Assignment 5

Dataset

The dataset you are going to use for this assignment is called Hitters_Fixed, which is a cleaned version of the Hitters dataset. First, you are going to read the Hitters dataset; then, you will clean it to get the Hitters_Fixed dataset. Take the following steps to do so:

- Install the “ISLR” package in R and then load it into R. The “ISLR” contains the Hitters dataset.
- Remove the missing data from the Hitters dataset and save the results into Hitters_Fixed. Run the following line of code to do so:

Hitters_Fixed = na.omit(Hitters)

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.2.2
```

```
Hitters_Fixed=na.omit(Hitters)
```

That is all. Now you can use the Hitters_Fixed dataset to do this assignment!

Note: If you want more info about the variables in the Hitters_Fixed dataset, you can get it using this link:

https://docs.google.com/document/d/1qKeEoWVnAkrlPDwuq7Uz65ybOzrXpBC2pplRbabcgJ0/edit?usp=sharing (https://docs.google.com/document/d/1qKeEoWVnAkrlPDwuq7Uz65ybOzrXpBC2pplRbabcgJ0/edit?usp=sharing)

Questions

We are going to use all the available data in the Hitters_Fixed dataset to find a model to predict **“Salary”** (i.e., Salary is the outcome variable)

```
str(Hitters_Fixed)
```

```
## 'data.frame':    263 obs. of  20 variables:
##  $ AtBat    : int  315 479 496 321 594 185 298 323 401 574 ...
##  $ Hits     : int  81 130 141 87 169 37 73 81 92 159 ...
##  $ HmRun    : int  7 18 20 10 4 1 0 6 17 21 ...
##  $ Runs     : int  24 66 65 39 74 23 24 26 49 107 ...
##  $ RBI      : int  38 72 78 42 51 8 24 32 66 75 ...
##  $ Walks    : int  39 76 37 30 35 21 7 8 65 59 ...
##  $ Years    : int  14 3 11 2 11 2 3 2 13 10 ...
##  $ CAtBat   : int  3449 1624 5628 396 4408 214 509 341 5206 4631 ...
##  $ CHits    : int  835 457 1575 101 1133 42 108 86 1332 1300 ...
##  $ CHmRun   : int  69 63 225 12 19 1 0 6 253 90 ...
##  $ CRuns    : int  321 224 828 48 501 30 41 32 784 702 ...
##  $ CRBI     : int  414 266 838 46 336 9 37 34 890 504 ...
##  $ CWalks   : int  375 263 354 33 194 24 12 8 866 488 ...
##  $ League   : Factor w/ 2 levels "A","N": 2 1 2 2 1 2 1 2 1 1 ...
##  $ Division : Factor w/ 2 levels "E","W": 2 2 1 1 2 1 2 2 1 1 ...
##  $ PutOuts  : int  632 880 200 805 282 76 121 143 0 238 ...
##  $ Assists  : int  43 82 11 40 421 127 283 290 0 445 ...
##  $ Errors   : int  10 14 3 4 25 7 9 19 0 22 ...
##  $ Salary   : num  475 480 500 91.5 750 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 2 1 2 2 1 1 1 2 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:59] 1 16 19 23 31 33 37 39 40 42 ...
##   ..- attr(*, "names")= chr [1:59] "-Andy Allanson" "-Billy Beane" "-Bruce Bochte" "-Bob B
oone" ...
```

