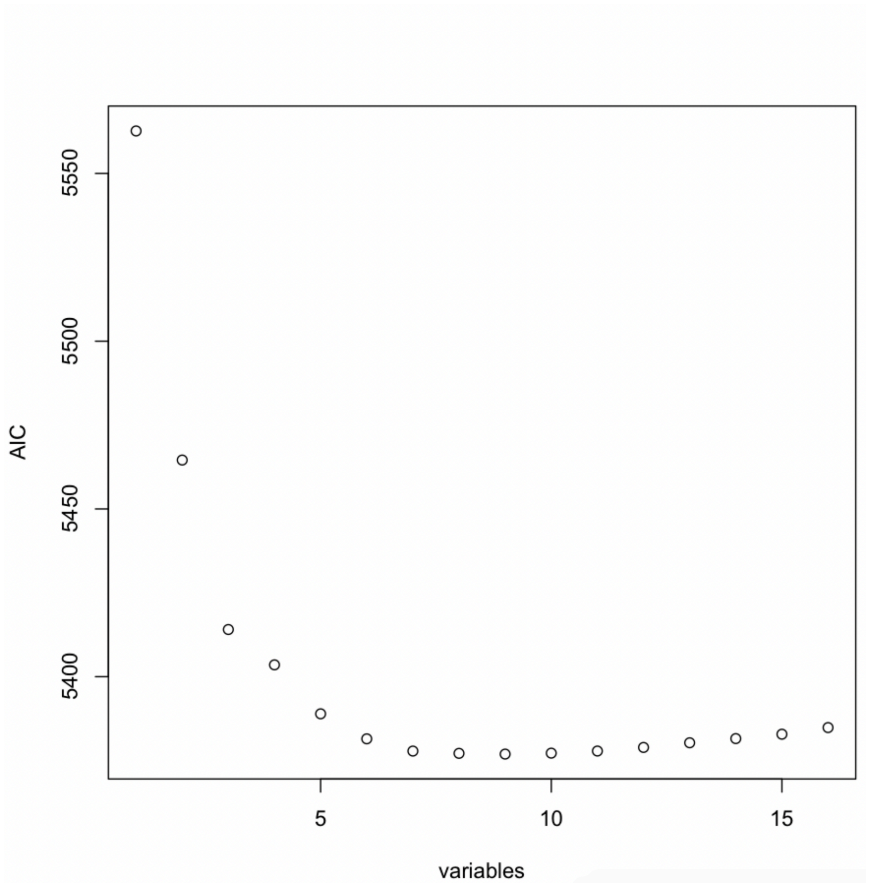
HW 6 using leaps for baseball salary data

