Stat 435 Intro to Statistical Machine Learning

Week 4: Bootstrap and Subset Selection

Richard Li

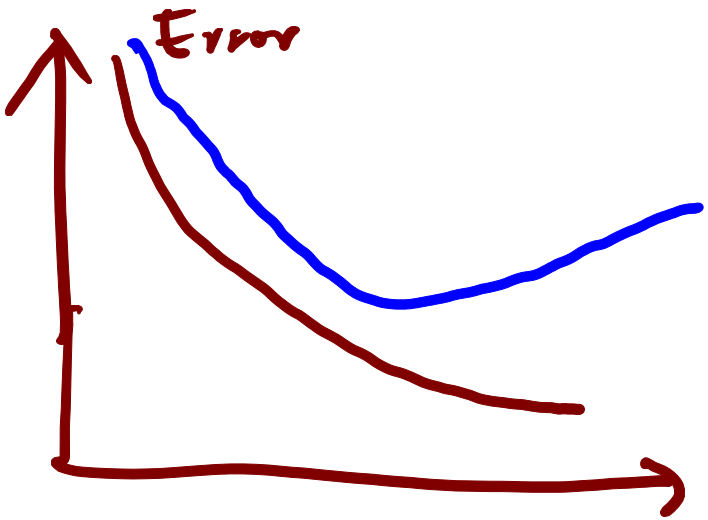
April 19, 2017

Today

- First half: Finish Chapter 5 Resampling
  - Bootstrap
- Second half: Beginning Chapter 6
  - Subset selection
  - Model selection criterion

How we get here

Training vs test error



Tools

Regression

Classification (LR, KNN, LDA, QDA)

minimize training error given a model

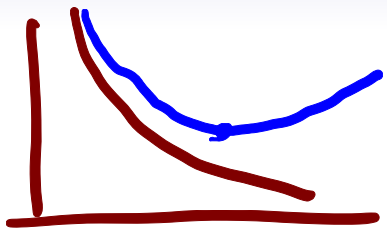
Goal: min testing error

Ideal: have a large test data

Reality: no test data

Where we go from here

Estimating test error when there's no test data



↓ Resampling

- C-V

↓ mathematical adjustments to the training error

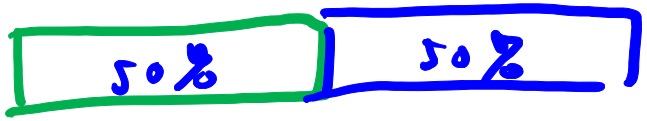
(Cp, ...)

A new task: measuring accuracy of parameters

- Bootstrap

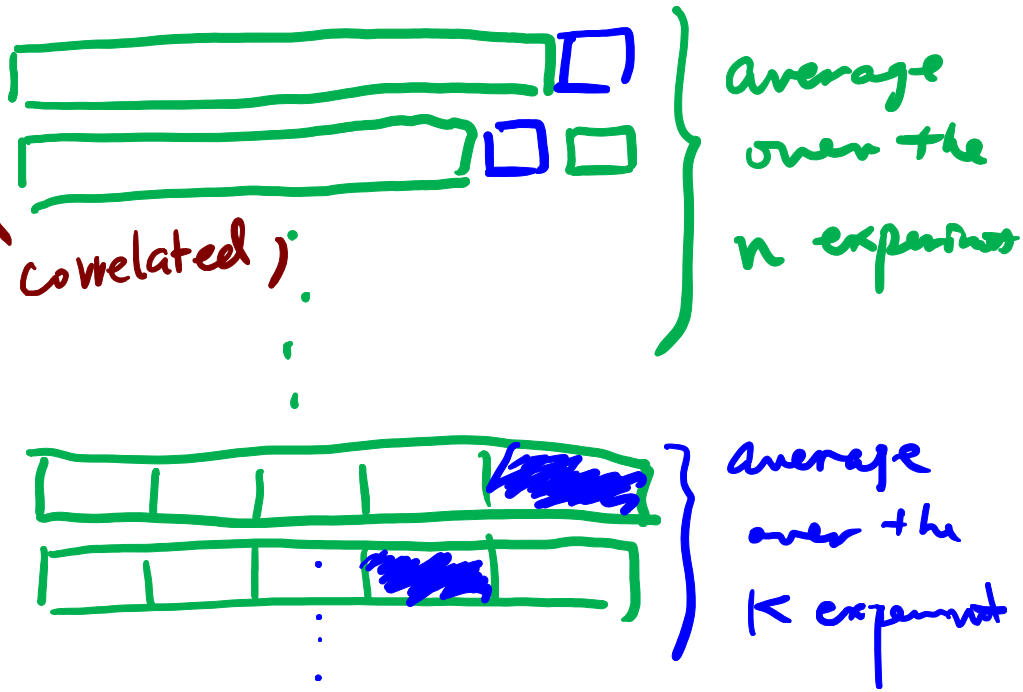
Recall Cross-validation

$Var(a+b) = var(a) + var(b) + 2cov(a,b)$



- Validation-set approach
  - only 50% traing
  - Test error over-estimated

- LOOCV
  - Bias low
  - high variance ($\hat{f}$ correlated)



- K-fold Cross-validation

$K = 5, 10$

Bootstrap

- A quick detour from testing error.
- When you obtain some coefficient $\hat{\theta}$, what to do next (as a statistician)?

