QTM 347 Machine Learning

Lecture 10: Subset selection and regularization

Ruoxuan Xiong

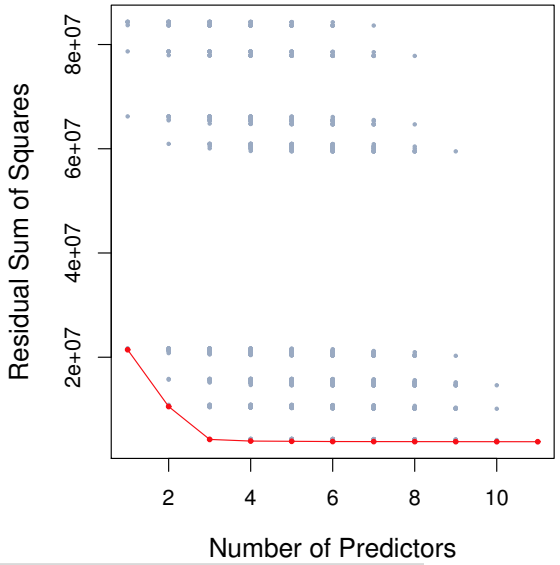
Suggested reading: ISL Chapter 6

Lecture plan

- Subset selection
- Ridge regression

Subset selection

- **Step 1**: For each k, select a subset of k predictors from the total p predictors
  - There are $\binom{p}{k} = \frac{p!}{k!(p-k)!}$ possible ways of choosing k predictors
  - Choose the subset with the smallest residual sum of squares (denoted by the red dots in the curve)
- **Step 2**: Use AIC, BIC or adjusted R^2 to select optimal k
  - Choose a dot in the red curve



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

- Step 1: No predictors (fit one model)
- Step 2: Select the best model with one predictor (fit p models)
- Step 3: Given the model with one predictor, select the best model with two predictors (fit $p - 1$ models)
- Step 4: Given the model with two predictors, select the best model with three predictors (fit $p - 2$ models)
- ...
- In each step, best is defined as having smallest RSS/MSE or highest R^2
- Select a single best model with the optimal number of predictors using cross-validated predictor error, AIC, BIC or adjusted R^2

Number of model fits in forward selection

- Step 1: No predictors (fit one model)
- Step 2: Select the best model with one predictor (fit p models)
- Step 3: Given the model with one predictor, select the best model with two predictors (fit $p - 1$ models)
- Step 4: Given the model with two predictors, select the best model with three predictors (fit $p - 2$ models)
- ...
- Fit $1 + p + (p - 1) + \cdots + 1 = 1 + \sum_{k=0}^{p-1}(p - k) = 1 + \frac{p(p+1)}{2}$ models in total
- Much fewer than $\sum_{k=0}^{p} \binom{p}{k} = 2^p$ in best subset selection

Forward vs. best subset selection

- Forward stepwise selection may fail to select the best possible k-variable model
- Forward stepwise selection is applicable to **high-dimensional settings** ($p > n$)

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income	rating, income,
	student, limit	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.*

Backward stepwise selection

- Step 1: All predictors (fit one model)
- Step 2: Select the best model with $p-1$ predictors (fit p models)
- Step 3: Given the model with $p-1$ predictors, select the best model with $p-2$ predictors (fit $p-1$ models)
- Step 4: Given the model with $p-2$ predictors, select the best model with $p-3$ predictors (fit $p-2$ models)
- ...
- In each step, best is defined as having smallest RSS/MSE or highest R^2
- Select a single best model with the optimal number of predictors using cross-validated predictor error, AIC, BIC or adjusted R^2

Number of model fits in backward selection

- Step 1: All predictors (fit one model)
- Step 2: Select the best model with $p-1$ predictors (fit p models)
- Step 3: Given the model with $p-1$ predictors, select the best model with $p-2$ predictors (fit $p-1$ models)
- Step 4: Given the model with $p-2$ predictors, select the best model with $p-3$ predictors (fit $p-2$ models)
- ...
- Fit $1+p+(p-1)+\cdots+1=1+\sum_{k=0}^{p-1}(p-k)=1+\frac{p(p+1)}{2}$ models in total
- Much fewer than $\sum_{k=0}^{p}\binom{p}{k}=2^p$ in best subset selection

Forward vs. backward selection

- You cannot apply backward selection when $p > n$
- Although it seems like they should, they need not produce the same sequence of models
- *Example.* $X_1, X_2 \sim N(0, \sigma^2)$ independent

$$X_3 = X_1 + 3X_2$$
$$Y = X_1 + 2X_2 + \epsilon$$

- Regress Y on X_1, X_2, X_3
  - Forward: $\{X_3\} \rightarrow \{X_3, X_2\} \rightarrow \{X_3, X_2, X_1\}$
  - Backward: $\{X_1, X_2, X_3\} \rightarrow \{X_1, X_2\} \rightarrow \{X_2\}$

Lecture plan

- Subset selection

- Ridge regression

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 ($n = 1$)

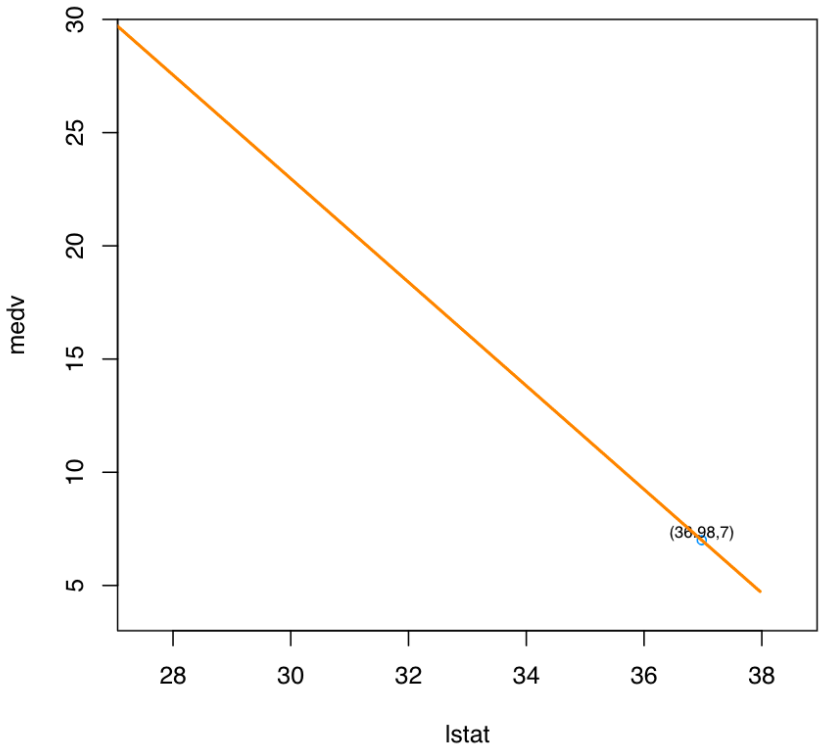
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
45.7461	0	18.1	0	0.693	4.519	100	1.6582	24	666	20.2	36.98	7

