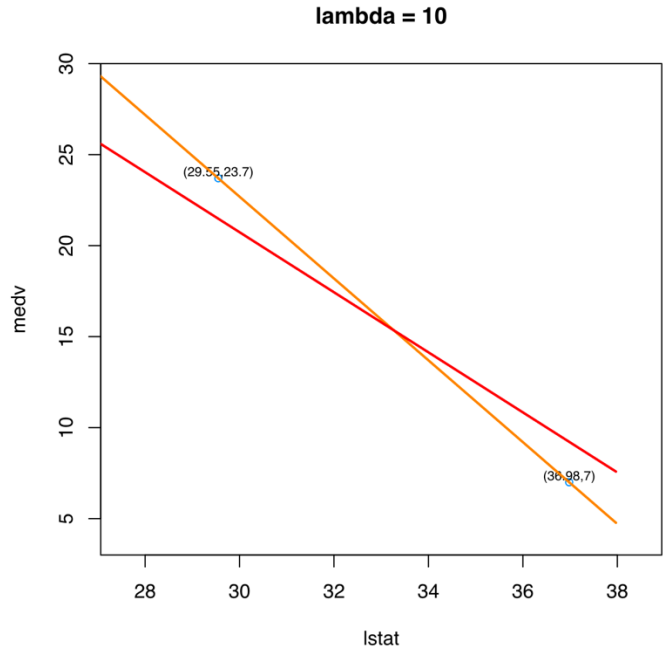


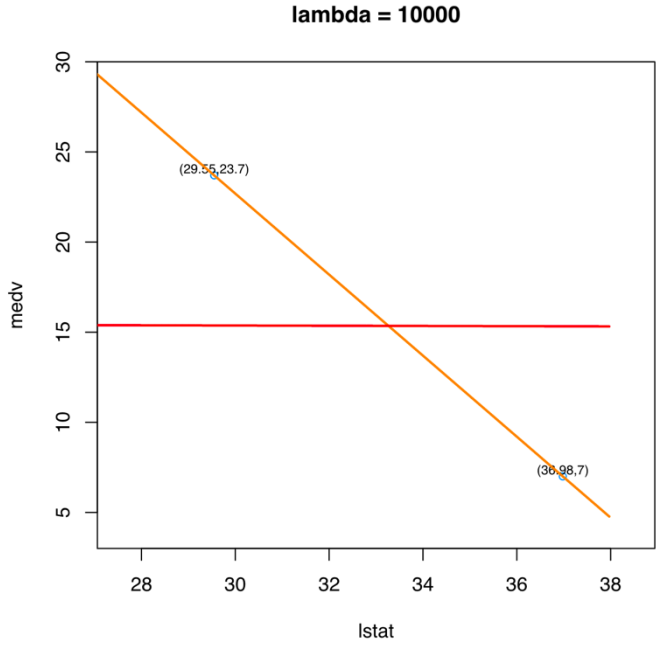
Our prediction becomes less sensitive to $lstat$ as λ increases

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

- How to choose the optimal λ?