- Assess uncertainty of $\hat{\theta}$

- Get confidence limits

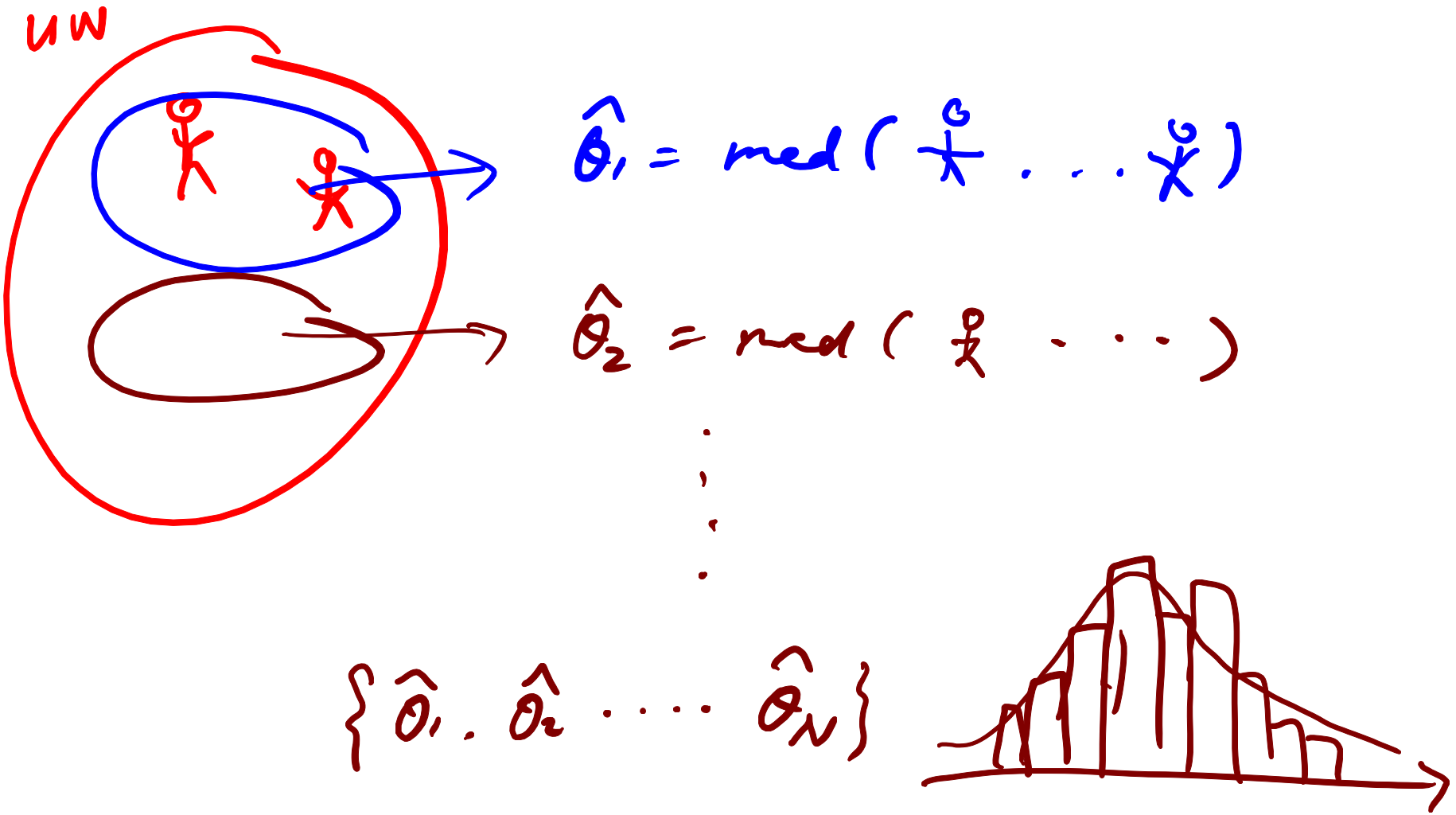
Bootstrap

- bootstrap: to pull oneself up by one's bootstraps

"The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps." –"The Surprising Adventures of Baron Munchausen" by Rudolf Erich Raspe

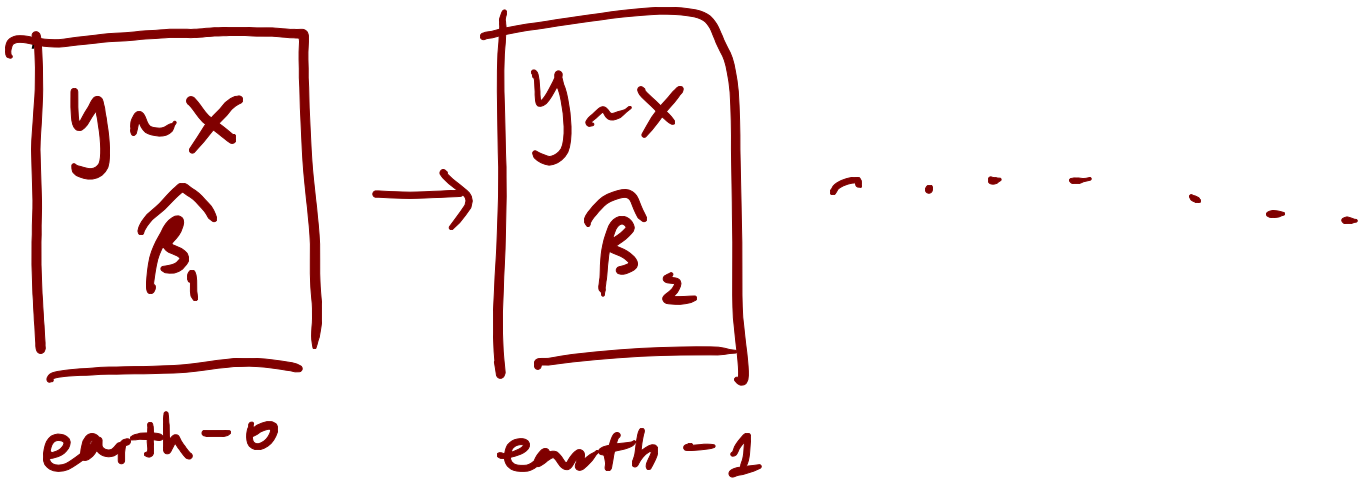
An ideal world where you don't need bootstrap

- $\hat{\theta}$: an estimator of median height of students attending UW



An ideal (meta)-world where you don't need bootstrap

- For practical problems, typically you don't have a large number of collections of samples.
- e.g., finding variance of a regression coefficient.