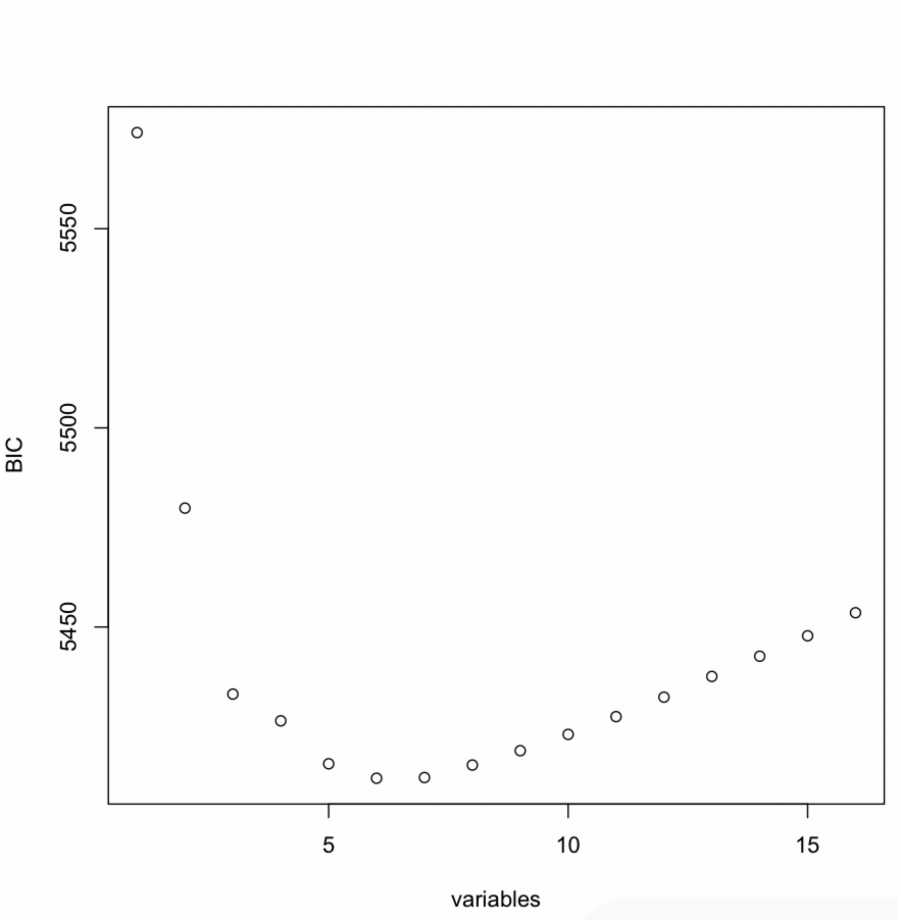
1.) so because we are trying to determine what models is best and we a considering all possible models , I used the leaps .AIC function given in ecampus to find what the best number a variables is , then I used the stand alone leaps command to determine what columns I needed to omit from the model the output I shown below:

```
> outputAIC = leaps.AIC(X,y)
[1] "AIC values"
 [1] 5562.674 5464.568 5414.059 5403.523 5388.926 5381.472 5377.825 5377.144
 [9] 5376.926 5377.207 5377.837 5378.910 5380.296 5381.541 5382.850 5384.824
[1] "BIC values"
 [1] 5574.134 5479.849 5433.159 5426.444 5415.666 5412.032 5412.206 5415.345
 [9] 5418.947 5423.048 5427.499 5432.391 5437.598 5442.663 5447.792 5453.585
>
```

From here we can see that the best model is one that has either 6,7,8, or 9 variables depending on whether you look at AIC or BIC . I would say the best model is one that uses 7 variables since from screenshot below was can see that the R^2 for both models is not that different . Also below is the plot of AIC and BIC relative to how many variables are being used in the model and again the findings match up. Because AIC tends to over estimate the number of variables but the R^2 for all 4 possible models kinda match up I think 7 variables is a safe bet.

```
      1     2     3     4     5     6     7    8     9     A     B     C     D
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
2  FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE  TRUE
3  FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE  TRUE
4  FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE  TRUE FALSE FALSE  TRUE
5  FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE  TRUE  TRUE FALSE  TRUE
6  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE TRUE FALSE  TRUE  TRUE FALSE  TRUE
7  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE TRUE FALSE  TRUE  TRUE FALSE  TRUE
8  FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE TRUE FALSE  TRUE  TRUE FALSE  TRUE
9  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE TRUE  TRUE  TRUE  TRUE FALSE  TRUE
10 FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE TRUE FALSE  TRUE  TRUE  TRUE  TRUE
11 FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE TRUE FALSE  TRUE  TRUE  TRUE  TRUE
12  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
13 FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
16  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
      E     F     G
1  FALSE FALSE FALSE
2  FALSE FALSE FALSE
3  FALSE  TRUE FALSE
4  FALSE  TRUE FALSE
5  FALSE  TRUE FALSE
6  FALSE  TRUE FALSE
7   TRUE  TRUE FALSE
8   TRUE  TRUE FALSE
9   TRUE  TRUE  TRUE
10  TRUE  TRUE  TRUE
11  TRUE  TRUE  TRUE
12  TRUE  TRUE  TRUE
13  TRUE  TRUE  TRUE
14  TRUE  TRUE  TRUE
15  TRUE  TRUE  TRUE
16  TRUE  TRUE  TRUE

$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"
 [6] "5"           "6"           "7"           "8"           "9"
[11] "A"           "B"           "C"           "D"           "E"
[16] "F"           "G"

$size
 [1]  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17

$r2
 [1] 0.4467879 0.5889593 0.6482650 0.6611081 0.6773946 0.6863191 0.6915310
 [8] 0.6939755 0.6959834 0.6975303 0.6987566 0.6995847 0.7001309 0.7008022
[15] 0.7014150 0.7014386
```


