STAT 641 APPLED DATA ANALAYSIS

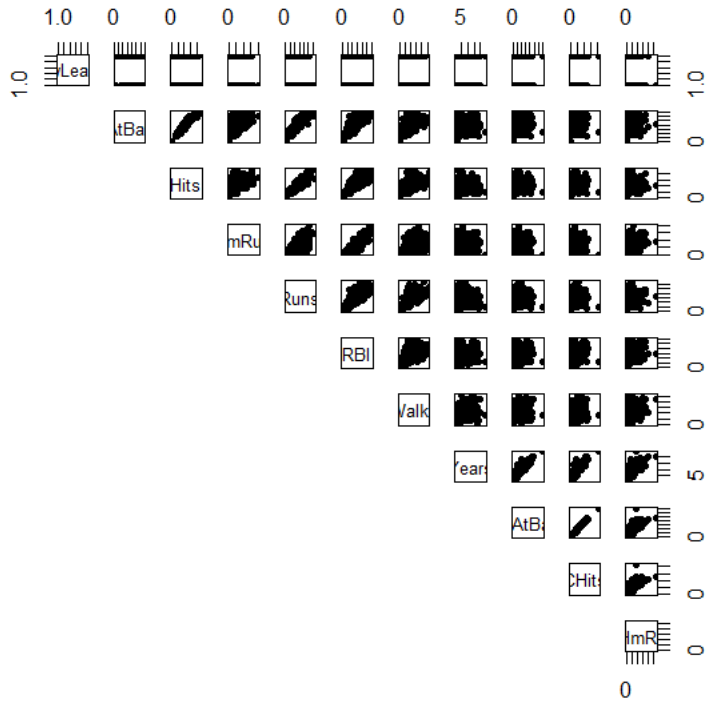
Final Project- Surya Suresh(V00998151)

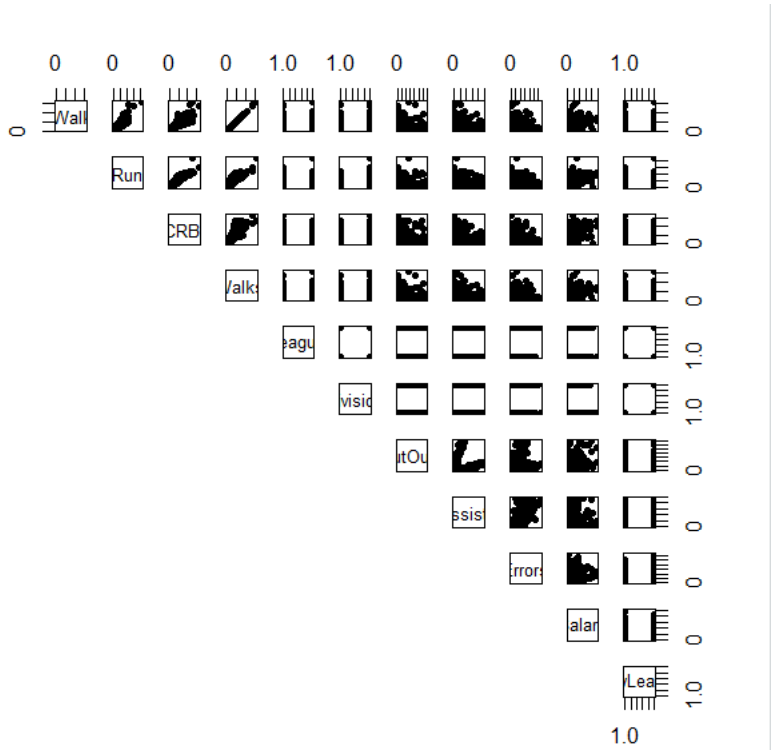
In this project we have taken the Major League Baseball Data (Hitter) for the statistical analysis. The original dataset includes 322 observations of major league players on the 20 variables.