$\{\hat{\beta}_1 \ldots \hat{\beta}_N\} \Rightarrow$ Histogram

A real world where bootstrap is useful

- Idea: the observed data contains all the information about the distribution of the parameter of interest.
- A procedure that can be applied to a wide range of statistical methods where measure of variability could be hard to compute.

When it looked like all was lost

- A simple linear regression example: predict `mpg` using `horsepower`.
- We only have one dataset, how to estimate $sd(\hat{\beta}_1)$.

Parametric Approach

Assume $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ ⟶ what if ?

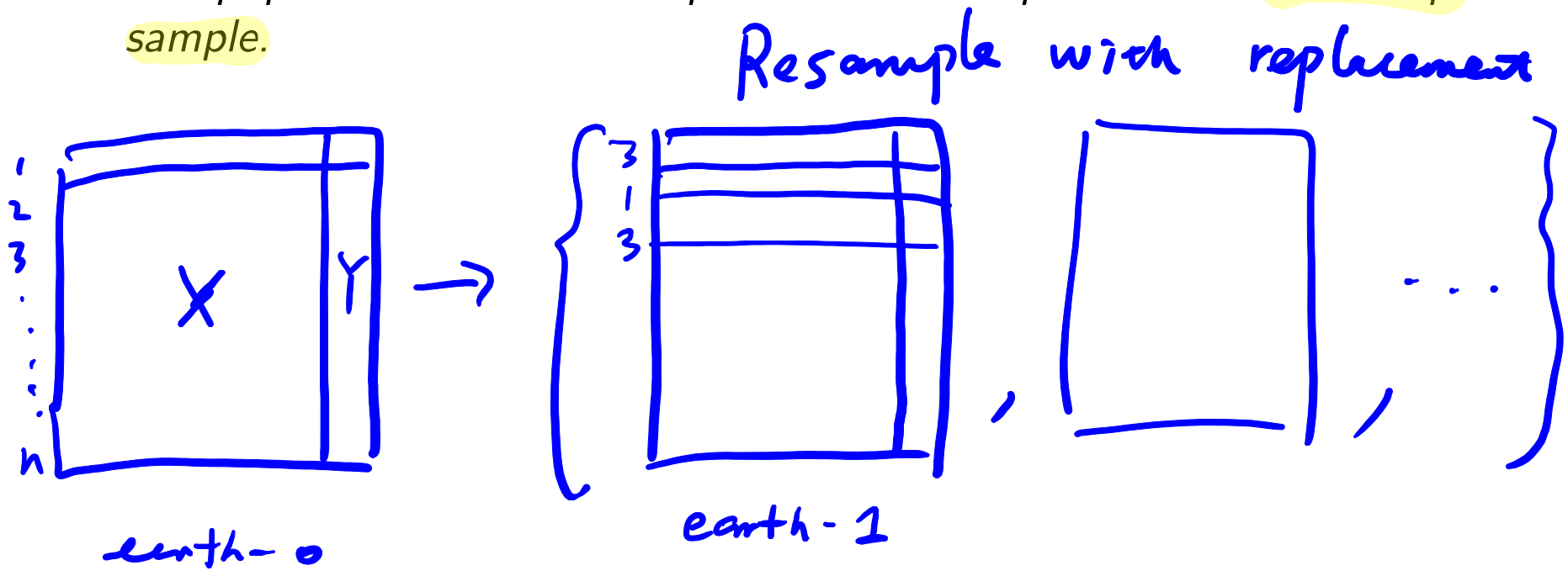
Derive $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$var(\hat{\beta}_1) = \cdots = \cdots \sigma^2$

$\hat{\sigma}^2 = \cdots$ plug in

Pick himself up by his own bootstraps

- Let's try create a multiverse.
- *The population is to the sample what the sample is to the bootstrap sample.*



What's going on

In a perfect multiverse,

- We generate samples from the original population many times.
- Each dataset of samples we generate are independent.

In a bootstrap-multiverse,

- We generate *bootstrap samples* from the sample many times.
- Each *bootstrap sample* are sampled with replacement and has the same size as the original sample.

Formal definition

Bootstrap on a dataset Z of size n

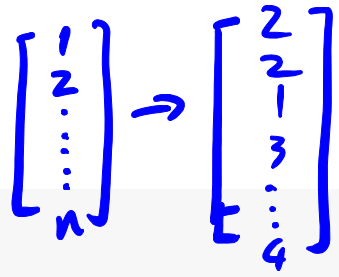
- Randomly select n observations in Z with replacement, call it Z^{*1}.
- Do analysis and calculate the estimator of interest θ^{*1}.
- Repeat these two steps B times, for some large B, and obtain
  - bootstrap datasets: Z^{*1}, Z^{*2}, ..., Z^{*B}.
  - bootstrap estimates: θ^{*1}, θ^{*2}, ..., θ^{*B}
- Bootstrap mean:

$$Mean_B(\hat{\theta}) = \frac{1}{B}\sum_{r}^{B}\hat{\theta}^{*r}$$

- Bootstrap standard error:

$$SE_B(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{r}^{B}(\hat{\theta}^{*r} - Mean_B(\hat{\theta}))^2}$$

Example




```
library(ISLR)
library(boot)
boot.fn=function(data,index){
  return(coef(lm(mpg~horsepower, data=data, subset=index)))
}
boot.fn(Auto ,1:392)

## (Intercept)  horsepower
##  39.9358610  -0.1578447

summary(lm(mpg~horsepower, data=Auto))$coef

##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 39.9358610 0.717498656  55.65984 1.220362e-187
## horsepower  -0.1578447 0.006445501 -24.48914  7.031989e-81
```