AIC
5400
5450
5500
5550
5
10
15
variables


BIC
5450
5500
5550
5
10
15
variables

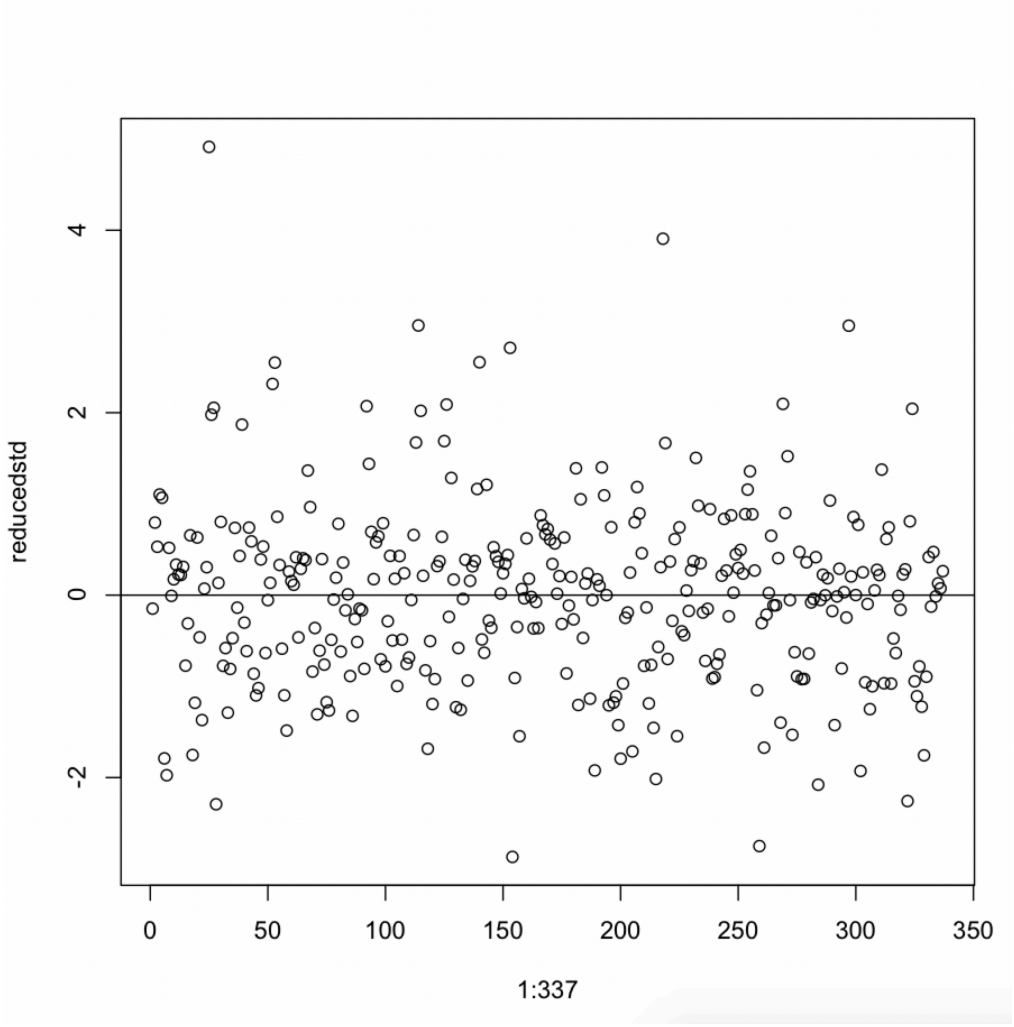
Now this model is one that excludes data from columns [2,7]-[10]-[13]-[17] so now get its summary stats below:

```
> summary(fitreduced)

Call:
lm(formula = baseball$salary ~ ., data = baseball[, c(8:9, 11:12,
    14:16)])

Residuals:
    Min      1Q  Median      3Q     Max
-1928.2  -450.0    16.3   328.0  3335.6

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -109.277     87.869  -1.244   0.2145
home.runs            25.447      9.506   2.677   0.0078 **
rbi                  19.335      2.967   6.516 2.71e-10 ***
strike.outs          -9.537      1.841  -5.180 3.88e-07 ***
stolen.bases         16.387      3.579   4.578 6.65e-06 ***
free.agent.eligible 1408.415    102.421  13.751  < 2e-16 ***
free.agent          -320.084    135.761  -2.358   0.0190 *
arbitration.eligible 818.333    110.855   7.382 1.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 696 on 329 degrees of freedom
Multiple R-squared:  0.6915,    Adjusted R-squared:  0.685
F-statistic: 105.4 on 7 and 329 DF,  p-value: < 2.2e-16

> 
```

