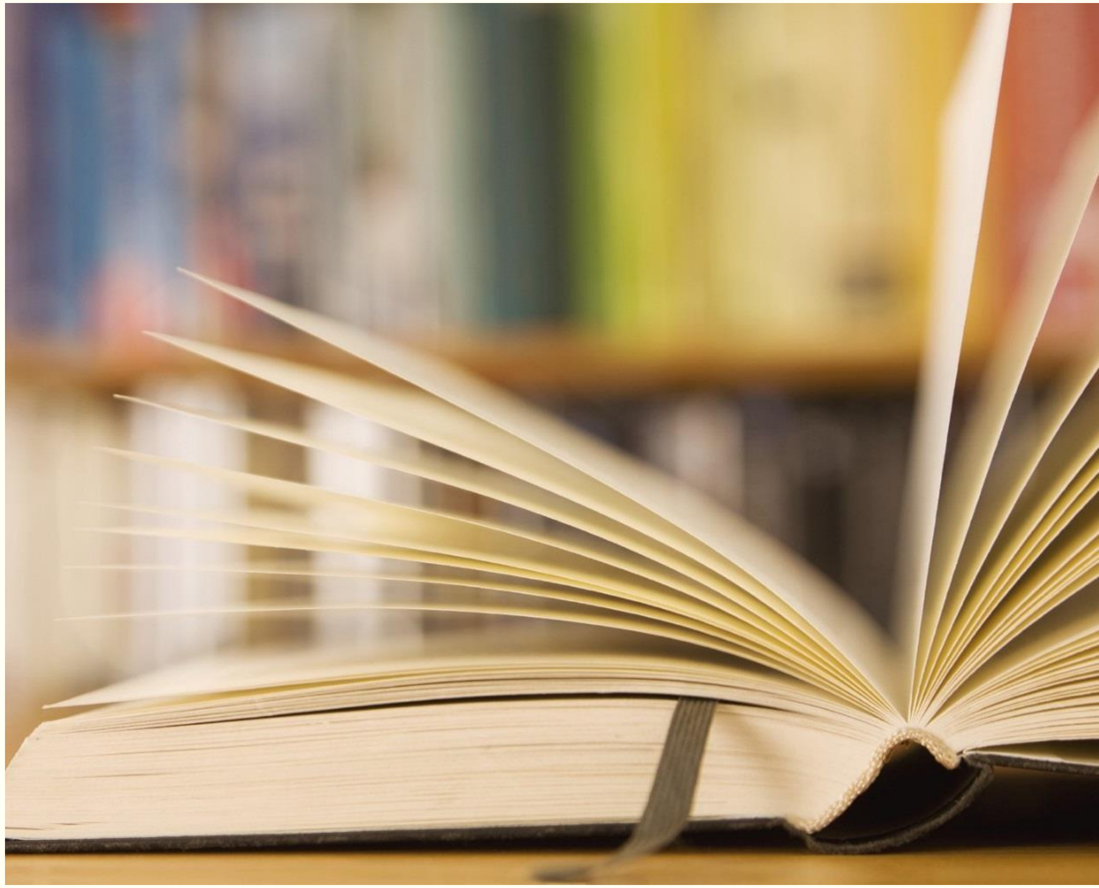
STA 4320 REVIEW



4 assumptions on error terms of least squares regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1) Normal
2) expectation 0
3) same variance ; $var(\varepsilon_i) = \sigma^2$ for all i
4) independent

note: ε_i and X_i are independent (unrelated)

ε_i cannot be reduced to 0

anova vs summary on lm command

- reg = lm(mpg ~ horsepower + acceleration + cylinders + displacement, dat = Auto) #from ISLR2 package
- anova(reg)

```
Analysis of Variance Table

Response: mpg
              Df  Sum Sq Mean Sq  F value    Pr(>F)    
horsepower     1 14433.1 14433.1 726.3343 < 2.2e-16 ***
acceleration   1   581.0   581.0  29.2360 1.124e-07 ***
cylinders      1   943.1   943.1  47.4620 2.282e-11 ***
displacement   1   171.7   171.7   8.6415  0.003483 ** 
Residuals    387  7690.1    19.9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H_0: model contains only intercept

H_1: model contains intercept and horse power

H_0: model consists of intercept, horsepower

H_1: model consists of intercept, horsepower, accelation

ANOVA for model selection

- We first specify a significance level α
- Then we test H_0: model contains only intercept
  H_1: model contains intercept and horsepower
- If pvalue $> \alpha$
- Then the larger model is not significant, and we stop at the smaller model
- If pvalue $< \alpha$, then continue with the next test

Example when $\alpha = 5\%$

H_0: intercept
H_1: intercept, horsepower
pval $< 2.2e-16$ $\Rightarrow$ reject H_0

$\Rightarrow$ H_0: intercept, horsepower
H_1: intercept, horsepower, acceleration
pval $= 1.124e-7$ $\Rightarrow$ reject H_0
$\Rightarrow$ ...