Example

```
boot.fn(Auto,sample(392,392,replace=T))

## (Intercept)  horsepower
##  39.4116717  -0.1547955

boot.fn(Auto,sample(392,392,replace=T))

## (Intercept)  horsepower
##  39.8142904  -0.1592939
```

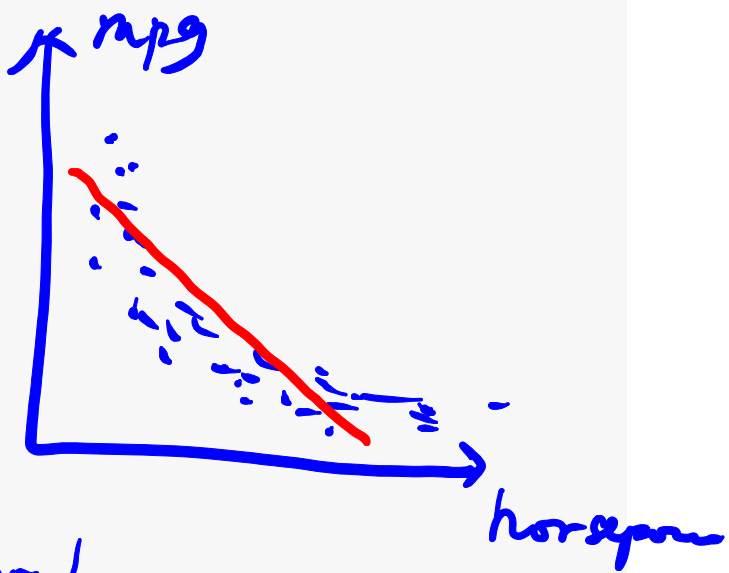
Example

```
boot(Auto, boot.fn, 1000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Auto, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##       original        bias    std. error
## t1* 39.9358610  0.0601108699 0.859624333
## t2* -0.1578447 -0.0005534737 0.007342598

summary(lm(mpg~horsepower, data=Auto))$coef

##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 39.9358610 0.717498656  55.65984 1.220362e-187
## horsepower  -0.1578447 0.006445501 -24.48914  7.031989e-81
```

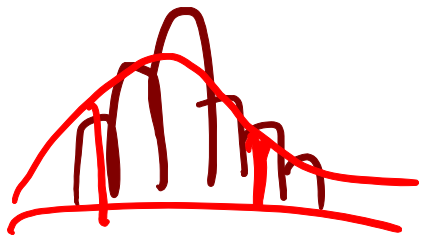


Better !

More in general

- In complex data situations, figuring out the appropriate way to generate bootstrap samples is not trivial.
- For example, for time series data, simply resample observations is a bad idea.

How bootstrap is used in practice



- Bootstrap standard error is the most common.
- Can be used to estimate confidence intervals: *Bootstrap Percentile* confidence interval.
- Can be used to estimate prediction error...
  - Difficult to use in practice, *why?*
  - *So, sticking to cross-validation is easier!*

training → bootstrap sample

validation → original sample.

testing error underestimated

problem: overlap.

Chapter 6: Model selection

Overview

Why we are still studying linear regression 4 weeks in?

- Simple but easy to interpret.
- Often good predictive performance.

But we can do something *alternative* to OLS.

Overview: Feature selection

Why?

- Prediction accuracy: especially when $p > n$.
- Interpretability: removing extra irrelevant features automatically.

Three classes of methods

- **Subset selection** (Today)
- **Shrinkage**
- **Dimension reduction**

Notice: all these methods can be extended outside of linear regression, e.g., logistic regression, etc.

Best subset selection

$X = n \times p$

$M_0 \quad y_i = a + \varepsilon_i$

$M_1 \quad y_i = a + b x_4 + \varepsilon_i$ (example)