```
#Salary is in thousands
```

```
summary(Hitters_Fixed)
```

```
##      AtBat            Hits           HmRun            Runs
##  Min.   : 19.0   Min.   :  1.0   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:282.5   1st Qu.: 71.5   1st Qu.: 5.00   1st Qu.: 33.50
##  Median :413.0   Median :103.0   Median : 9.00   Median : 52.00
##  Mean   :403.6   Mean   :107.8   Mean   :11.62   Mean   : 54.75
##  3rd Qu.:526.0   3rd Qu.:141.5   3rd Qu.:18.00   3rd Qu.: 73.00
##  Max.   :687.0   Max.   :238.0   Max.   :40.00   Max.   :130.00
##       RBI             Walks            Years            CAtBat
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0
##  1st Qu.: 30.00   1st Qu.: 23.00   1st Qu.: 4.000   1st Qu.:  842.5
##  Median : 47.00   Median : 37.00   Median : 6.000   Median : 1931.0
##  Mean   : 51.49   Mean   : 41.11   Mean   : 7.312   Mean   : 2657.5
##  3rd Qu.: 71.00   3rd Qu.: 57.00   3rd Qu.:10.000   3rd Qu.: 3890.5
##  Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0
##      CHits            CHmRun            CRuns             CRBI
##  Min.   :   4.0   Min.   :  0.00   Min.   :   2.0   Min.   :   3.0
##  1st Qu.: 212.0   1st Qu.: 15.00   1st Qu.: 105.5   1st Qu.:  95.0
##  Median : 516.0   Median : 40.00   Median : 250.0   Median : 230.0
##  Mean   : 722.2   Mean   : 69.24   Mean   : 361.2   Mean   : 330.4
##  3rd Qu.:1054.0   3rd Qu.: 92.50   3rd Qu.: 497.5   3rd Qu.: 424.5
##  Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.0
##      CWalks        League  Division    PutOuts          Assists
##  Min.   :   1.0   A:139    E:129    Min.   :   0.0   Min.   :  0.0
##  1st Qu.:  71.0   N:124    W:134    1st Qu.: 113.5   1st Qu.:  8.0
##  Median : 174.0                     Median : 224.0   Median : 45.0
##  Mean   : 260.3                     Mean   : 290.7   Mean   :118.8
##  3rd Qu.: 328.5                     3rd Qu.: 322.5   3rd Qu.:192.0
##  Max.   :1566.0                     Max.   :1377.0   Max.   :492.0
##      Errors          Salary       NewLeague
##  Min.   : 0.000   Min.   :  67.5   A:141
##  1st Qu.: 3.000   1st Qu.: 190.0   N:122
##  Median : 7.000   Median : 425.0
##  Mean   : 8.593   Mean   : 535.9
##  3rd Qu.:13.000   3rd Qu.: 750.0
##  Max.   :32.000   Max.   :2460.0
```

Question 1) Conduct a regression analysis using “Hits” as the only predictor. Answer the following questions:

```
cor(Hitters_Fixed$Salary, Hitters_Fixed$Hits)
```

```
## [1] 0.4386747
```

Correlation seems low. Doesn’t give me hopes in the practical part of this regression analysis

```
salary_simple_out = lm(Salary ~ Hits, data = Hitters_Fixed)

summary(salary_simple_out)
```

```
## 
## Call:
## lm(formula = Salary ~ Hits, data = Hitters_Fixed)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -893.99 -245.63  -59.08  181.12 2059.90 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  63.0488    64.9822   0.970    0.333    
## Hits          4.3854     0.5561   7.886 8.53e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 406.2 on 261 degrees of freedom
## Multiple R-squared:  0.1924, Adjusted R-squared:  0.1893 
## F-statistic: 62.19 on 1 and 261 DF,  p-value: 8.531e-14
```

a. Based on your results, discuss BOTH the statistical significance and the quality of the linear equation that uses “Hits” to predict the player’s salary.

First Impressions: PV of 8.53e-14 indicates statistical significance. R-squared being .1924 is very ugly. An RSE of 406.2 (measured in thousands of dollars) sounds rather significant. The practical significance of this linear regression already feels doomed, but lets be sure