Choose λ by cross-validation

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

Ridge regression for more than one predictor

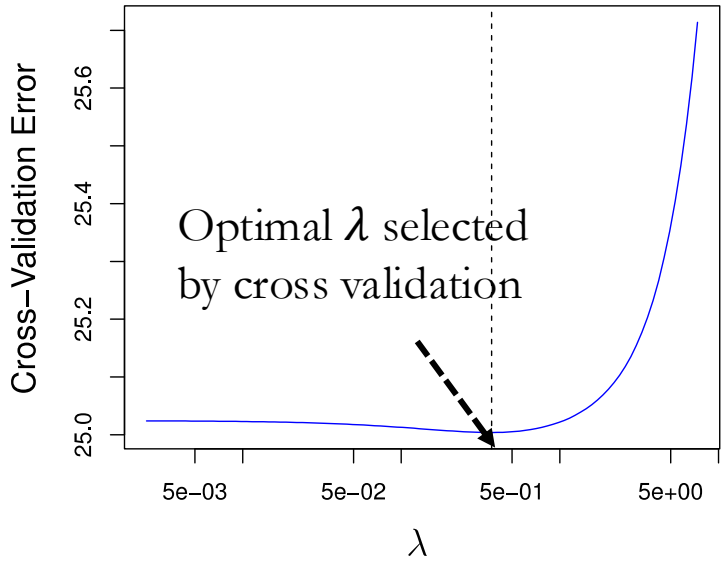
- Ridge regression minimizes

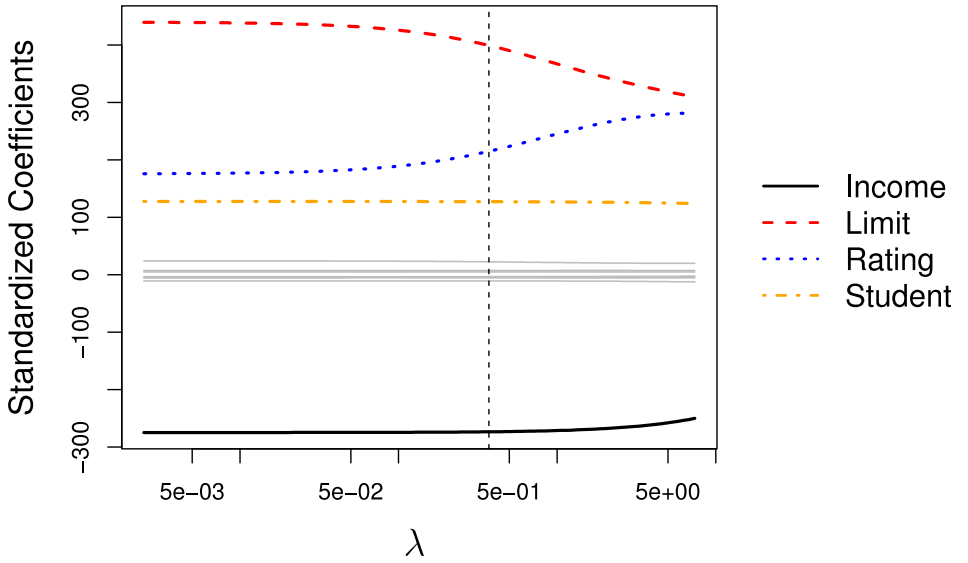
$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{i,j}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

  - $X_{i,j}$: 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

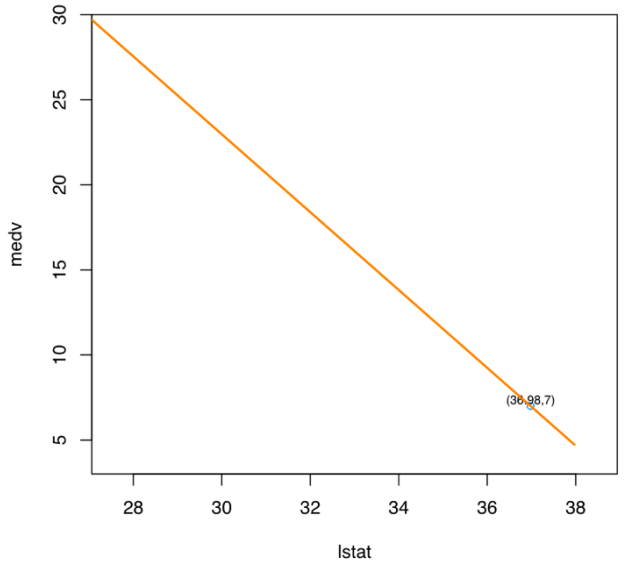


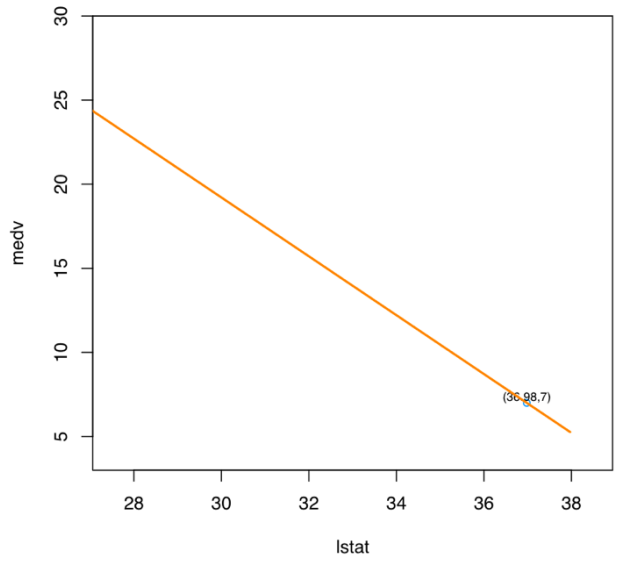


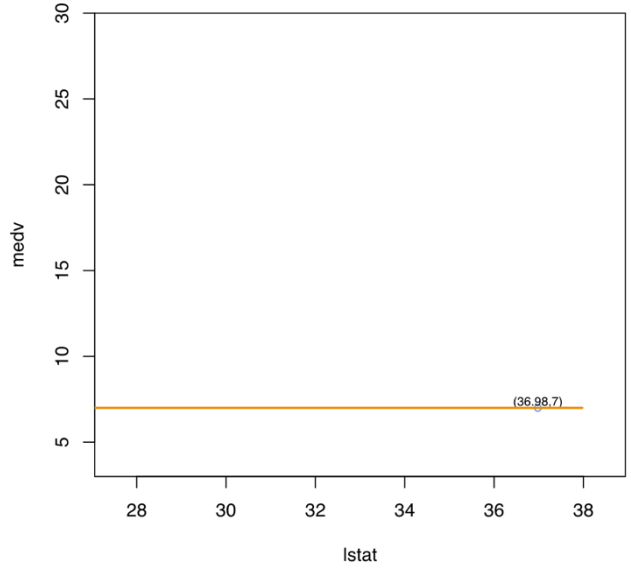
Quiz: Which is the ridge regression fit?