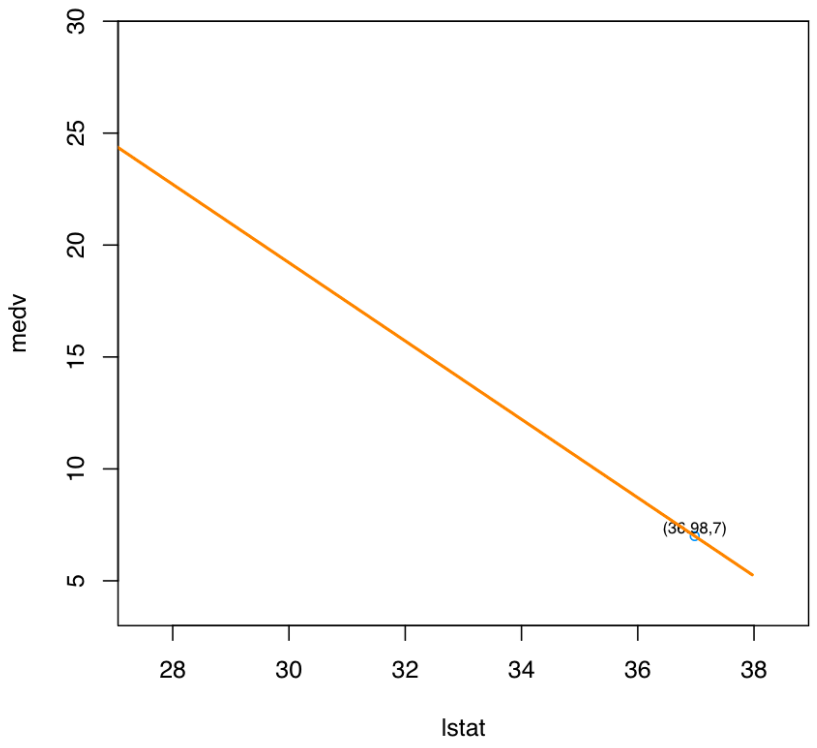
- But we want to estimate the coefficients in the linear model ($p = 2$)

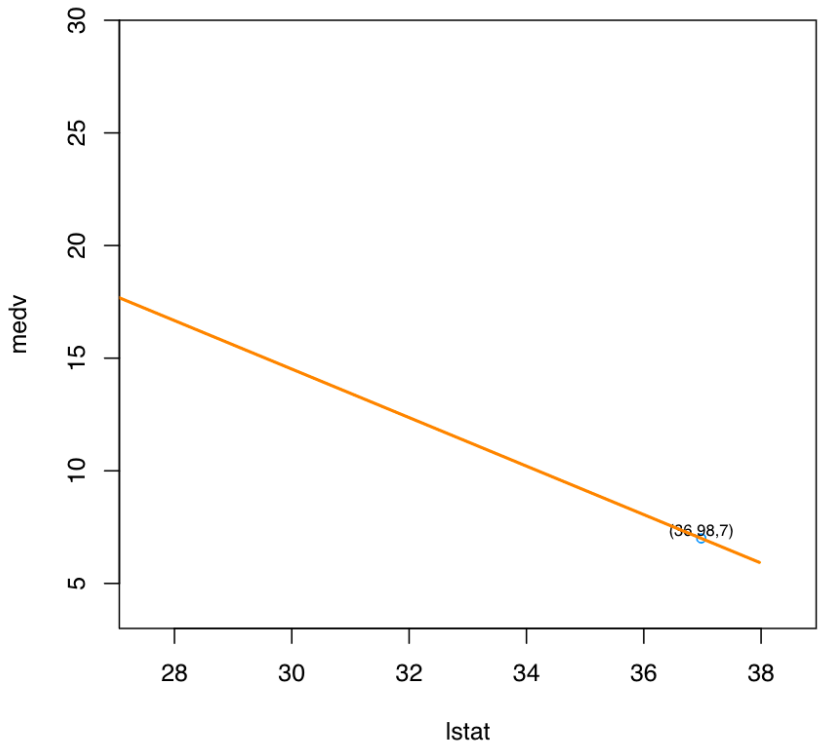
$$medv = \beta_0 + lstat \cdot \beta_1 + \varepsilon$$

- 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 ($n = 2$)

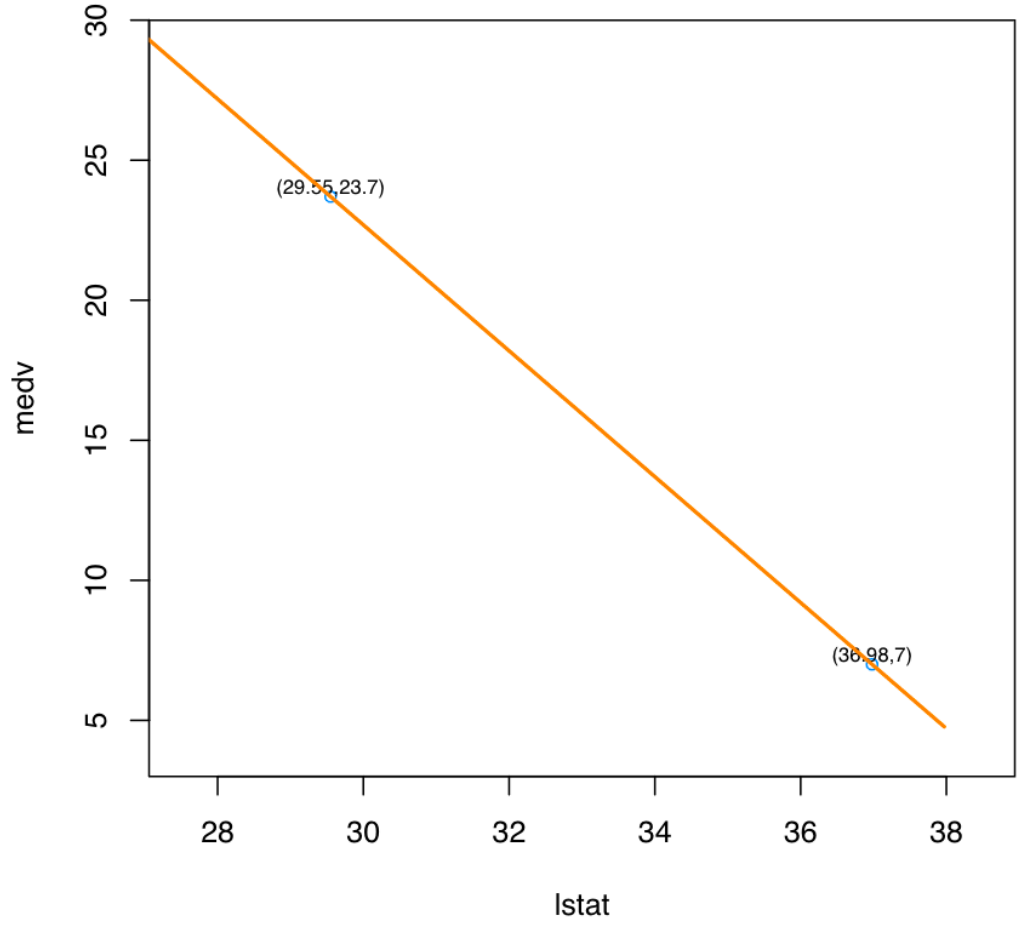
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
0.28955	0	10.59	0	0.489	5.412	9.8	3.5875	4	277	18.6	29.55	23.7
45.74610	0	18.10	0	0.693	4.519	100.0	1.6582	24	666	20.2	36.98	7.0