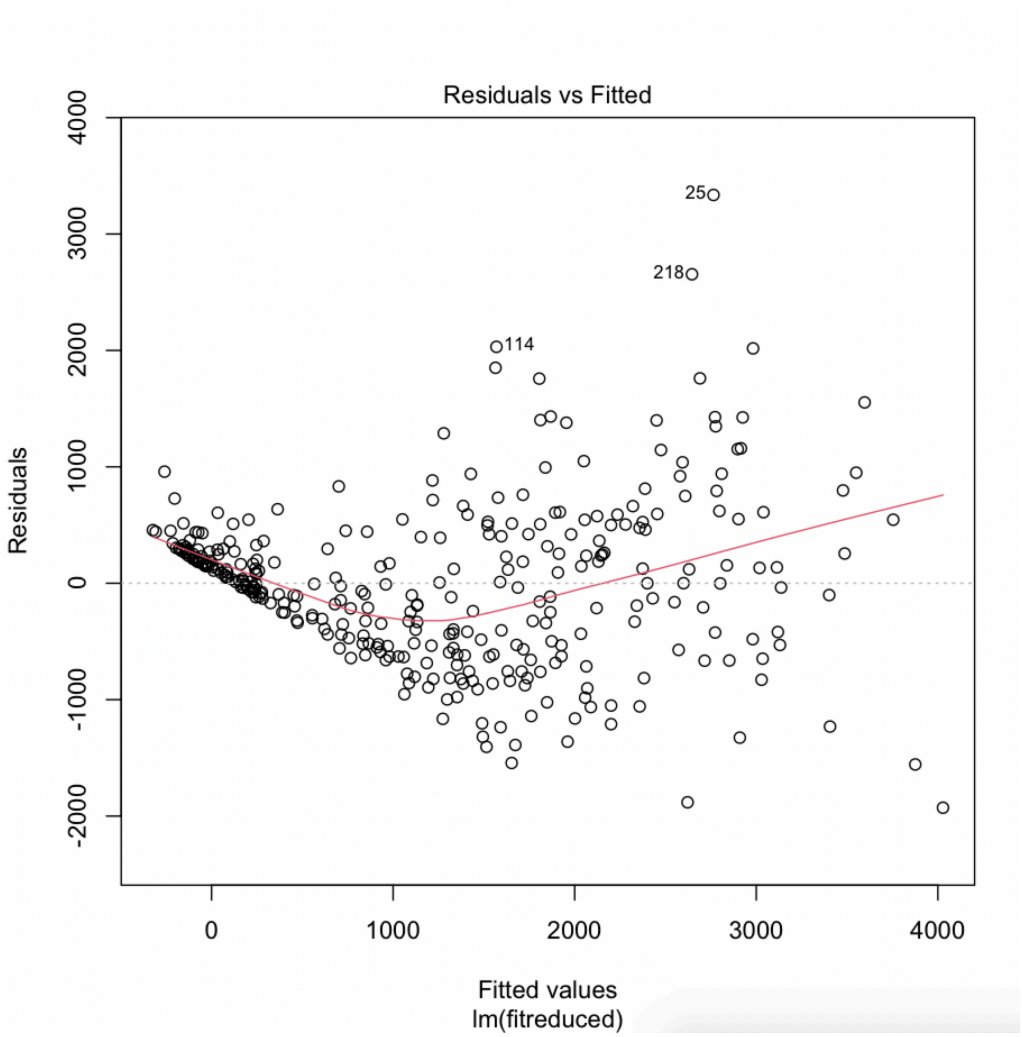
As you can see all the p-values are really small meaning that we need all 7 variables for the model.