$\Rightarrow$ select all variables

Example when $\alpha = 0.1\%$

H_0: intercept

H_1: intercept, horsepower

pval $< 2.2e{-16}$ $\Rightarrow$ reject H_0

$\Rightarrow$ ...

$\Rightarrow$ H_0: intercept, horsepower, acceleration, cylinders

H_1: intercept, horsepower, acceleration, cylinders, displacement

pval $= 0.003 > \alpha$

fail to reject H_0

$\Rightarrow$ selected model:

intercept, horsepower, acceleration, cylinders

Ridge and LASSO regression CV plots

- The optimal lambda is the one giving the smallest Mean Squared Error

- Log(lambda) is located at the first dotted line

Ridge and LASSO regression coefficient vs lambda plot

Hitters dataset ridge regression plot

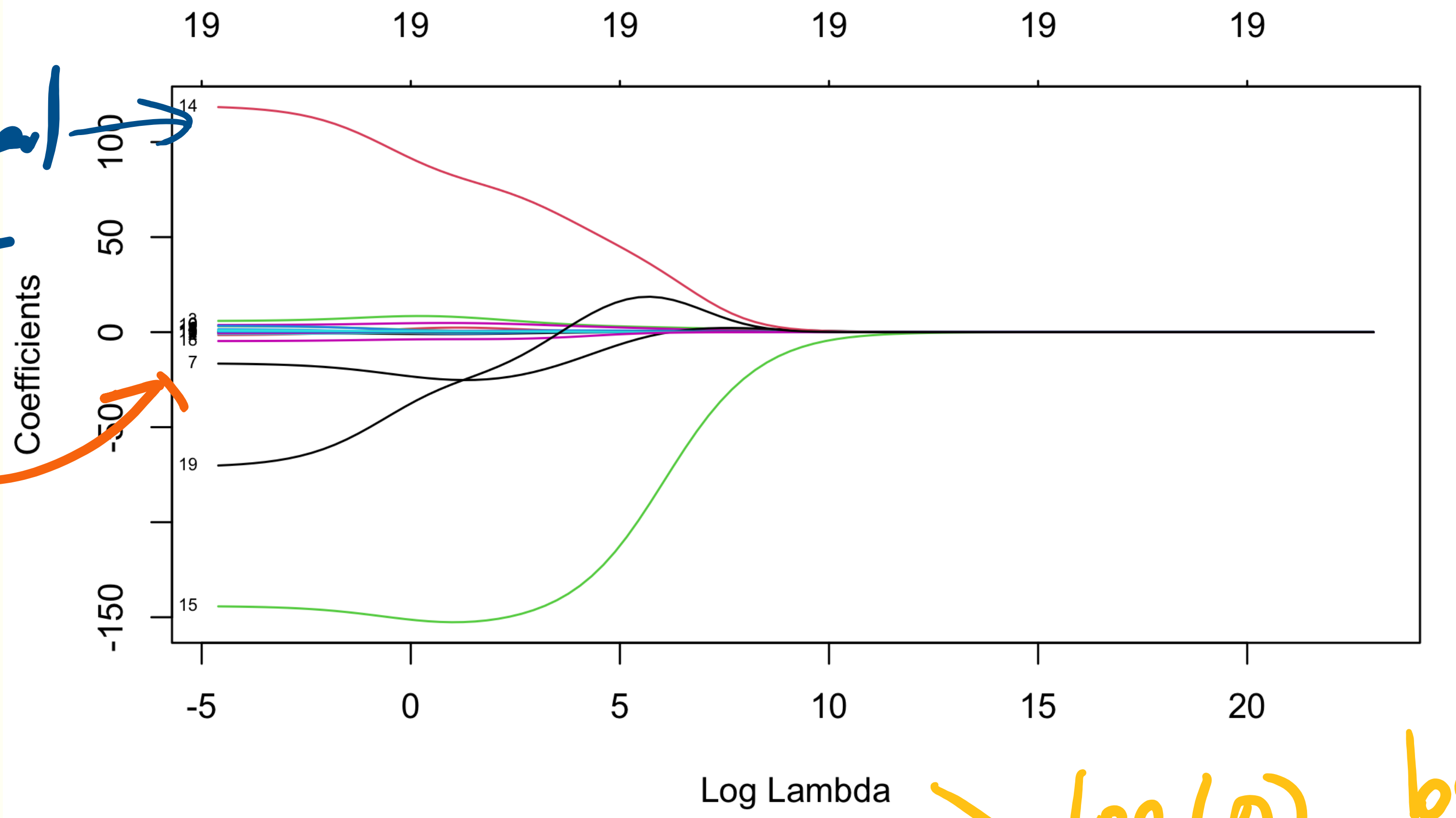
- plot(ridge_mod, "lambda", label = TRUE)

log(λ)=-10, there are 19 slope estimates non-zero

when log(λ)=20 there are 19 slope estimates that are non zero

It is league categorical variable

7 is years



log(λ) because λ can get big

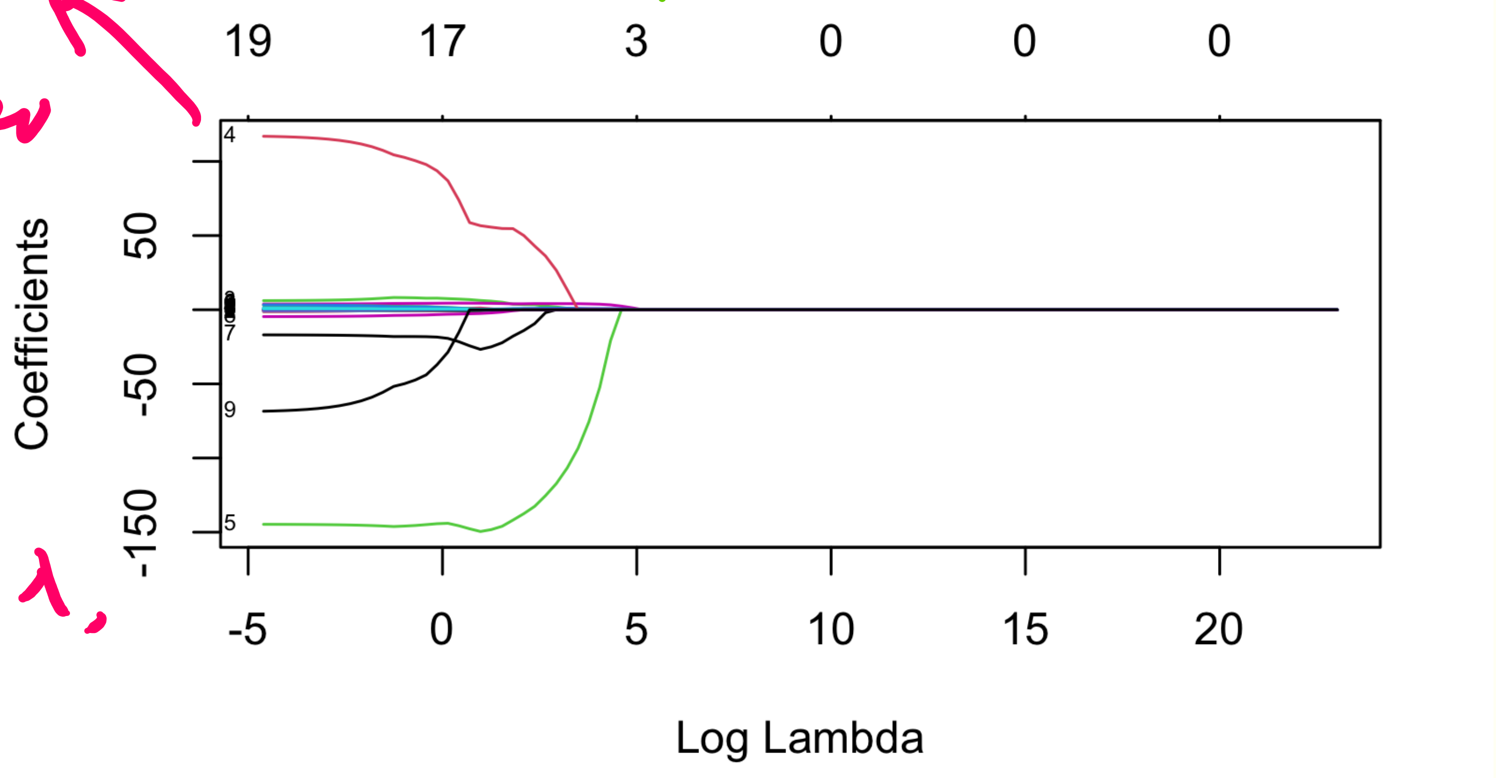
Hitters dataset lasso regression plot

- plot(lasso_mod, "lambda", label = TRUE)