```
RSE = 406.2

RSE/mean(Hitters_Fixed$Salary)
```

```
## [1] 0.7579406
```

That is an absolutely horrid number to receive from a coefficient of variation. This confirms my suspicions. While the equation derived from the simple linear regression of hits to predict salary is statistically significant, it is most definitely not practically significant.

Overall, hits is a poor predictor of salary, despite its statistical significance, due to the coefficient of variability being extremely high when being calculated using a mean of salary.

b. Write out the equation that you would use to predict salary based on the player’s number of hits in the 1986 season.

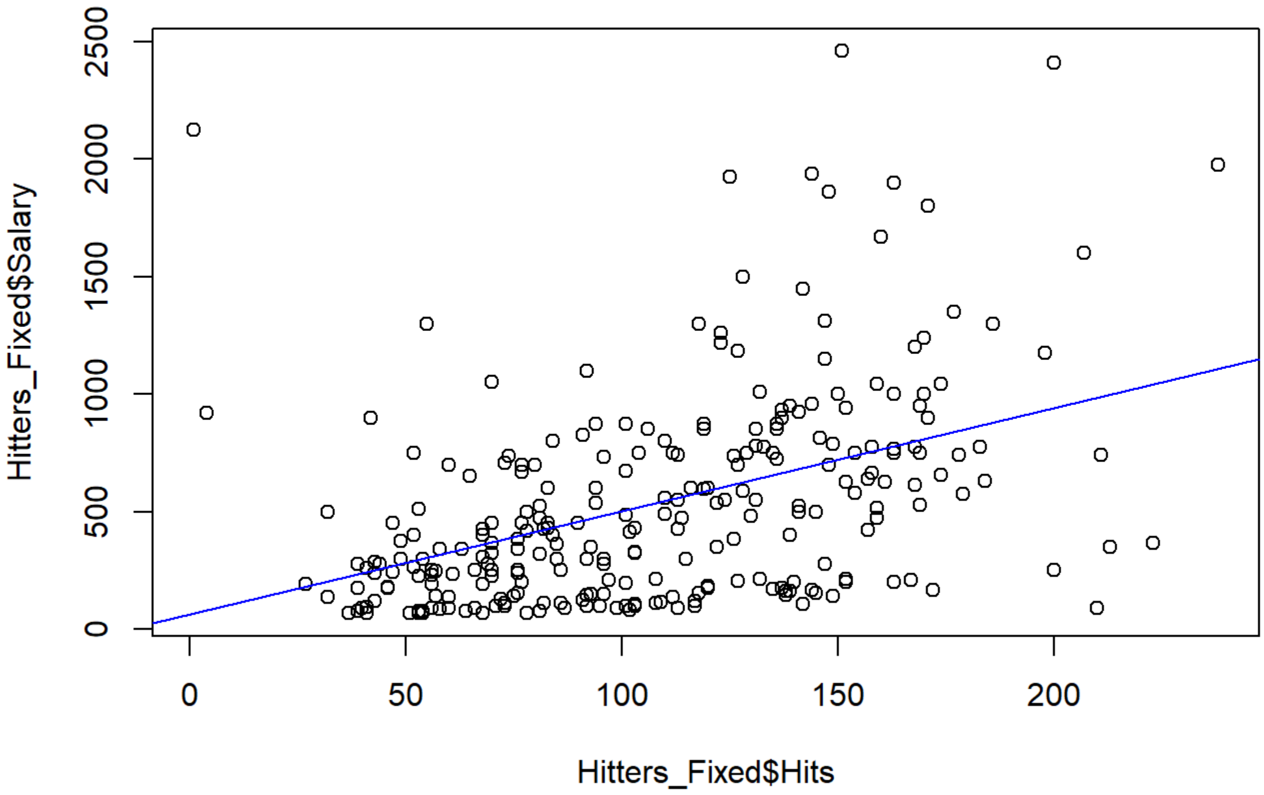
Predicted Salary = 63.05 - 4.39(Number of Hits)

Question 2) Use the same regression analysis from question 1 to answer this question.

Assume that Assumption 2 regarding the validity of the linear regression analysis is satisfied. Run an analysis to check for the validity of all other assumptions. Comment on your findings.

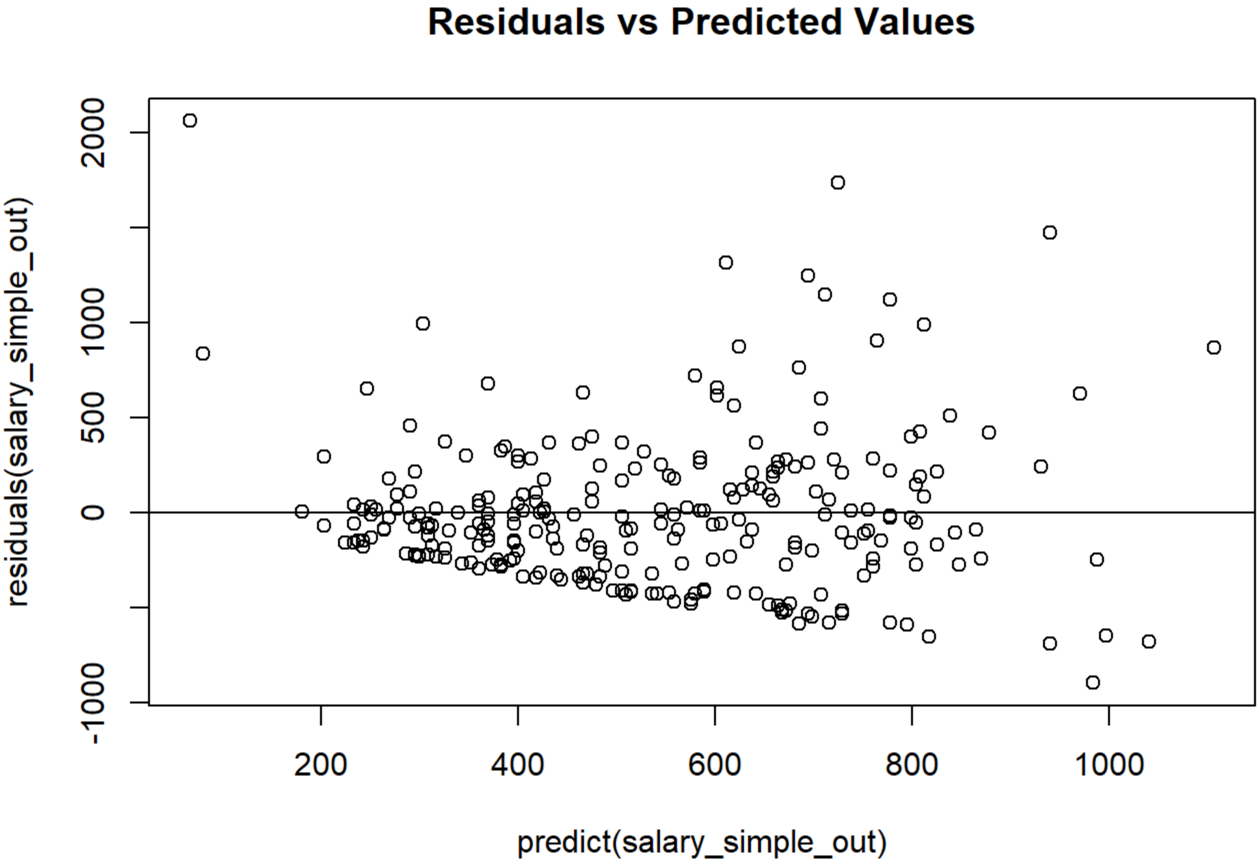
```
plot(Hitters_Fixed$Hits, Hitters_Fixed$Salary)