- 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







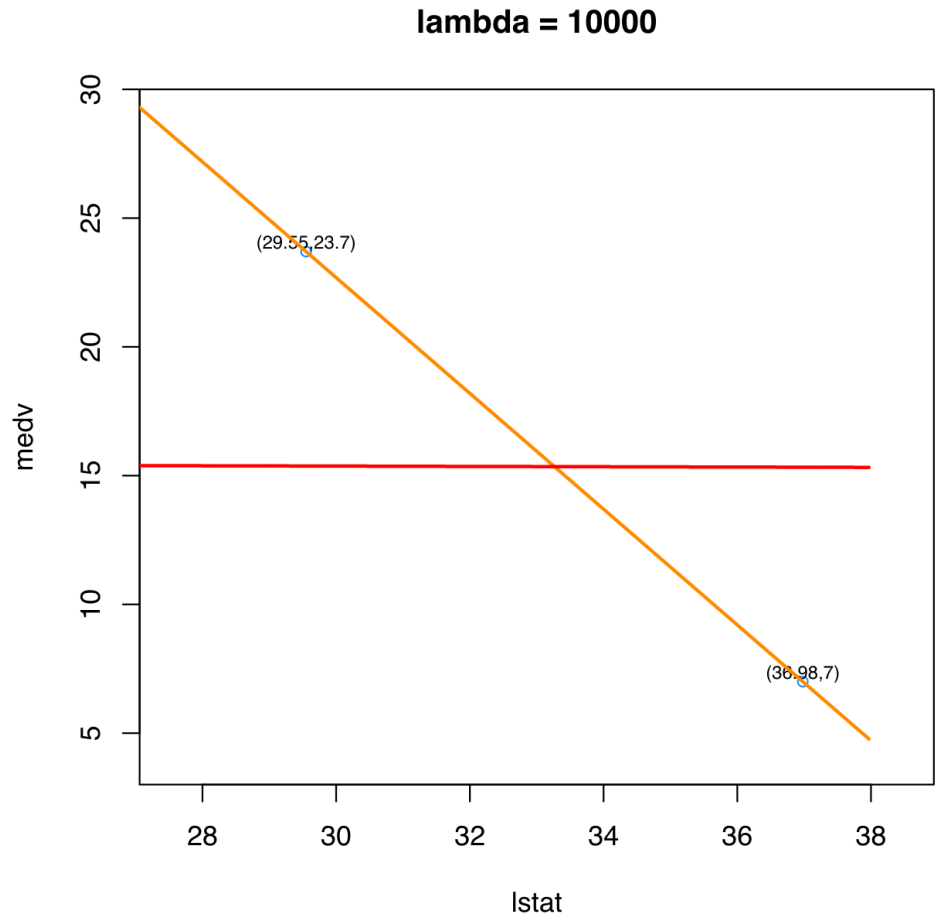
Lecture plan

- Ridge regression

- Lasso

Motivation

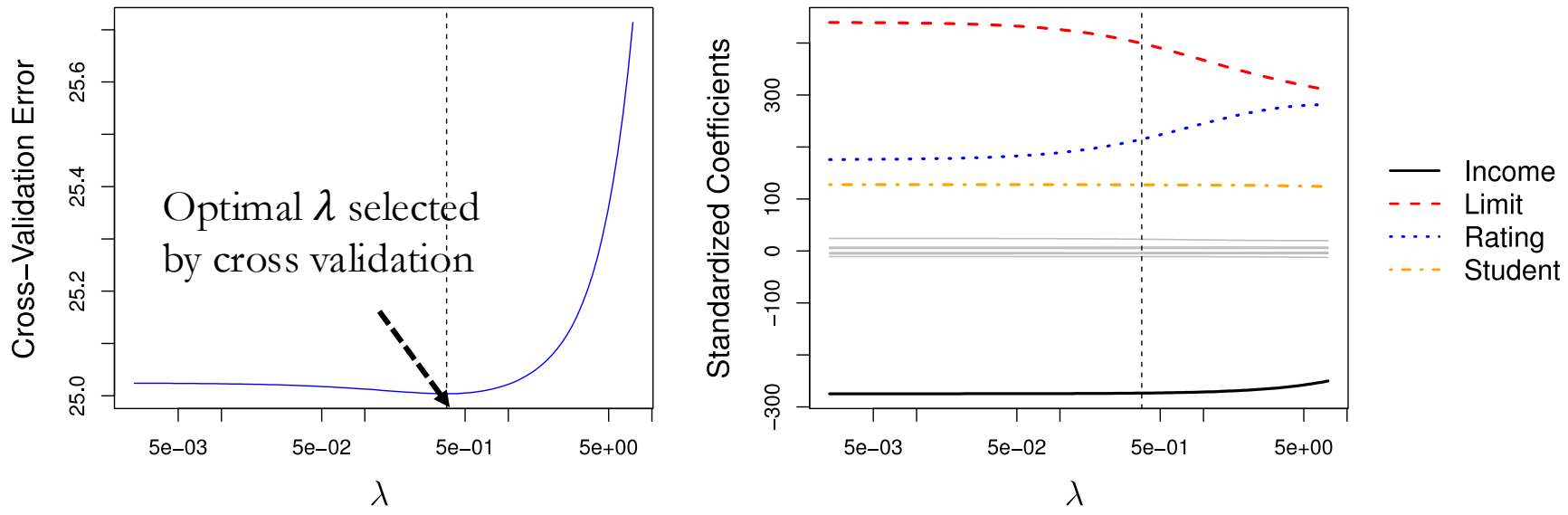
- Ridge regression shrinks coefficients to approximately zero, but not exactly zero
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$
  - When $\lambda = 10{,}000$, $\hat{\beta}_1^R = -0.0062$