4 means "runs": for smaller λ, it is big in absolute value; for larger λ, it shrinks to 0

when $\log(\lambda)=5$ only 3 slopes are nonzero

when $\log(\lambda)=20$ all slopes are set to 0



Bootstrap

x = c(1, 2, 3)

sample(x, 3, replace = TRUE)

1, 1, 2 possible

1, 1, 1 possible

3, 2, 1 possible

1, 2, 4 impossible

4 is not in the original data

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

$(1,2), (3,4), (5.6)$ possible

$(1,2), (1,2), (1,2)$ possible

$(1,2), (3,4), (7,8)$ impossible

Best one component model

- Consists of intercept and CRBI

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

CRBI has the highest R^2 in all one-component models.
So the best 1-component model is {intercept, CRBI

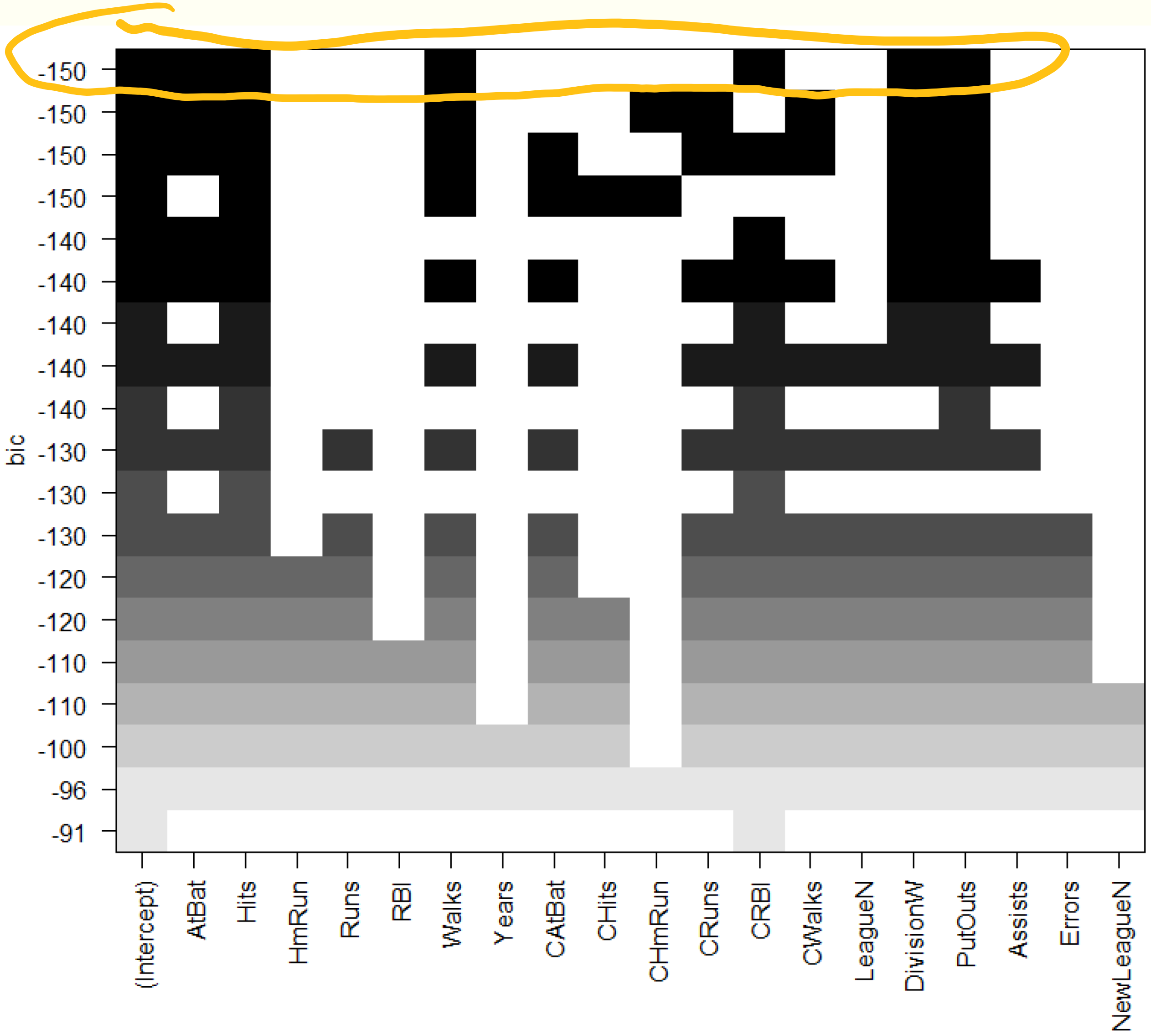
Best 3 component model

- Consists of intercept and CRBI

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

Best 3-component model consists of intercept, Hits, CRBI, PutOuts as they together gives highest R2 for 3-component models.

Determining the overall best model



BIC: smaller is better

the items on the top row form the best model based on BIC:

intercept
AtBat
Hits
Walks
CRBI
DivisionW