2.) there seems to be only 2 player who’s residuals have a magnitude of greater than three , these players correspond top player 25 and 218. If we look at the data itself we see that players Bobby Bonilla and Danny Tartabull both have relatively high salaries compared to the other players .

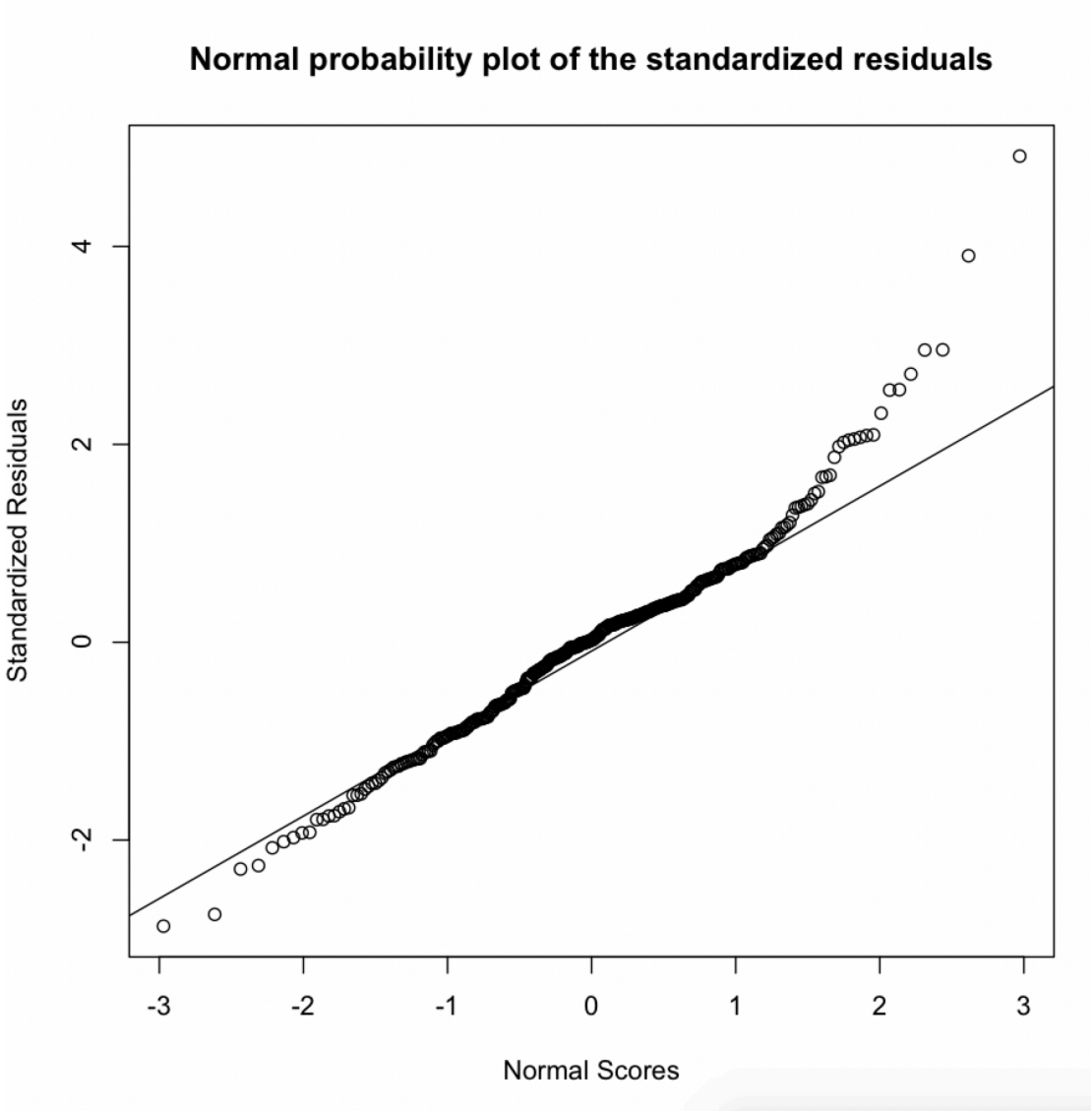
This would make them outliers because their respective responses are unusually high. This makes sense when looking at the plot because you would expect those outliers to be farther away from zero than most.