What if we want 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

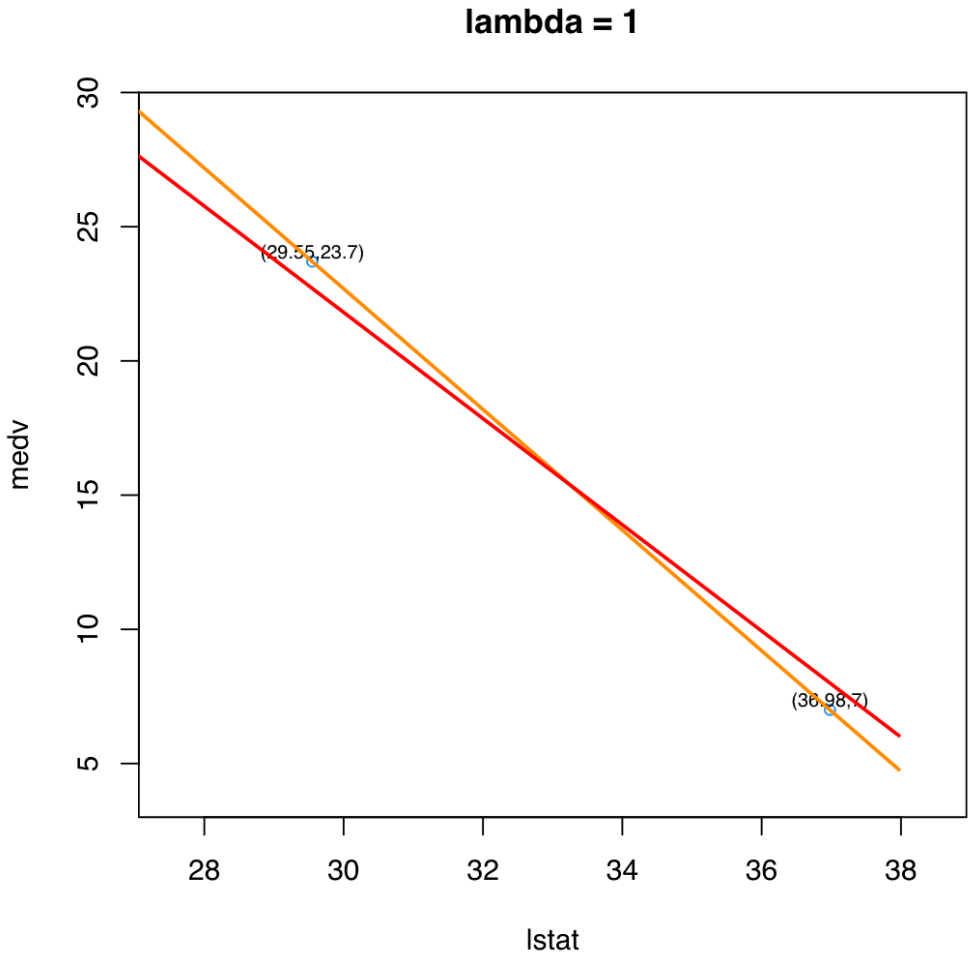
- Lasso: least absolute shrinkage and selection operator