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Best subset selection

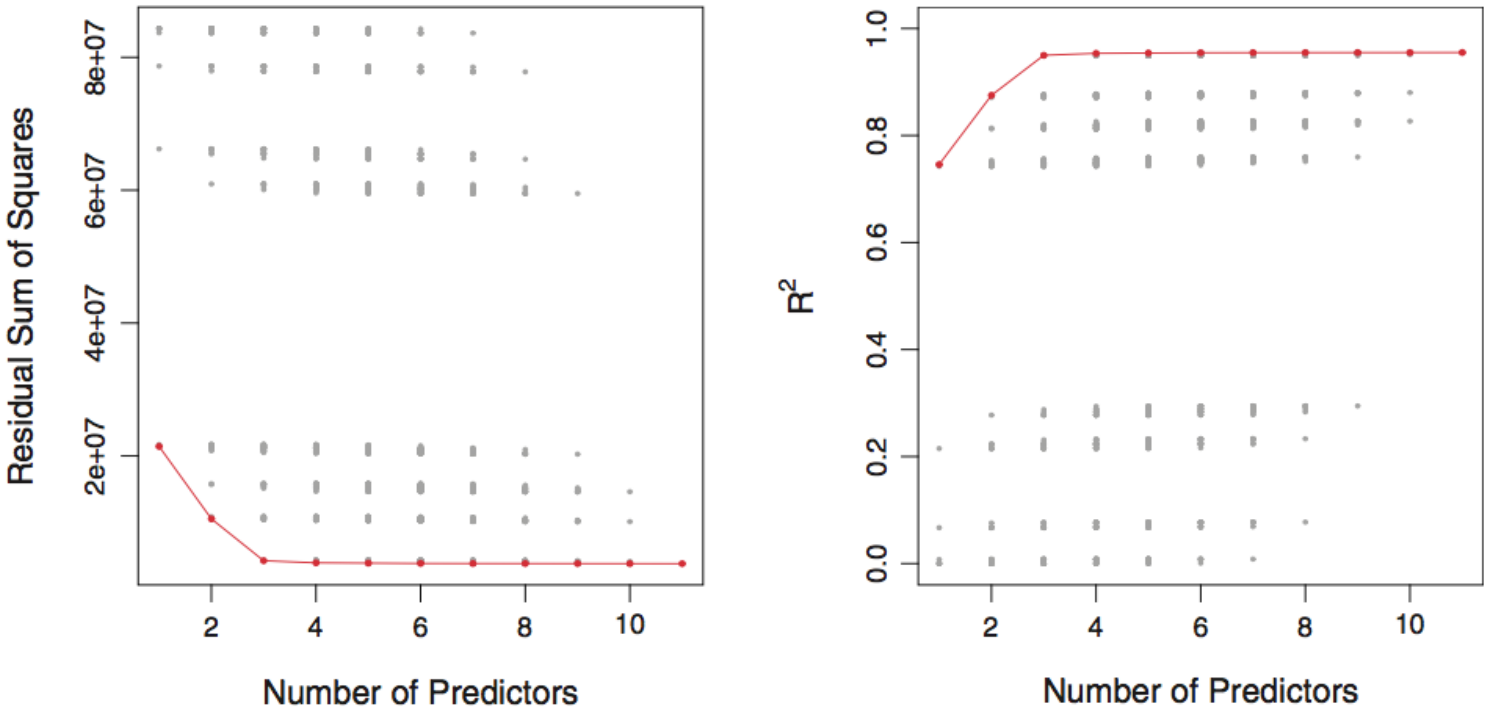


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Forward selection

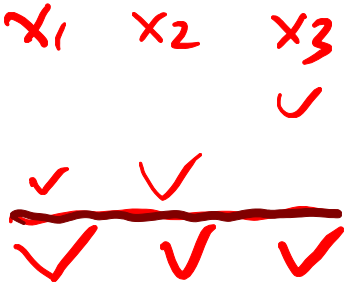
M_1 $y_i = \alpha + \beta x_4 + \varepsilon_i$

M_2 $y_i = \alpha + \beta_1 x_4 + \beta_2 x_{10} + \varepsilon_i$

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Not the absolute Best model $\Longrightarrow$



Backward selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Comparisons

- Best subset selection need to evaluate 2^p models.
- Step-wise selection evaluates $1 + p(p+1)/2$ models.
- Backward selection can only be applied if $n > p$.
- *Next we talk about criterion of model comparison*.

Choosing optimal model

- How to compare test error rate when there's no test data.
- A few approaches to approximate using training error
  - Mallow's C, C_p
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Adjusted R^2

C_p

- d: number of parameters
- Mallow's C_p:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

AIC

Lower the better

- Mallow's C_p:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- Akaike information criterion (AIC) is defined as

$$AIC = -2\log L + 2d$$

- Equivalent to C_p in linear models (not in other models!)

$$-2\log L = \frac{RSS}{\hat{\sigma}^2}$$

BIC

Lower the better

- Bayesian information criterion (BIC) is defined as

$$BIC = \frac{1}{n}(RSS + \log(n) d \hat{\sigma}^2)$$

- For C_p, AIC and BIC, we want small values!

$\log(n) > 2$

$\Rightarrow n > 7$

BIC penalize large models more than AIC/Cp

Adjusted R^2

higher the better

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$\text{Adjusted}\, R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

↘ as d ↗

$\frac{RSS}{n-d-1}$ ↘

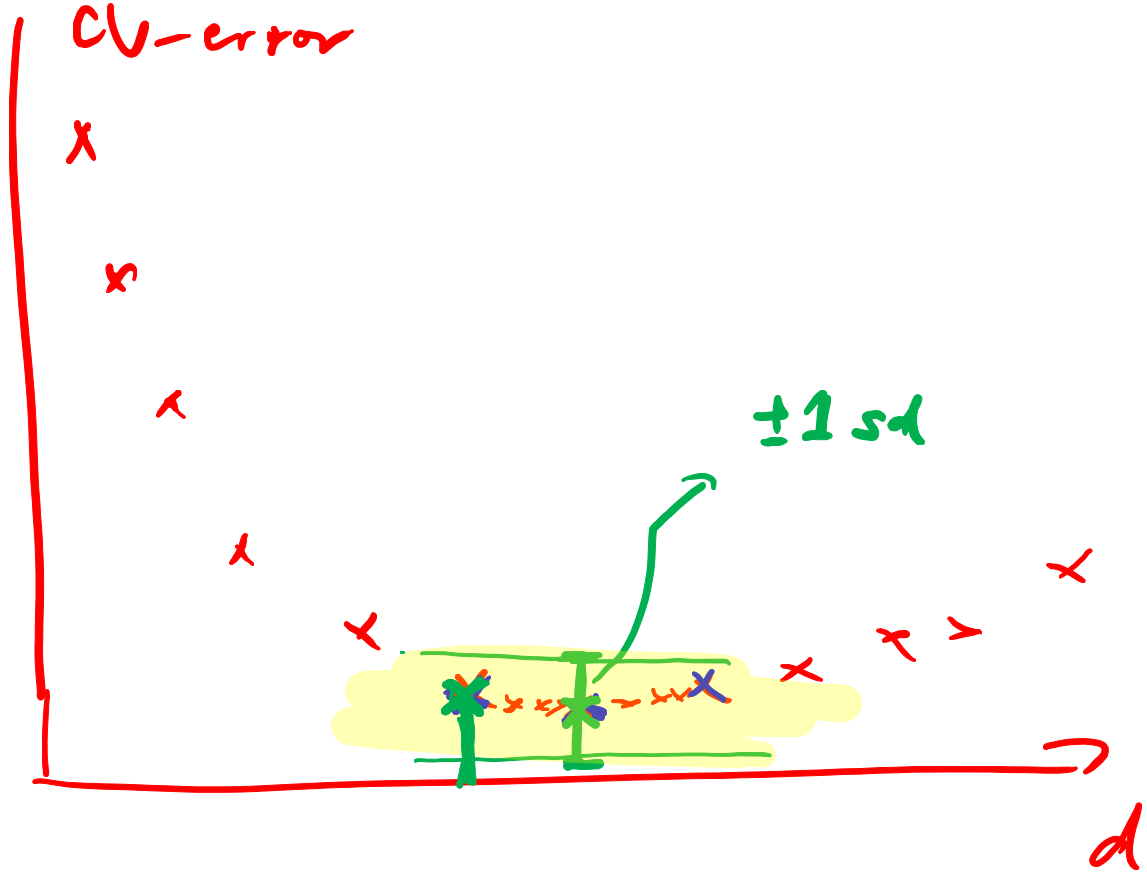
- We want adjusted R^2 to be large.
- It does not need $\hat{\sigma}^2$, so it can be used for $p > n$!
- However, difficult to generalize to other models.