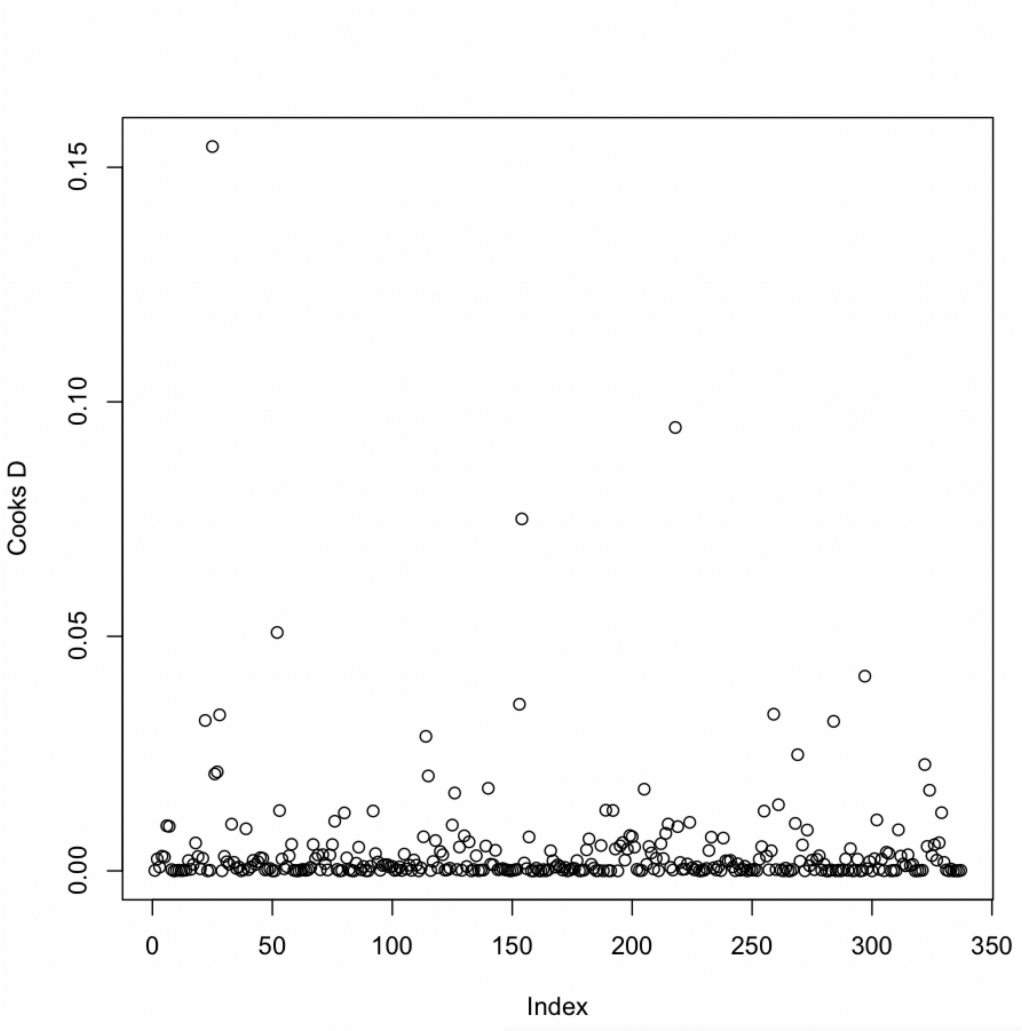
3.) Here the plot is almost the same as the same plot from last weeks homework where I said that it might be that salary is not linearly related to the predictors as suggested by the slight negative trend when looking the the smaller predicted values . It is also true as the response gets higher the tends to also get higher so it makes sense why we are seeing some non-constant variance.



3.) Here when looking at this plot what we a re looking for is that the plot should be fairly linear We are trying to find out if the the error terms do follow a normal distribution which from the plot I would conclude that yes they do. This is almost the same plot as last weeks plot with slight variations.



5.) After plotting the cooks distance we can see that there doesn't seems to be any leverage points that drastically affect the estimates of the betas. There is some outliers that effect the estimate of $\hat{\sigma}$ and those the same ones I pointed out in the previous questions.


Cooks D
0.00
0.05
0.10
0.15
0
50
100
150
200
250
300
350
Index