- 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

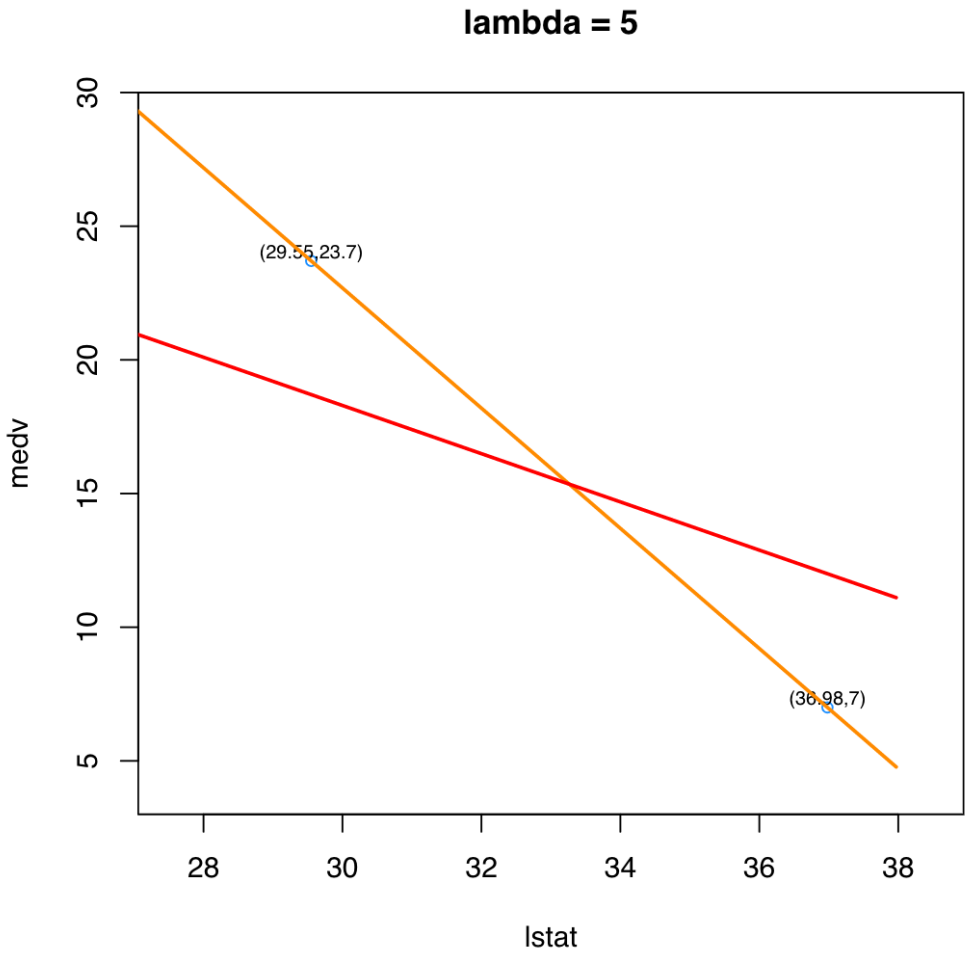
Role of λ in Lasso

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



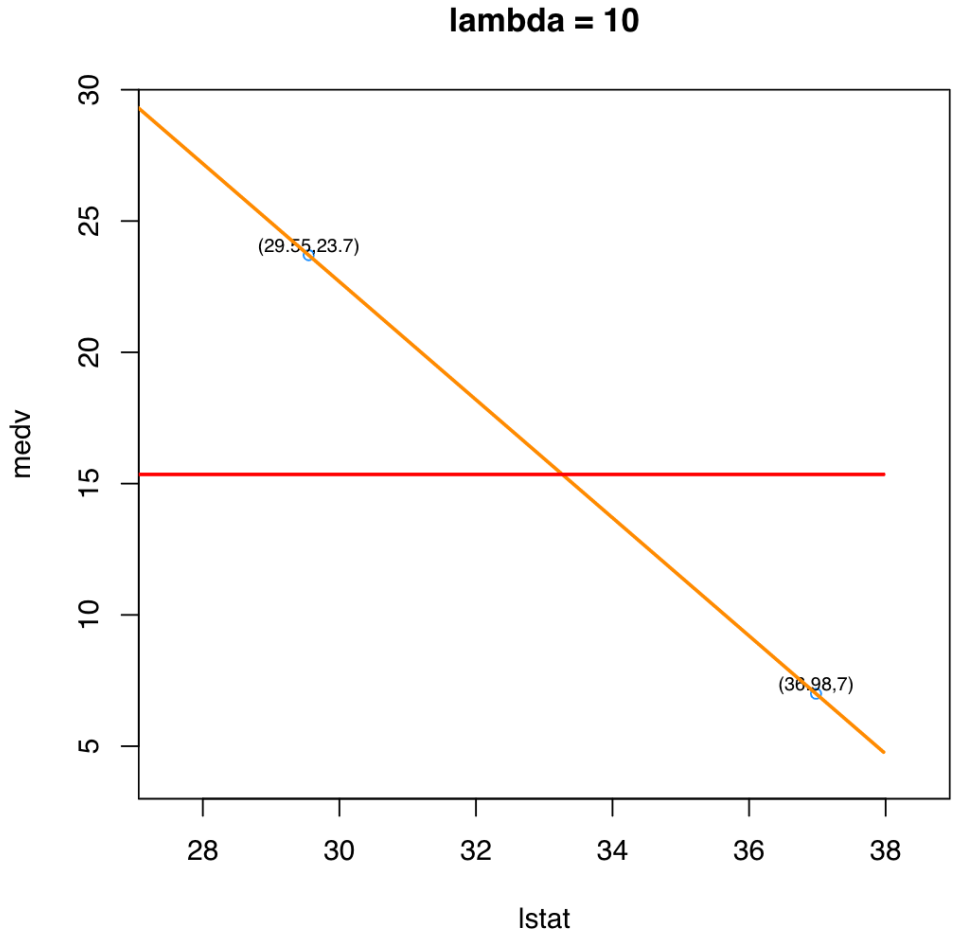
Role of λ in Lasso