Final thoughts

- Remember when we first talk about estimating test error, we can do cross-validation.
- Cross-validation has advantage since it directly estimate test error.
- It does not need to estimate σ^2, which can be difficult to calculate.
- Also model size d is not trivial in many settings. *Seems silly?*

A final final thought

One-standard error rule



Predict baseball player's salary

```
library(ISLR)
data(Hitters)
names(Hitters)
```

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"     "Years"
##  [8] "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CRBI"      "CWalks"    "League"
## [15] "Division"  "PutOuts"   "Assists"   "Errors"    "Salary"    "NewLeague"
```

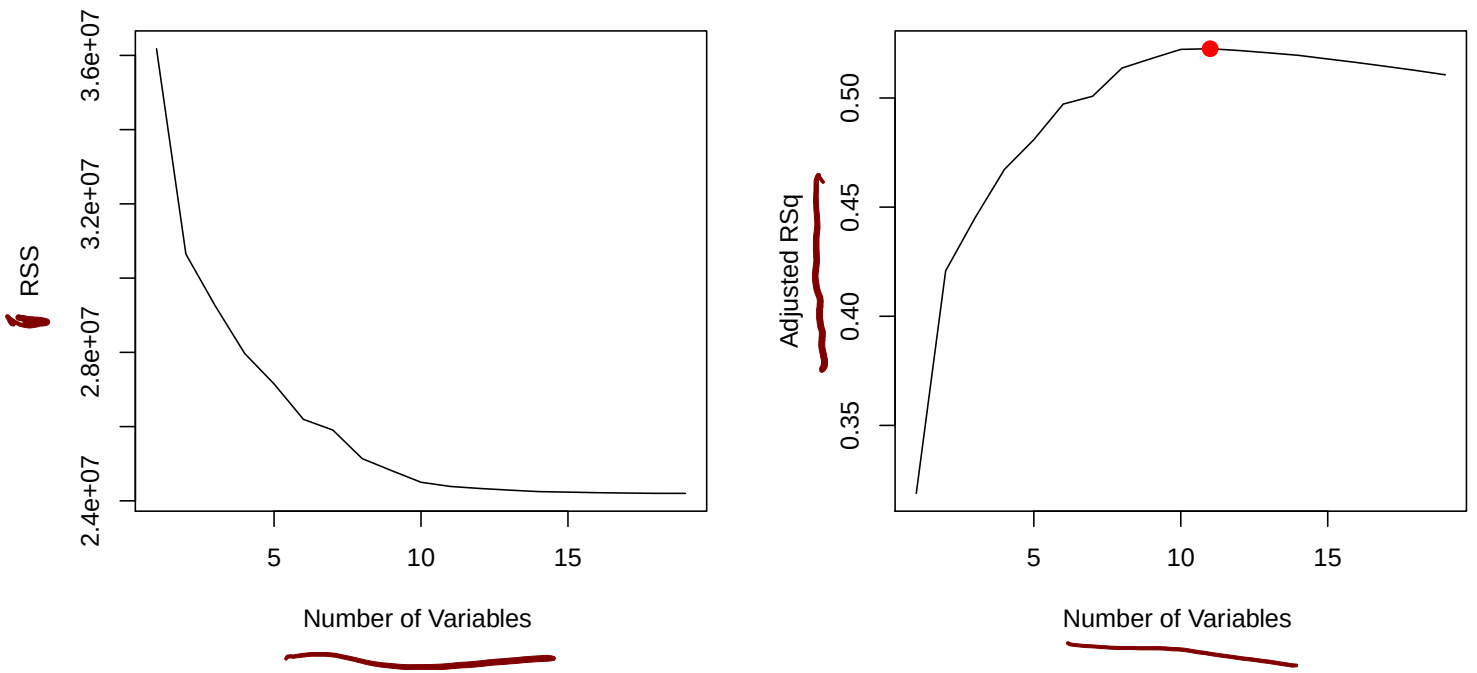
Best subset regression

```
library(leaps)
regfit=regsubsets(Salary~.,Hitters)
summary(regfit)$outmat
```