abline(salary_simple_out, col = "blue")
```



I would argue that there is a lack in linearity between hits and salary. There are too many scattered points that don't follow a linear relationship, thus voiding assumption one. This alone would probably be enough to conclude that further testing for the validity of our linear regression would be uncessary, but as an exercise in validation let us continue to test our other assumptions.

```
plot(predict(salary_simple_out), residuals(salary_simple_out))
abline(h=0)
title("Residuals vs Predicted Values")
```



There seems to be a lack of even distribution in the relationship of x and y in the above plot (Residuals vs Predicted Values), voiding assumption four. There is a decline in linearity as salary increases, and even that does not account for the scattered points we receive above the x axis. I'll proceed to test for assumption three using Shapiro, but again, all previous information of this regression analysis would normally cause me to proceed without conducting further validity testing.

Ho: The residuals follow a normal distribution Ha: The residuals do not follow a normal distribution

```
shapiro.test(residuals(salary_simple_out))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(salary_simple_out)
## W = 0.90073, p-value = 3.795e-12
```

Here we can see that the Shapiro test rejects the null hypothesis due to a low PV. Indicating that assumption three (normality) is not valid for this linear regression analysis.

Overall, we can see that despite proving statistical significance the simple linear regression to predict salary using hits is a poor example of linear regression when checking for validity and when examining in the most simple terms practical signifiance.

Question 3) Apply the best subset selection method to find a good multiple linear equation. DO NOT INCLUDE in the analysis the following four predictors:

CHits, CAtBat, CRuns, CRBI

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.2.1
```

```
best_subset_salary = regsubsets(Salary~.-CHits -CAtBat -CRuns -CRBI, data = Hitters_Fixed, nv
max = length(Hitters_Fixed) - 5)

summary(best_subset_salary)
```

```
## Subset selection object
## Call: regsubsets.formula(Salary ~ . - CHits - CAtBat - CRuns - CRBI, 
##     data = Hitters_Fixed, nvmax = length(Hitters_Fixed) - 5)
## 15 Variables  (and intercept)
##            Forced in Forced out
## AtBat          FALSE      FALSE
## Hits           FALSE      FALSE
## HmRun          FALSE      FALSE
## Runs           FALSE      FALSE
## RBI            FALSE      FALSE
## Walks          FALSE      FALSE
## Years          FALSE      FALSE
## CHmRun         FALSE      FALSE
## CWalks         FALSE      FALSE
## LeagueN        FALSE      FALSE
## DivisionW      FALSE      FALSE
## PutOuts        FALSE      FALSE
## Assists        FALSE      FALSE
## Errors         FALSE      FALSE
## NewLeagueN     FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: exhaustive
##           AtBat Hits HmRun Runs RBI Walks Years CHmRun CWalks LeagueN DivisionW
## 1  ( 1 )  " "   " "  " "   " "  " " " "   " "   "*"    " "    " "     " "      
## 2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   "*"    " "    " "     " "      
## 3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   "*"    " "    " "     " "      
## 4  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   "*"    " "    " "     " "      
## 5  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   "*"    " "    " "     "*"      
## 6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "    " "     "*"      
## 7  ( 1 )  "*"   "*"  " "   " "  " " "*"   "*"   "*"    " "    " "     "*"      
## 8  ( 1 )  "*"   "*"  " "   " "  " " "*"   "*"   "*"    " "    " "     "*"      
## 9  ( 1 )  "*"   "*"  " "   " "  " " "*"   "*"   "*"    " "    "*"     "*"      
## 10  ( 1 ) "*"   "*"  "*"   " "  " " "*"   "*"   "*"    " "    "*"     "*"      
## 11  ( 1 ) "*"   "*"  "*"   " "  " " "*"   "*"   "*"    " "    "*"     "*"      
## 12  ( 1 ) "*"   "*"  "*"   " "  "*" "*"   "*"   "*"    " "    "*"     "*"      
## 13  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    " "    "*"     "*"      
## 14  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"    "*"     "*"      
## 15  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"    "*"     "*"      
##           PutOuts Assists Errors NewLeagueN
## 1  ( 1 )  " "     " "     " "    " "       
## 2  ( 1 )  " "     " "     " "    " "       
## 3  ( 1 )  "*"     " "     " "    " "       
## 4  ( 1 )  "*"     " "     " "    " "       
## 5  ( 1 )  "*"     " "     " "    " "       
## 6  ( 1 )  "*"     " "     " "    " "       
## 7  ( 1 )  "*"     " "     " "    " "       
## 8  ( 1 )  "*"     "*"     " "    " "       
## 9  ( 1 )  "*"     "*"     " "    " "       
## 10  ( 1 ) "*"     "*"     " "    " "       
## 11  ( 1 ) "*"     "*"     "*"    " "       
## 12  ( 1 ) "*"     "*"     "*"    " "       
## 13  ( 1 ) "*"     "*"     "*"    " "
## 14  ( 1 ) "*"     "*"     "*"    " "
## 15  ( 1 ) "*"     "*"     "*"    "*"
```