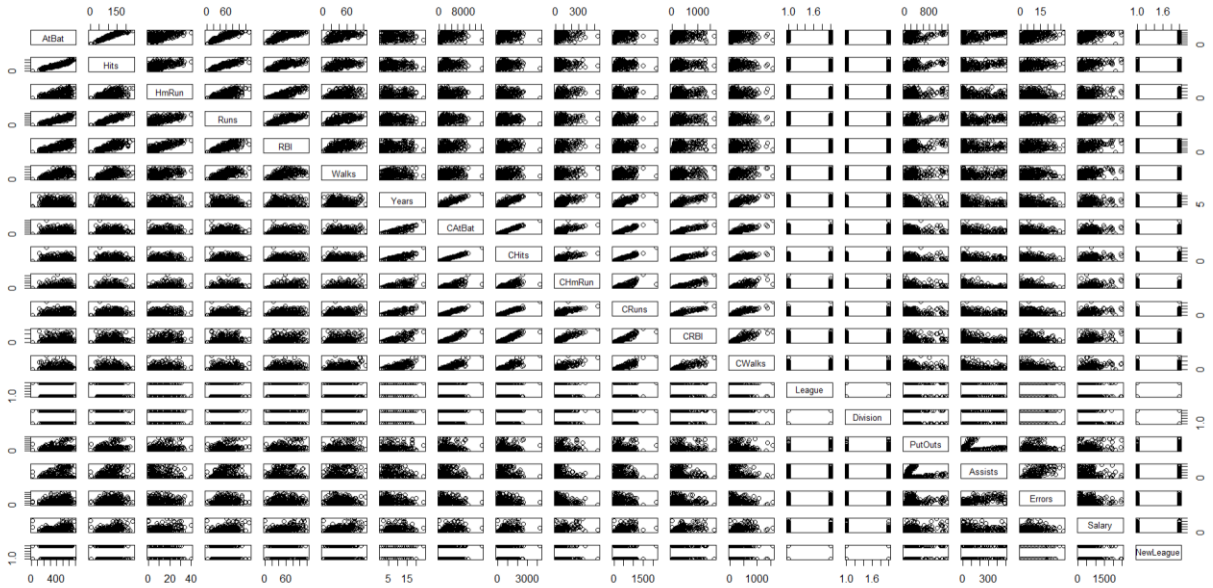
We have used a multiple linear regression model to identify the significant attributes which impacts to the annual salary (Salary) considered as the dependent variable(response)and the other 19 attributes as independent variables (predictors). The model has 0.9581 R-squared and p-value is 2.2×10^{-16}.

1) Find the linear relationship between response variable & predictors using scatter plots.

   After we remove the NA values, we tried to fit the model. Then we tried to plot the model using Scatter plots. Since the Hitter data set is large, we tried to plot the predictors in 2 plots.




0 0 0 0 1.0 1.0 0 0 0 0 1.0
Wall
Run
CRB
Walks
eagu
visio
utOu
ssist
rrors
alar
Lea
0
0 0 0 0 0 0 0 0 0 0 1.0
1.0


AtBat
Hits
HmRun
Runs
RBI
Walks
Years
CAtBat
CHits
CHmRun
CRuns
CRBI
CWalks
League
Division
PutOuts
Assists
Errors
Salary
NewLeague

We may conclude that CRuns, League and Chits are the most important in estimating Salary. From the plot, we can also see multicollinearity amongst dependent variables. Over-fitting can occur when two variables are highly connected; hence this type of variable should be removed.

Both the variables' "Hits" and "Salary" are related. After analyzing the P-Value, it is less than 0.05. Therefore, we can conclude that the relationship between "Hits" and "Salary" is statistically significant. The quality of the linear equation: By using the linear equation to predict salary we can reduce the error by 19%

2) CRuns(159.031220) and CRBI(132.703500) seems to have the most serious variance inflation. So, we removed that to fit the model again.
Note that we have a lot of coefficients with large values. The VIFs indicate that we have a high degree of collinearity.