- 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

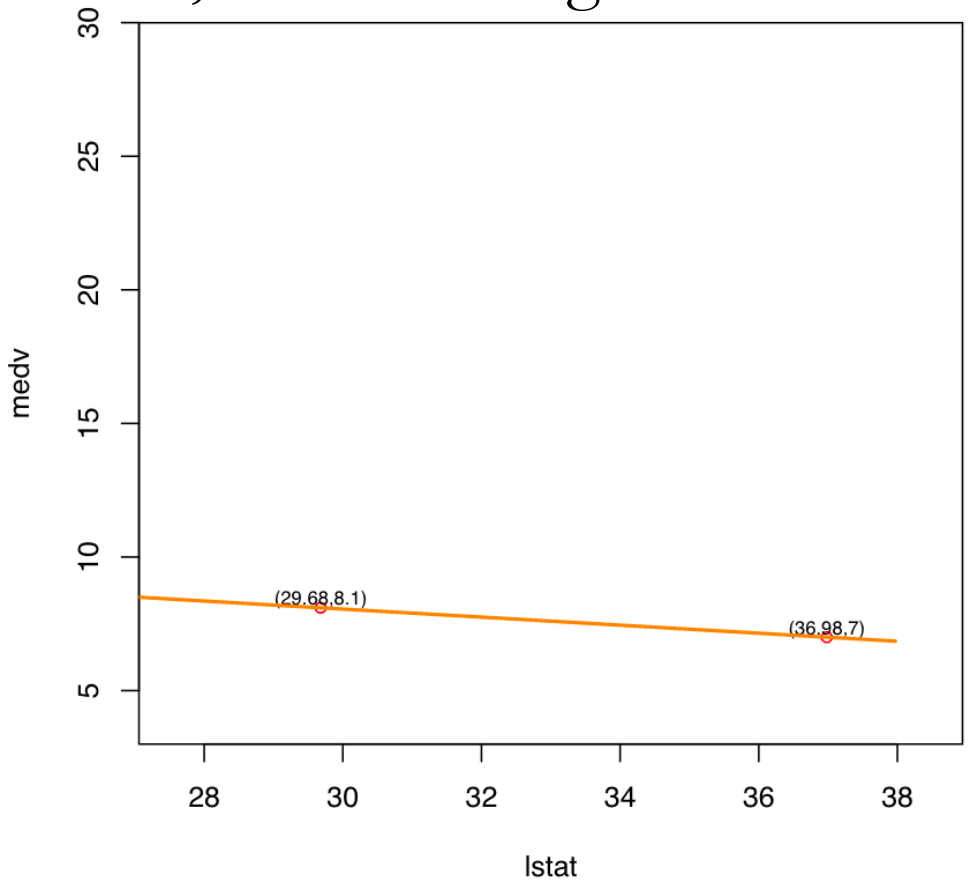
Example



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

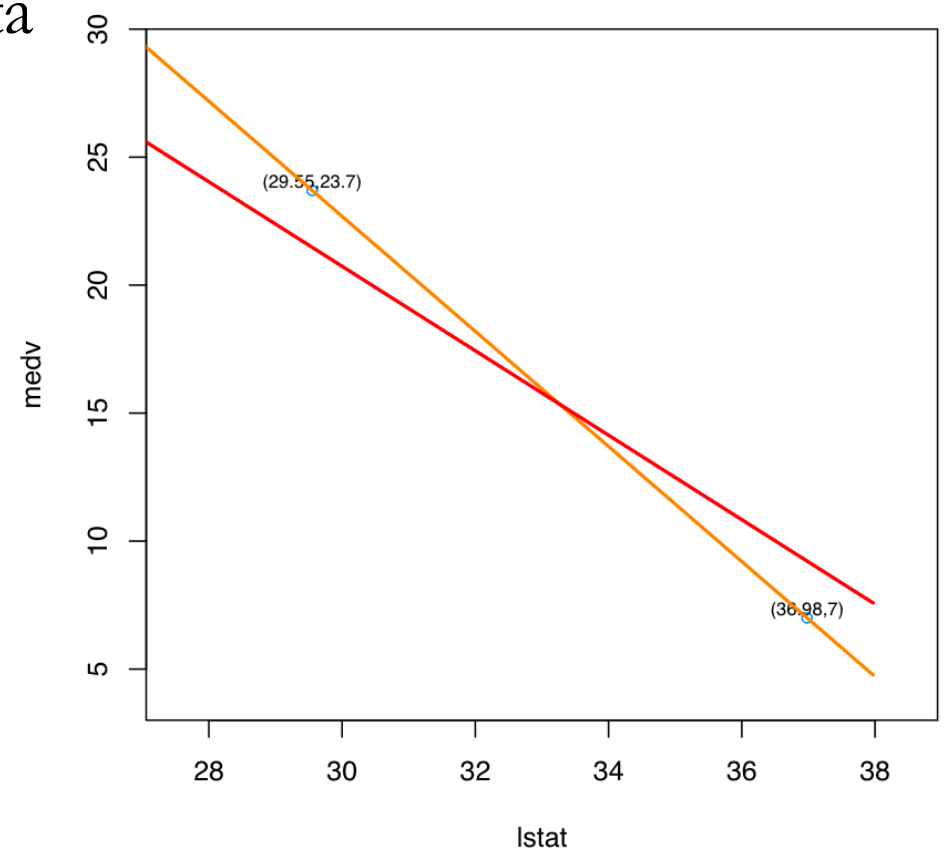
Example

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



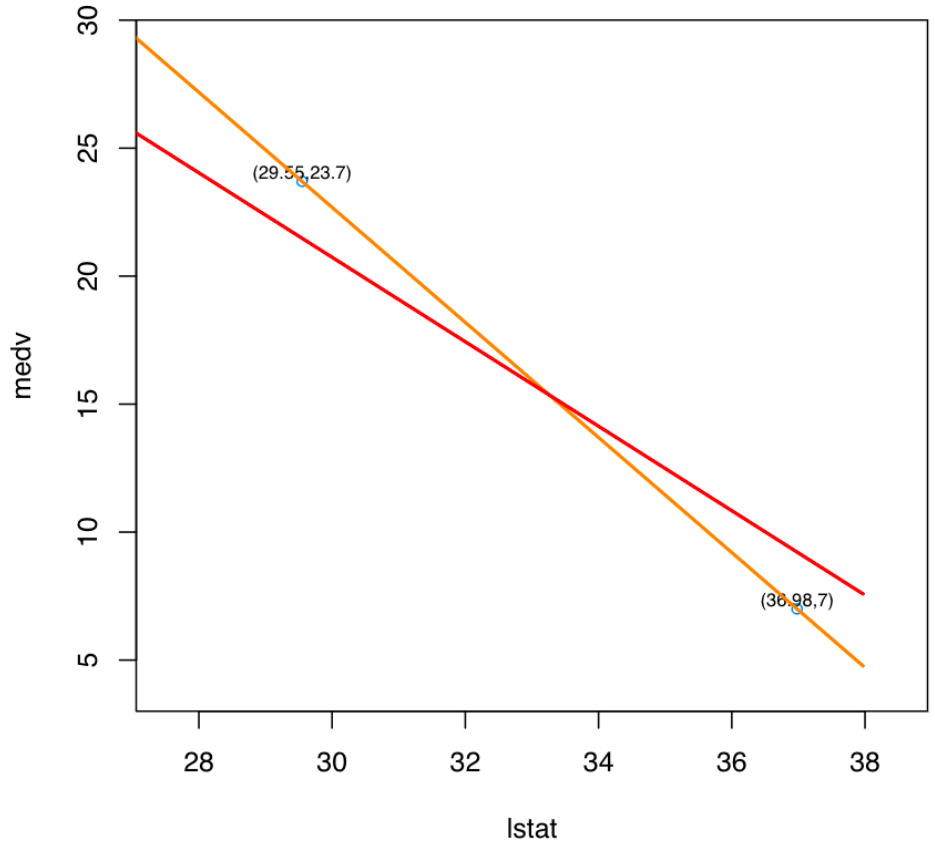