- 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

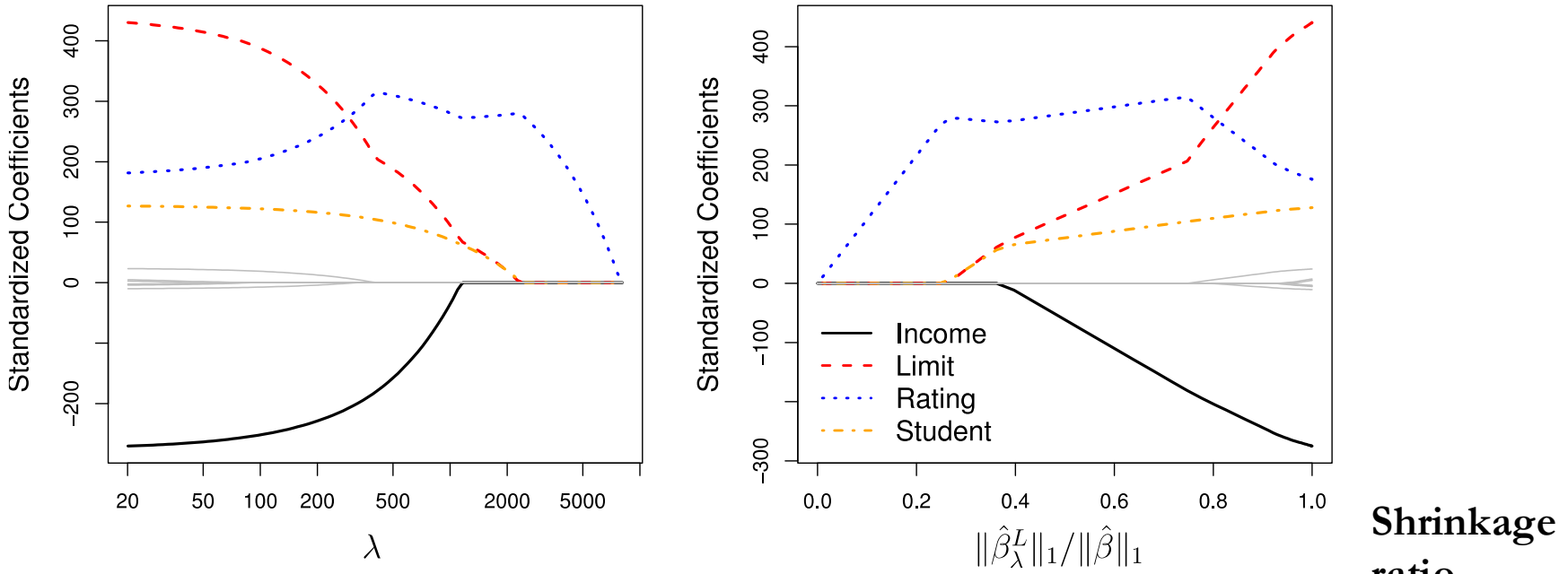
- Lasso minimizes

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

  - $X_{i,j}$: 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

Example: Credit card data set (lasso)