```
##          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
## 1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 4  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 5  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*"  " "
## 7  ( 1 ) " "   "*"  " "   " "  " " "*"   " "   "*"    "*"   "*"    " "   " "  " "
## 8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   "*"    "*"   " "  "*"
##          LeagueN DivisionW PutOuts Assists Errors NewLeagueN
## 1  ( 1 ) " "     " "       " "     " "     " "    " "
## 2  ( 1 ) " "     " "       " "     " "     " "    " "
## 3  ( 1 ) " "     " "       "*"     " "     " "    " "
## 4  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 5  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 6  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 7  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 8  ( 1 ) " "     "*"       "*"     " "     " "    " "
```

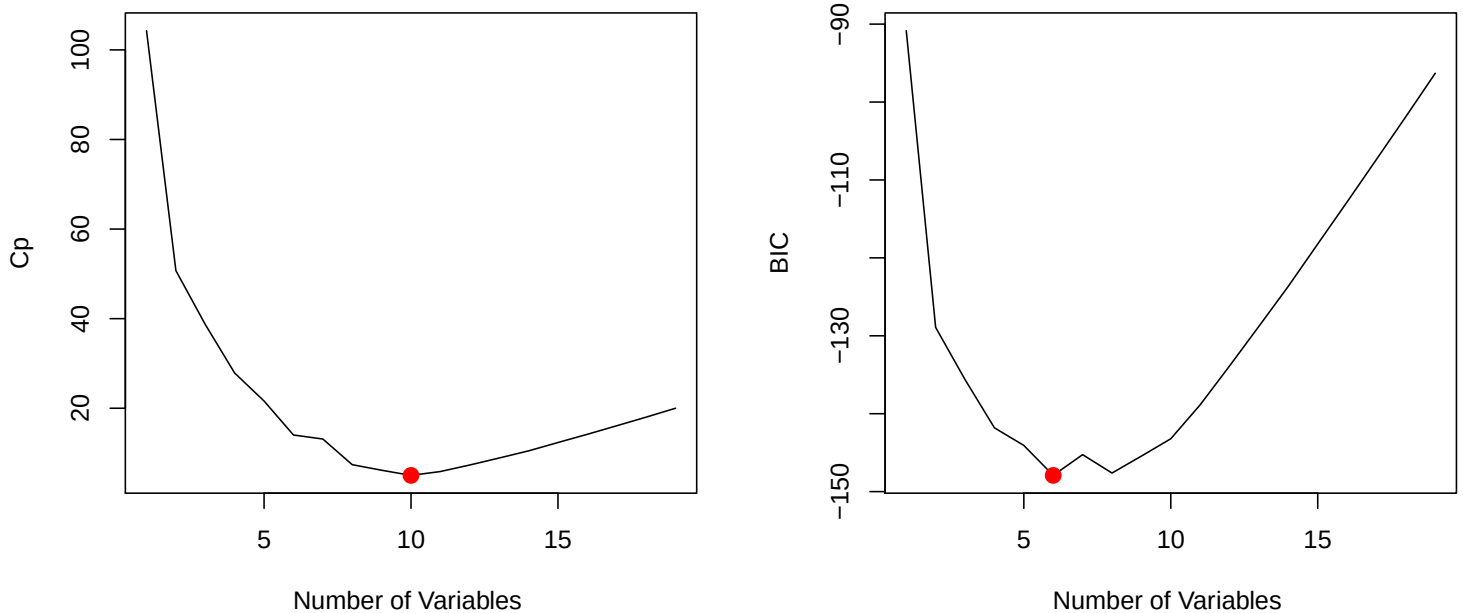
Best subset regression

```
regfit.full=regsubsets(Salary~.,Hitters, nvmax = 19)
reg.summary=summary(regfit.full)
par(mfrow=c(1,2))
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS", type="l")
plot(reg.summary$adjr2 ,xlab="Number of Variables ", ylab="Adjusted RSq",type="l")
k=which.max(reg.summary$adjr2)
points(k,reg.summary$adjr2[k], col="red",cex=2,pch=20)
```



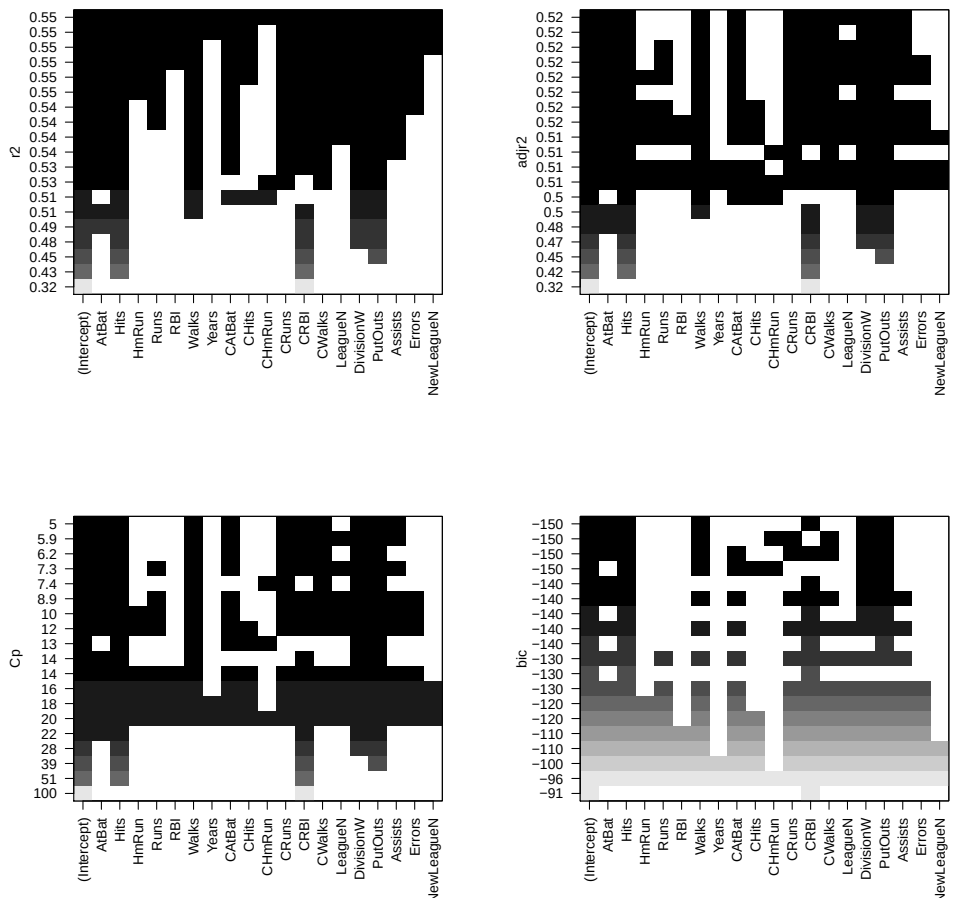
Best subset regression