Ridge regression

- 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**

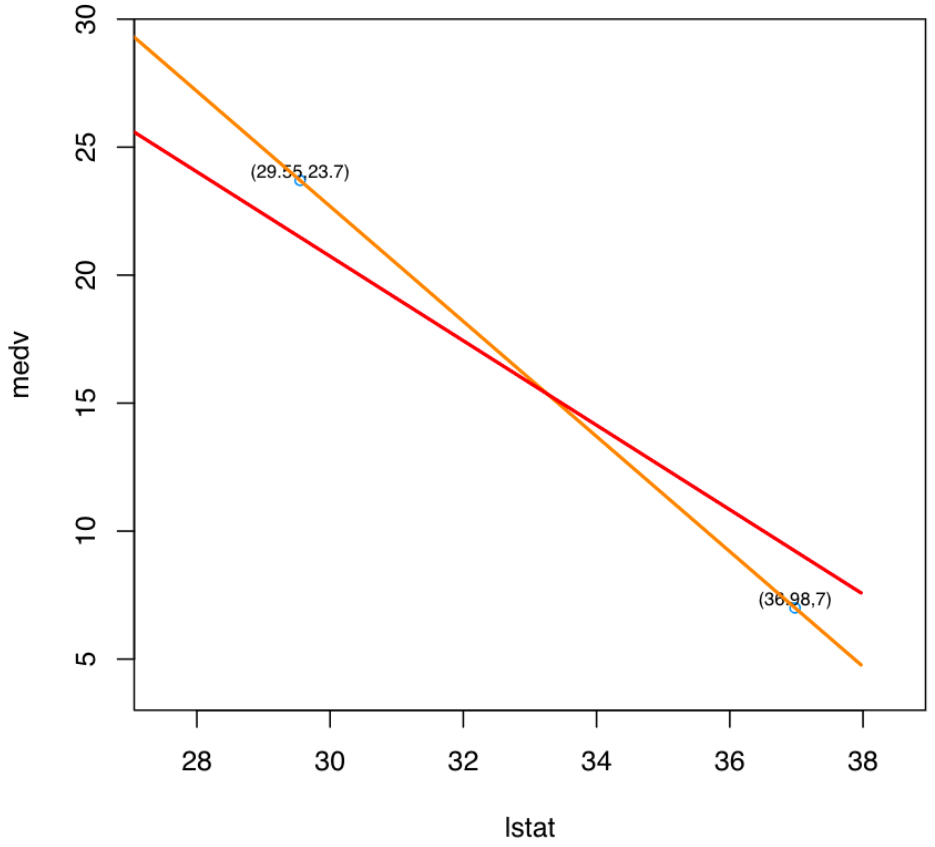


Fitting ridge regression

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

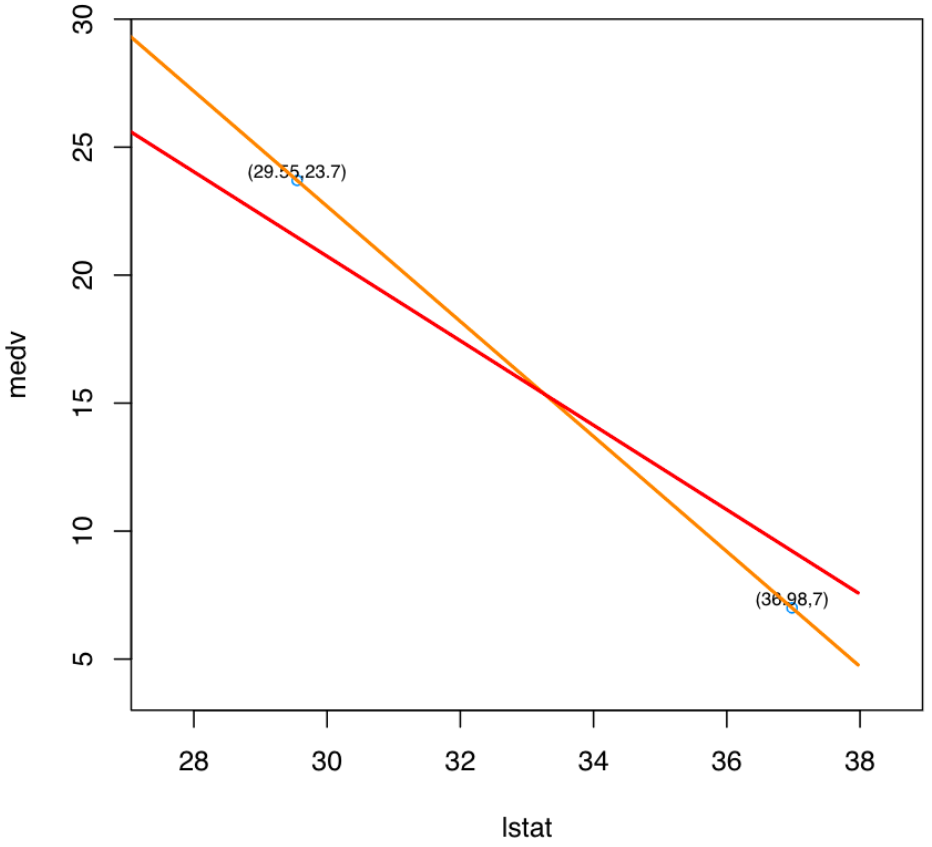
- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$