3) From Best subset regression and stepwise selection (forward, backward, both),we see that all variables except CRBI and Hits are significant.

We see that using forward stepwise selection, the best one variable model contains only CRBI, and the best two-variable model additionally includes Hits. For this data, the best one-variable through six-variable models is each identical for best subset and forward selection. However, the best seven-variable models identified by forward stepwise selection, backward stepwise selection, and best subset selection are different.

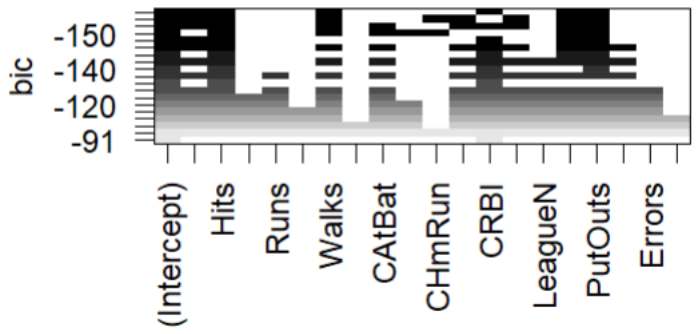
An asterisk indicates that a given variable is included in the corresponding model. For instance, this output indicates that the best two-variable model contains only Hits and CRBI

he "best" model with its corresponding subset size kk is then selected, according to several indicators such as Cp, BIC, R^2

```
Subset selection object
Call: regsubsets.formula(Salary ~ ., data = Hitters, nbest = 1, nvmax = 20)
19 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
CAtBat         FALSE      FALSE
CHits          FALSE      FALSE
CHmRun         FALSE      FALSE
CRuns          FALSE      FALSE
CRBI           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 19
Selection Algorithm: exhaustive
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks LeagueN
```

```
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 )  "*"   "*"  "*"  "*"  " " " " "*"  " "  "*"  "*"  " "  "*"  "*"  "*"  "*"
16 ( 1 )  "*"   "*"  "*"  "*"  "*" "*" "*"  " "  "*"  "*"  " "  "*"  "*"  "*"  "*"
17 ( 1 )  "*"   "*"  "*"  "*"  "*" "*" "*"  " "  "*"  "*"  " "  "*"  "*"  "*"  "*"
18 ( 1 )  "*"   "*"  "*"  "*"  "*" "*" "*"  "*"  "*"  "*"  " "  "*"  "*"  "*"  "*"
19 ( 1 )  "*"   "*"  "*"  "*"  "*" "*" "*"  "*"  "*"  "*"  "*"  "*"  "*"  "*"  "*"
```

```
         DivisionW 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 )  "*"       "*"      "*"     "*"    " "
16  ( 1 )  "*"       "*"      "*"     "*"    " "
17  ( 1 )  "*"       "*"      "*"     "*"    "*"
18  ( 1 )  "*"       "*"      "*"     "*"    "*"
19  ( 1 )  "*"       "*"      "*"     "*"    "*"
```



Here from all these results, R^2 increases from 32% (one-variable) to 55% (all variables). Although different models share a BIC $\approx -150 \approx -150$, there is a single six-variable model containing AtBat, Hits, Walks, CRBI, DivisionW, PutOuts

4) The new model excludes CHits, CAtBat, CRuns, CRBI fits better. Apply the best subset selection method to find a good multiple linear equation.