```
par(mfrow=c(1,2))
plot(reg.summary$cp ,xlab="Number of Variables ",ylab="Cp", type="l")
k=which.min(reg.summary$cp)
points(k,reg.summary$cp[k], col="red",cex=2,pch=20)
plot(reg.summary$bic ,xlab="Number of Variables ", ylab="BIC",type="l")
k=which.min(reg.summary$bic)
points(k,reg.summary$bic[k], col="red",cex=2,pch=20)
```



Best subset regression

```
par(mfrow=c(2,2))
plot(regfit.full,scale="r2")
plot(regfit.full,scale="adjr2")
plot(regfit.full,scale="Cp")
plot(regfit.full,scale="bic")
```



Forward selection

```
regfit.fwd=regsubsets(Salary ~ ., data=Hitters, nvmax=8, method = "forward")
summary(regfit.fwd)$outmat
```

```
##          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
## 1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 4  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 5  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
## 6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*"  " "
## 7  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*"  "*"
## 8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   "*"  "*"
##          LeagueN DivisionW PutOuts Assists Errors NewLeagueN
## 1  ( 1 ) " "     " "       " "     " "     " "    " "
## 2  ( 1 ) " "     " "       " "     " "     " "    " "
## 3  ( 1 ) " "     " "       "*"     " "     " "    " "
## 4  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 5  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 6  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 7  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 8  ( 1 ) " "     "*"       "*"     " "     " "    " "
```

Backward selection

```
regfit.bwd=regsubsets(Salary ~ ., data=Hitters, nvmax=8, method = "backward")
summary(regfit.bwd)$outmat
```

```
##          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
## 1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
## 2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
## 3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
## 4  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
## 5  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "  " "
## 6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "  " "
## 7  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "  "*"
## 8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   "*"  "*"
##          LeagueN DivisionW PutOuts Assists Errors NewLeagueN
## 1  ( 1 ) " "     " "       " "     " "     " "    " "
## 2  ( 1 ) " "     " "       " "     " "     " "    " "
## 3  ( 1 ) " "     " "       "*"     " "     " "    " "
## 4  ( 1 ) " "     " "       "*"     " "     " "    " "
## 5  ( 1 ) " "     " "       "*"     " "     " "    " "
## 6  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 7  ( 1 ) " "     "*"       "*"     " "     " "    " "
## 8  ( 1 ) " "     "*"       "*"     " "     " "    " "
```

Why not the same as the previous case?

What's next? A preview of some fancy stuff to come

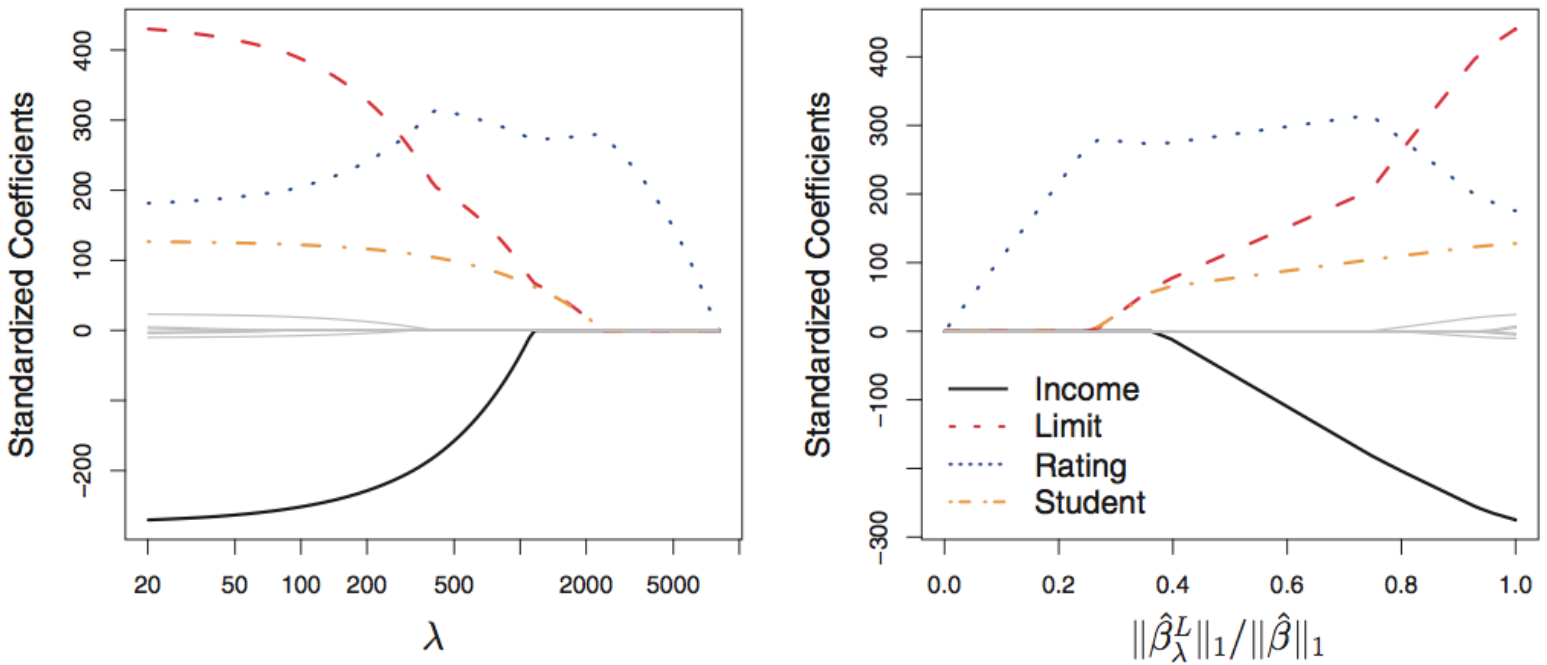


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.