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



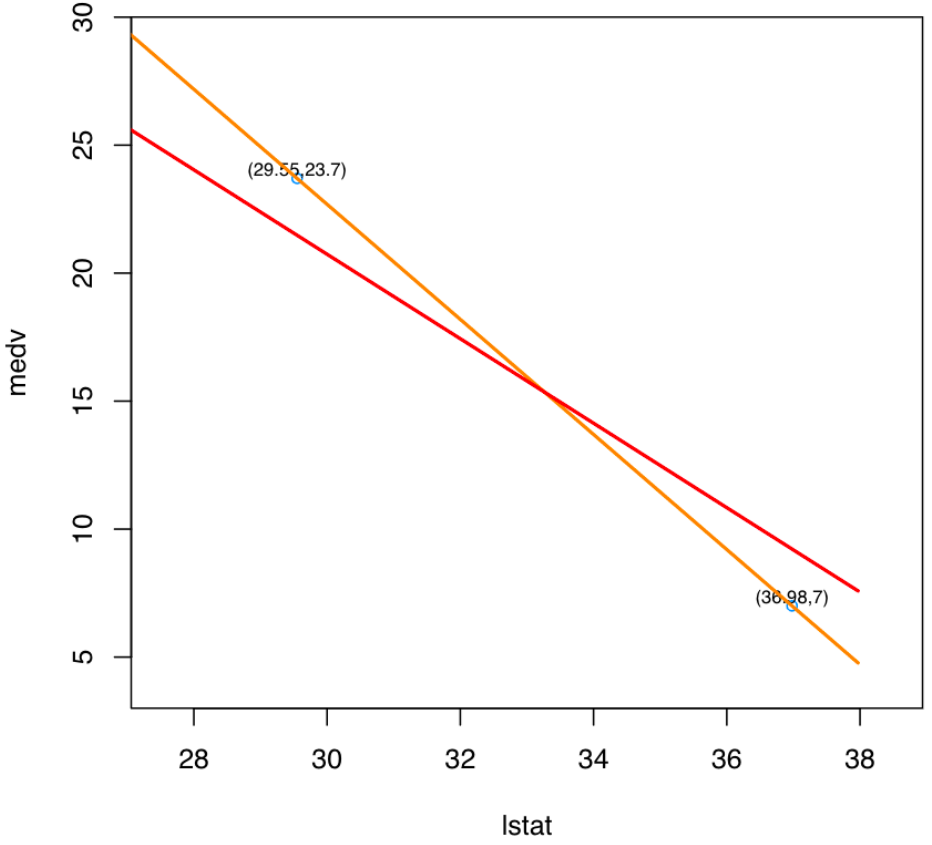
Objective value of least squares solution

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



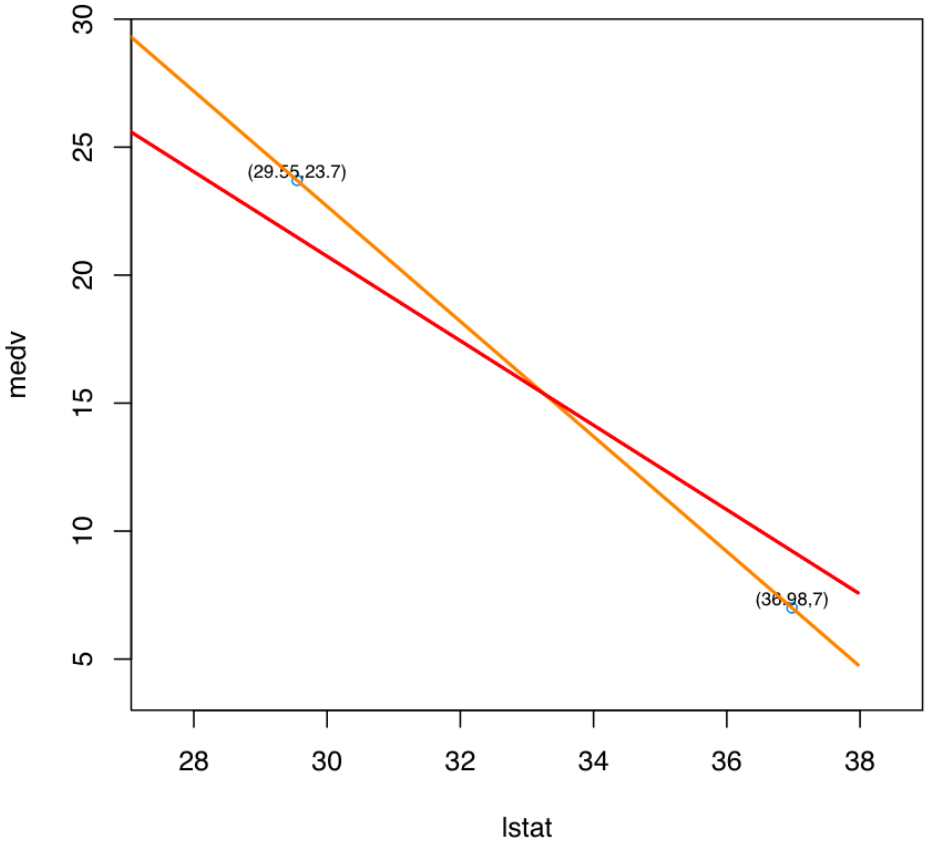
Objective value of ridge regression solution