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

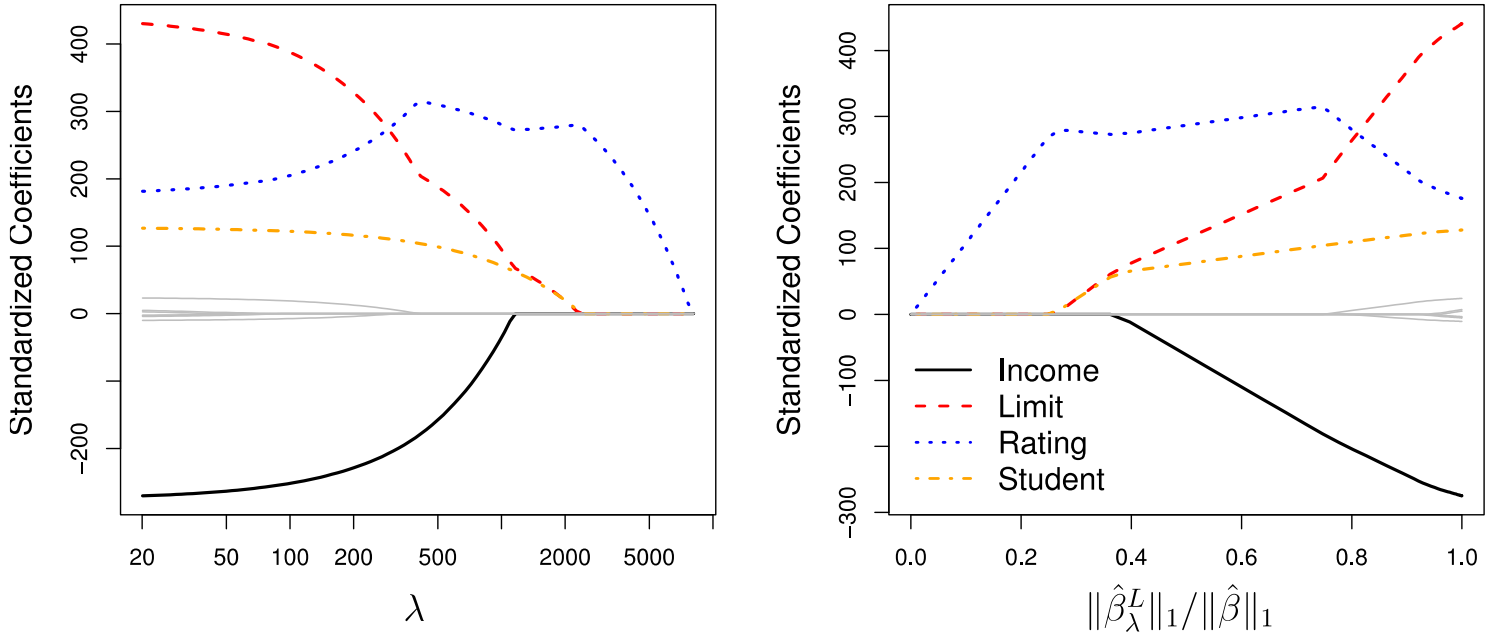


Shrinkage ratio

- **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

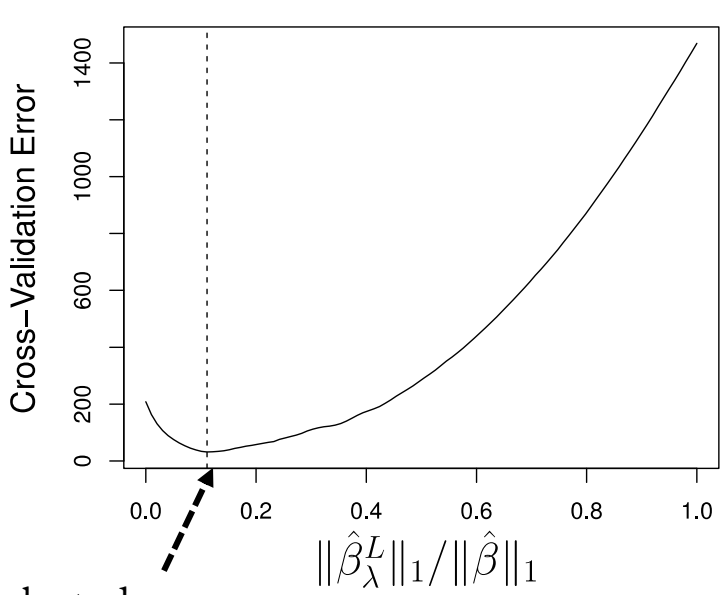
Choose λ by cross-validation

- The procedure is the same for ridge and lasso

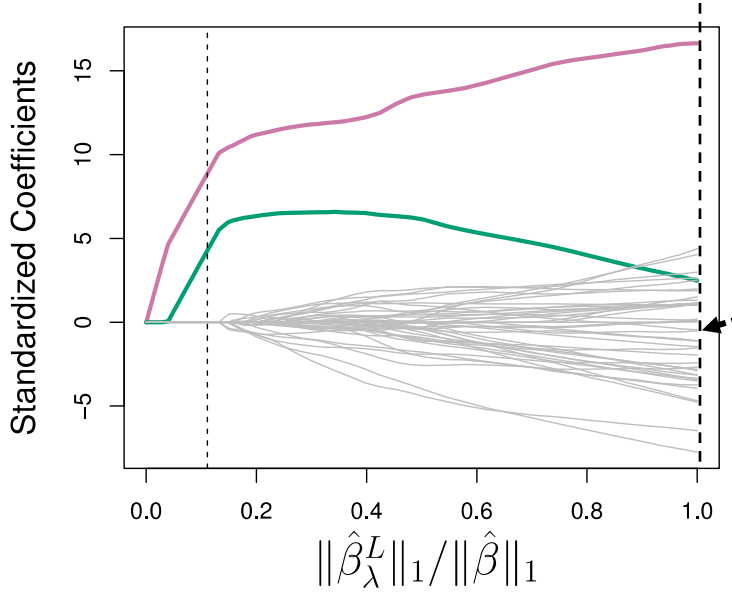
1. Choose a grid of λ values
2. Compute the cross-validation error for each λ value
3. Select the λ with the smallest cross-validation error
4. Refit the model using all observations and selected λ