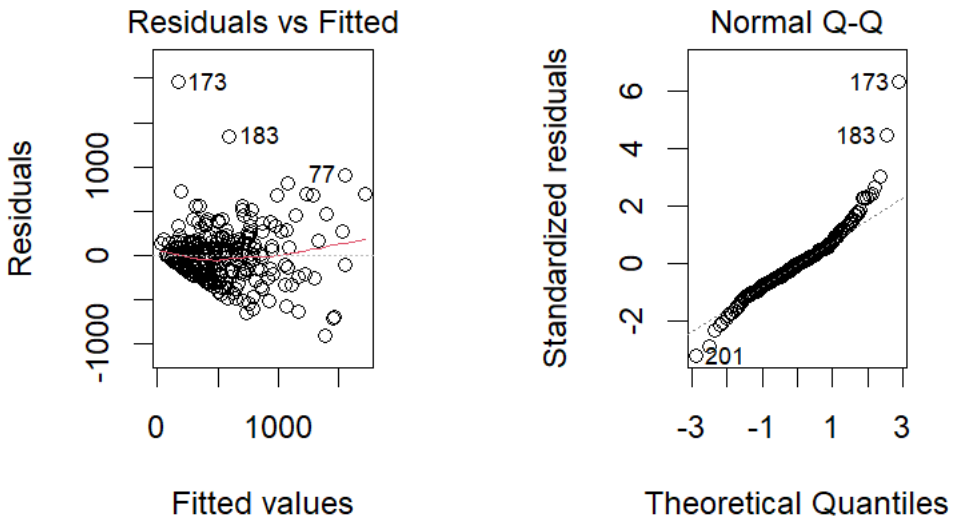
The new fitted model ,Salary = HmRun (-2.67729) + Runs (1.48651) + RBI 0.57047) + Walks(4.49084)+ Years(-17.29272)+ Chits(0.48859)+ CHmRun(1.44429)+ CWalks(-0.54827)+ LeagueN(83.52068)+ DivisionW(-141.80841)+ PutOuts(0.27017)+ Assists(0.19205)+ Errors -5.77686)+ NewLeagueN(-56.93810)

The new fitted models:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -58.12494   83.21607  -0.698 0.485530
Hits           3.41053    1.55669   2.191 0.029391 *
HmRun         -4.03473    5.96215  -0.677 0.499211
Runs          -0.55764    2.69917  -0.207 0.836494
RBI           -0.21957    2.52014  -0.087 0.930642
Walks          2.10936    1.70611   1.236 0.217495
Years          8.94573    9.33712   0.958 0.338954
CHmRun         1.95853    0.55715   3.515 0.000522 ***
CWalks         0.06769    0.21579   0.314 0.754021
LeagueN       76.51654   83.93952   0.912 0.362882
DivisionW   -131.08542   42.26472  -3.102 0.002148 **
PutOuts        0.27151    0.08148   3.332 0.000993 ***
Assists        0.08733    0.22590   0.387 0.699381
Errors        -4.18658    4.62206  -0.906 0.365930
NewLeagueN   -27.93730   83.17258  -0.336 0.737234
---
```

The null hypothesis H0: The coefficients associated with the variables are zero.
The alternate hypothesis H1: The coefficients are not.
R- squared is 0.9581,
F-statistic value is 292.4
p-value: < 2.2e-16 , p value is significantly smaller than 0.05, so we can say that there is no significance .

5) Residual diagnosis methods.



Unlike other variables, leverage does not involve the response(Salary) variable.

From Residual vs Fitted model, which shows, the residual has non -linear patterns or not. There could be a nonlinear relationship between predictor variables and the response .
So here the residuals are almost equally distributed among the horizontal line, so residual shows nonlinear relationships or patterns. The variance is not completely constant; therefore, the assumption of constant variance is not fully satisfied.
Therefore, the assumption is not satisfied, and a quadratic model should be used instead of a linear one.

Here from the QQ plot we can say that residuals don't follows a straight line, therefore, they are not normally distributed.

Scale location plot. This plot shows if residuals are spread equally along the ranges of predictors. But here we can't find a horizontal line with equally spread point.

In Residual vs leverage plot, we tried to find outlying values at the upper right corner or at the lower right corner. 201 and 189 are the possible outliers.