- 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



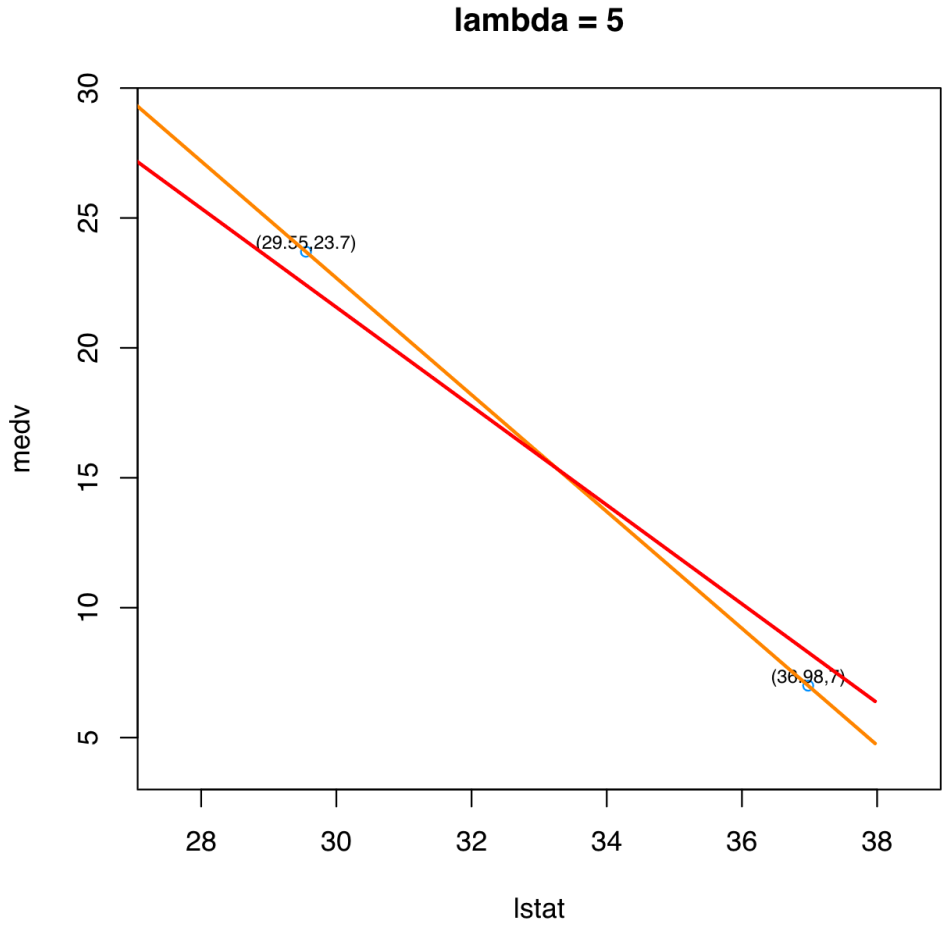
Ridge regression is less sensitive to $lstat$