PutOuts

- We can use:
- coef(regfit_best, 6)
- To show the best 6 model coefficients (6 is determined from BIC)

- regsubsets(Salary ~ ., data = dat, nvmax = 19, method = "backward")

"backward" : backward stepwise selection
"forward" : forward
empty / missing : best subset selection

Result of coef(regfit_best, 6)

- (Intercept) AtBat Hits Walks CRBI DivisionW PutOuts
- 91.5117981 -1.8685892 7.6043976 3.6976468 0.6430169 -122.9515338 0.2643076

Comparing best subset selection to forward/backward stepwise selection

- Best subset:

It does find the best subset of a given size based on a given criteria

For small p (maximum model size), best subset is preferred:

1) it gives the best model (for given criteria)

2) computational cost is not too high for small p

Comparing best subset selection to forward/backward stepwise selection

- Forward /backward stepwise selection:

It is computationally fast.
For large p (maximum model size),
forward/backward stepwise seletion
is best:
1) because best subset is too slow
large means $p > 25$
2) stepwise seletion usually gives good model

Brief support.

- please provide a brief support
- For example, the proportion is 43.2%, because this is the R2 value.

KNN regression (midterm 2 question 3 c)

Let $\vec{a}, \vec{b} \in \mathbb{R}^p$

L_2 Distance = Euclidean

$$= \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

L1 Distance = Manhattan

$$= \sum_{i=1}^{p} |a_i - b_i|$$

- L2 example already on midterm 2
- If using L1 distance

$$d_{L_1}\left(\begin{bmatrix}2\\1\end{bmatrix},\begin{bmatrix}0\\0\end{bmatrix}\right) = |2-0|+|1-0|=3$$

$$d_{L_1}\left(\begin{bmatrix}2\\1\end{bmatrix},\begin{bmatrix}5\\0\end{bmatrix}\right) = |2-5|+|1-0|=4$$

$$d_{L_1}\left(\begin{bmatrix}2\\1\end{bmatrix},\begin{bmatrix}5\\5\end{bmatrix}\right) = |2-5|+|1-5|=7$$

$$d_{L_1}\left(\begin{bmatrix}2\\1\end{bmatrix},\begin{bmatrix}0\\5\end{bmatrix}\right) = |2-0|+|1-5|=6$$

Here, $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 5 \\ 0 \end{bmatrix}$ are the 2-closest neighbors of $\vec{x_0} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$$\hat{f}(\vec{x_0}) = \frac{6+7}{2} = 6.5$$

$f\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right)$ $f\left(\begin{bmatrix} 5 \\ 0 \end{bmatrix}\right)$