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**



Optimal λ selected by cross-validation

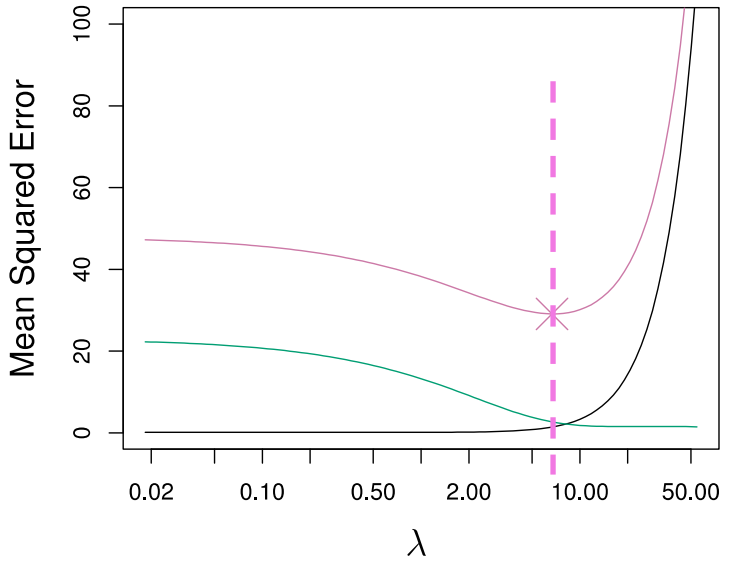


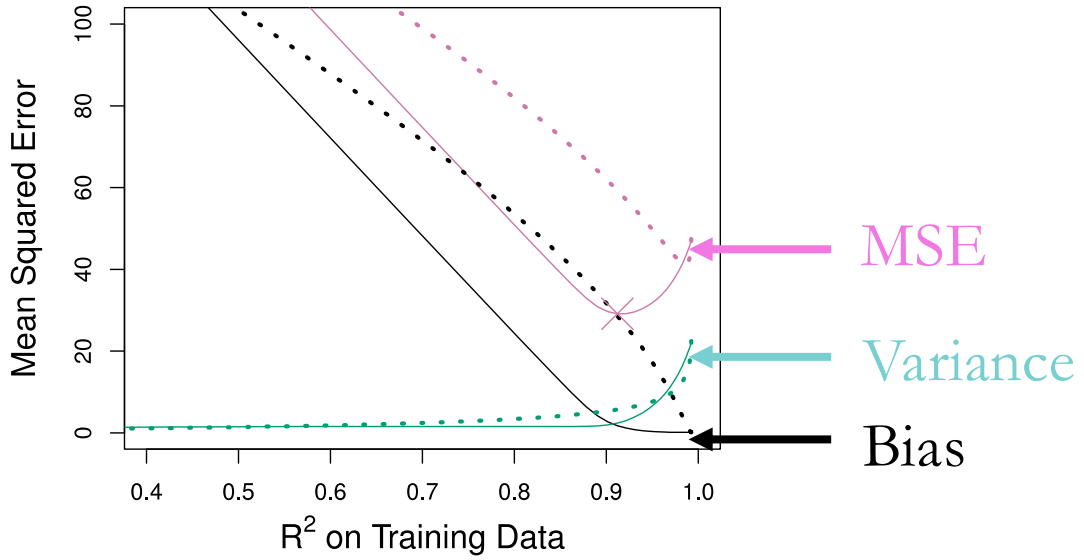
Linear regression solution

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



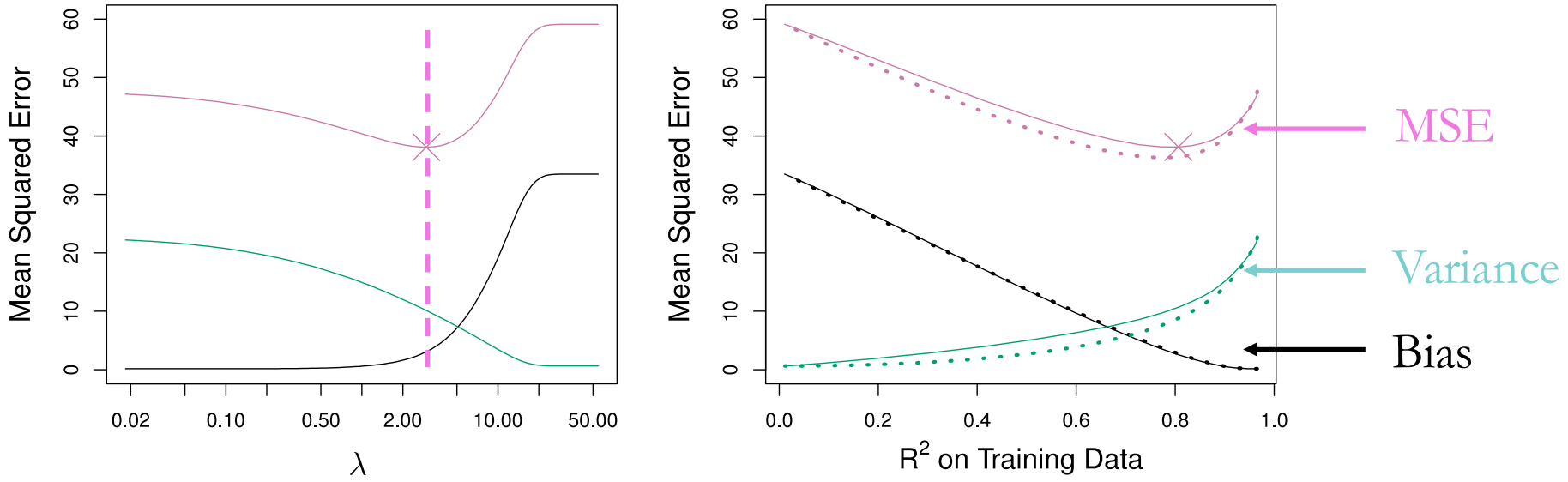


  - The **bias**, variance, and MSE are all lower for the lasso

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** ☺