- 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 $medv$



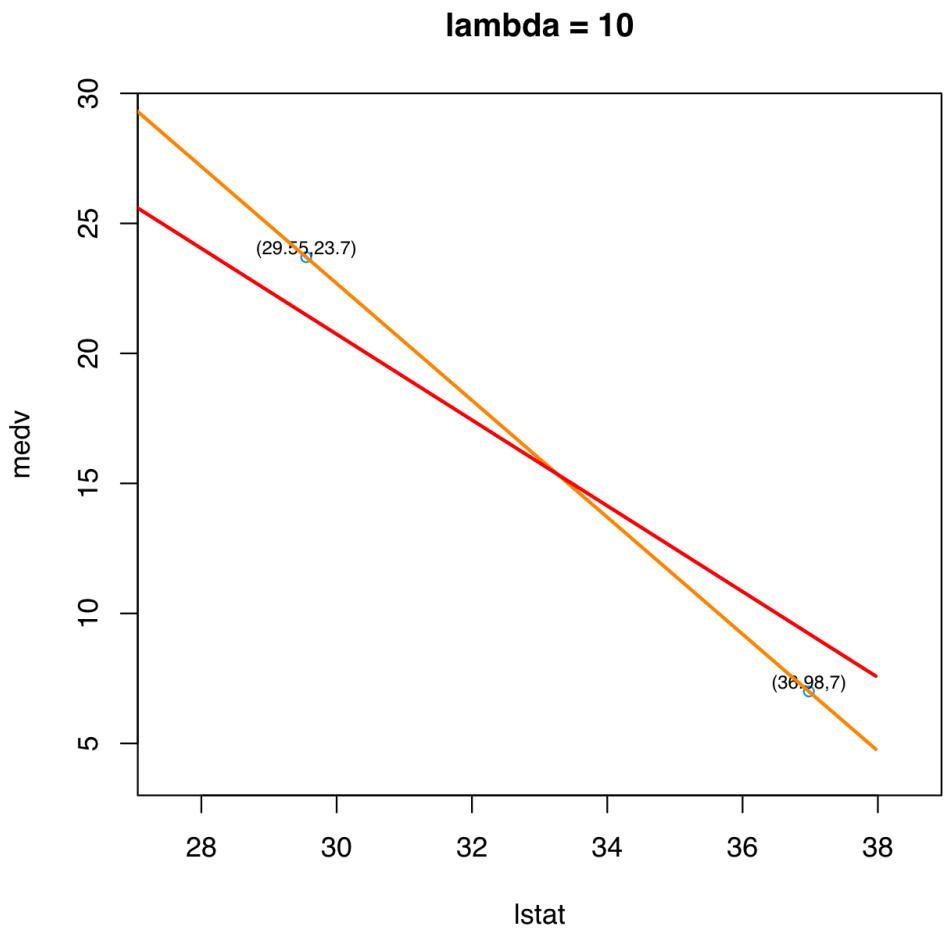
Role of λ in ridge regression

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



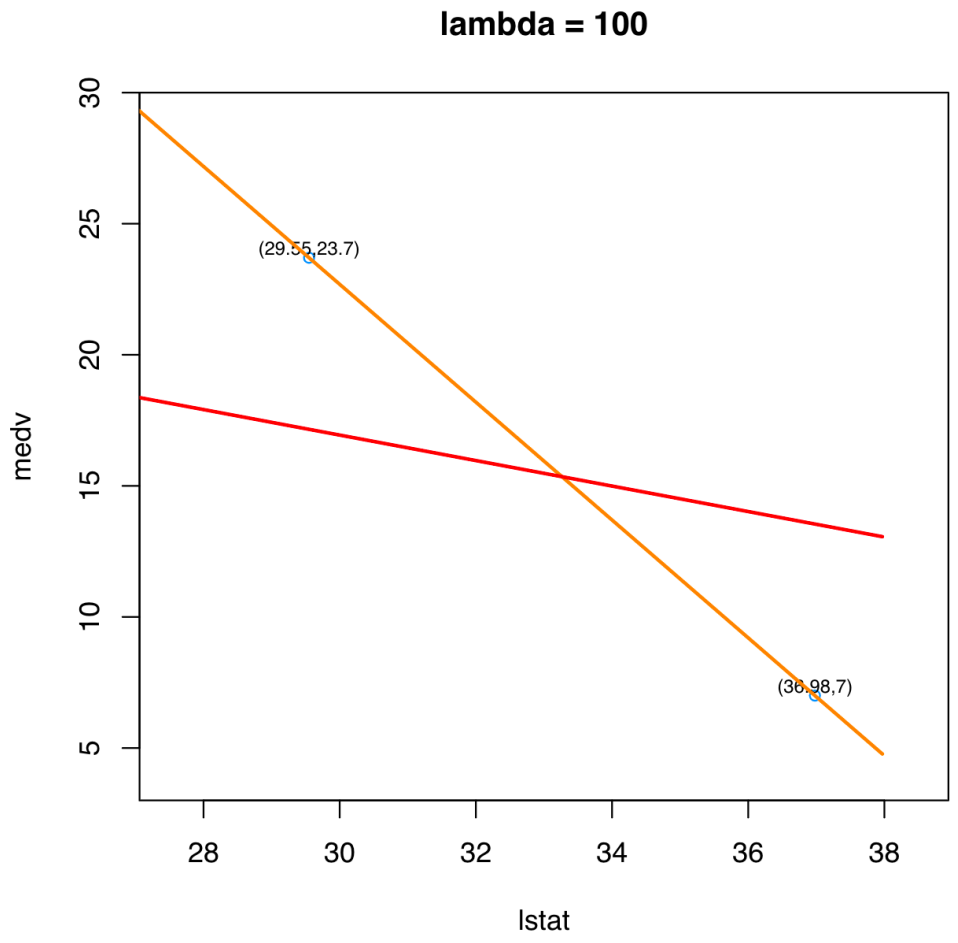
Role of λ in ridge regression

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



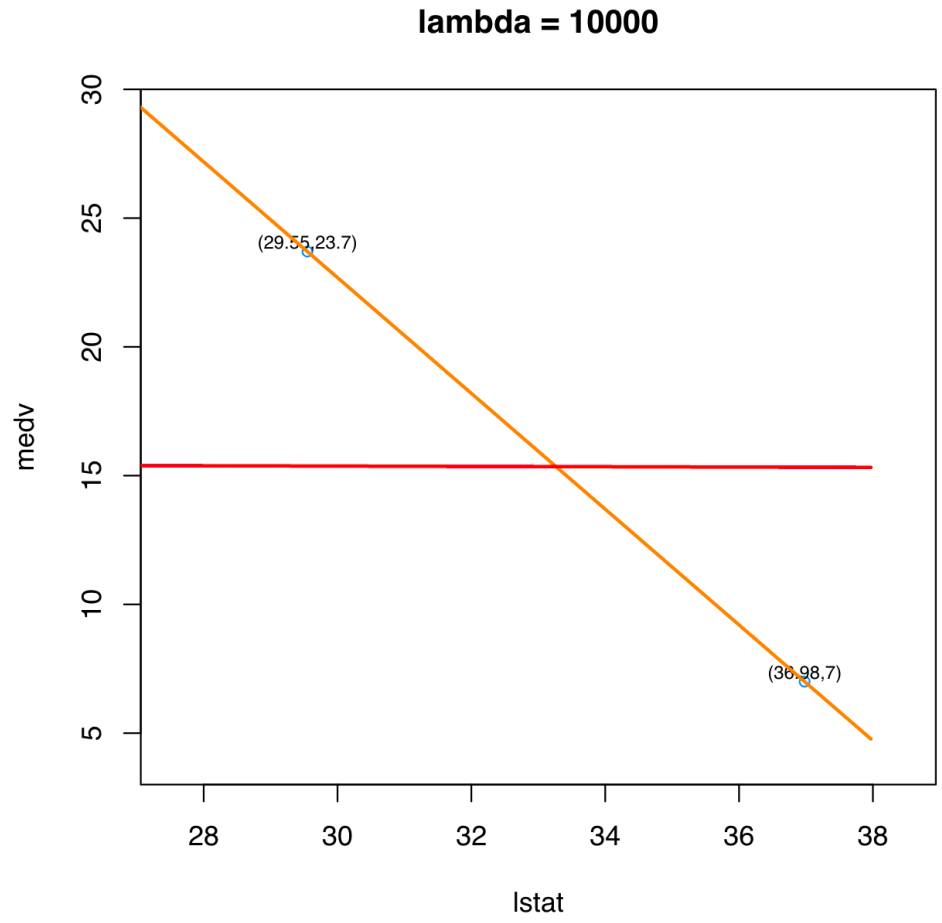
Role of λ in ridge regression

- 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

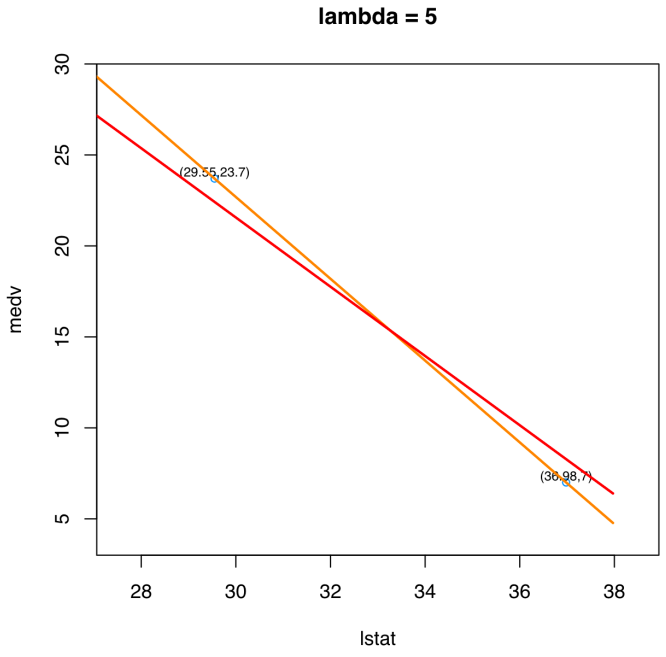
- Ridge regression minimizes
  - $\sum_{i=1}^{n}(medv_i - \beta_0 - lstat_i \cdot \beta_1)^2 + \lambda \cdot \beta_1^2$