Here we have our possible four equations to predict salary using all predictors except: CHits, CAtBat, CRuns, and CRBI. Lets find out which of these equations is best using adjusted R-squared.

```
summary(best_subset_salary)$adjr2
```

```
##  [1] 0.2727764 0.3902841 0.4136565 0.4410945 0.4578190 0.4699300 0.4759338
##  [8] 0.4781049 0.4790087 0.4780628 0.4766292 0.4745954 0.4725383 0.4704214
## [15] 0.4682810
```

```
which.max(summary(best_subset_salary)$adjr2)
```

```
## [1] 9
```

Here we can see that our 9th equation is our best predictor using multiple linear regression with the aforementioned columns. The R-squared for this equation is .479 and it contains AtBat, Hits, Walks, Years, CHmRun, LeagueN, DivisionW, PutOuts, and Assists.

a. Write the equation that includes the predictors that you consider appropriate. JUSTIFY why you selected this equation.

```
coef (best_subset_salary, 9)
```

```
##  (Intercept)        AtBat         Hits        Walks        Years       CHmRun 
##   40.2798451   -2.2448046    8.9948138    3.4639048   10.5717841    2.0277819 
##      LeagueN    DivisionW      PutOuts      Assists 
##   49.6591458 -115.9355338    0.2836580    0.2098445
```

Our equation is as follows:

Predicted Salary = 40.28 - 2.24(AtBat) + 8.99(Hits) + 3.46(Walks) + 10.57(Years) + 2.03(CHRun) + 49.66(LeagueN) - 115.94(DivisionW) + 0.28(PutOuts) + 0.21(Assists)

This equation was selected because of all the possible equations of our multiple linear regression, a combination of nine predictors composed of the previously aforementioned columns resulted in the largest adjusted R-squared value.

b. What percentage of the total variability in the salary values does the equation that you chose in 3 a) eliminate?

```
summary(lm(Salary~.-CHits -CAtBat -CRuns -CRBI -Errors -Runs -RBI -CWalks -NewLeague, data =
Hitters_Fixed))
```

```
##
## Call:
## lm(formula = Salary ~ . - CHits - CAtBat - CRuns - CRBI - Errors -
##     Runs - RBI - CWalks - NewLeague, data = Hitters_Fixed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1090.85  -167.12   -38.22   139.01  1988.36
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.25665   83.04272   0.545 0.586249
## AtBat         -2.16065    0.57674  -3.746 0.000222 ***
## Hits           8.98514    1.72157   5.219 3.76e-07 ***
## HmRun         -2.62633    3.56915  -0.736 0.462511
## Walks          3.43572    1.25711   2.733 0.006720 **
## Years          9.04203    6.70588   1.348 0.178749
## CHmRun         2.18492    0.45805   4.770 3.12e-06 ***
## LeagueN       45.82658   41.73727   1.098 0.273262
## DivisionW   -116.08259   40.58315  -2.860 0.004586 **
## PutOuts        0.28616    0.07772   3.682 0.000283 ***
## Assists        0.16975    0.17035   0.996 0.320000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 325.9 on 252 degrees of freedom
## Multiple R-squared:  0.498,  Adjusted R-squared:  0.4781
## F-statistic:    25 on 10 and 252 DF,  p-value: < 2.2e-16
```

An R-squared of .498 indicates that 49.8% of variability is eliminated by our equation.

c. Is the prediction error of this equation low or high? Justify your answer (Note: Just give me a well thought, very short, and simple answer for why you consider the error low or high)

```
RSE = 325.9

RSE/mean(Hitters_Fixed$Salary)
```

```
## [1] 0.6081065
```

This is still a considerably high value for a coefficient of variation. Ideally, a coefficient of variation should be less than or equal to ten percent.