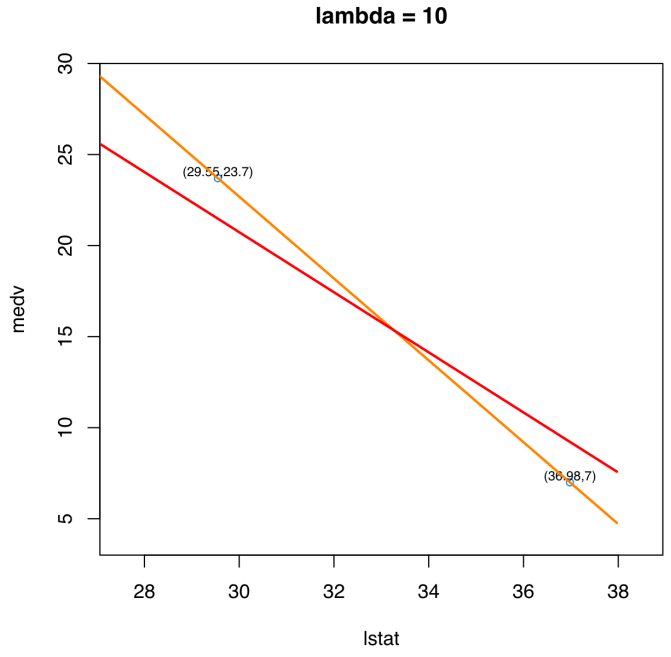


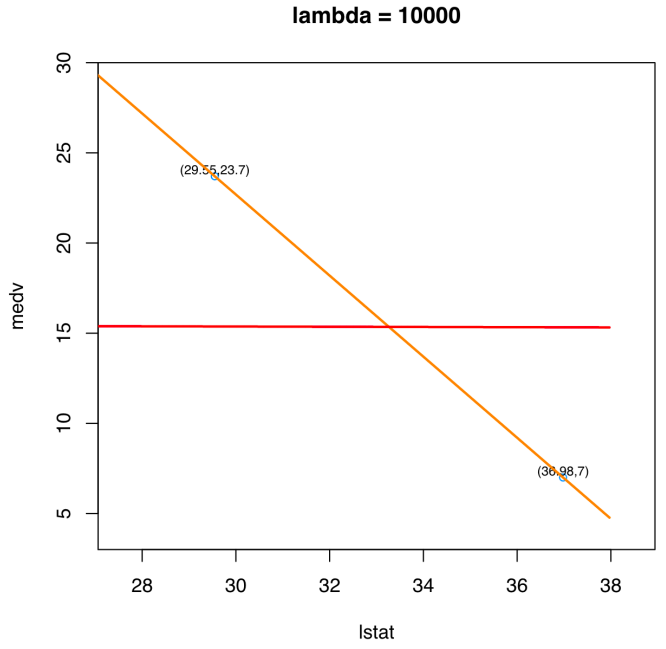
Our prediction becomes less sensitive to $lstat$ as λ increases

- 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

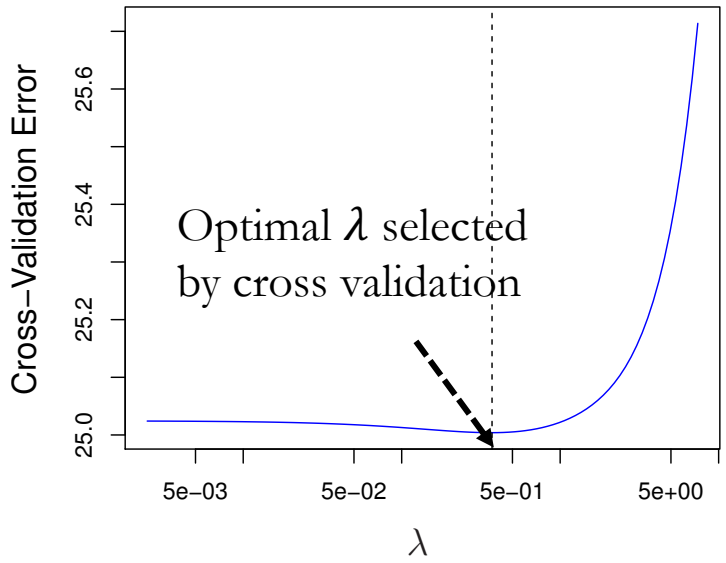
- Ridge regression minimizes

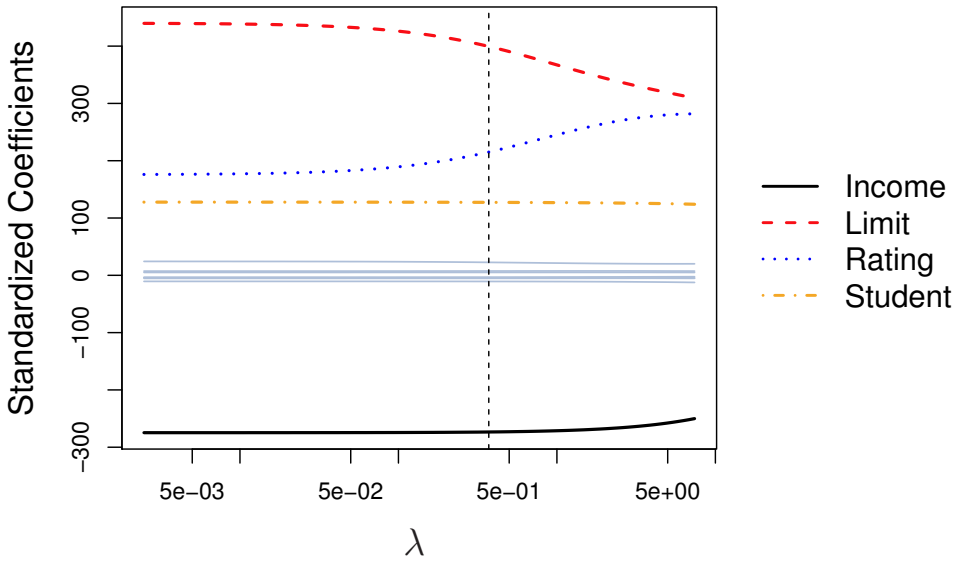
$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{i,j}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

  - $X_{i,j}$: 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