Heet Shah
Shah_heet@bentley.edu

Hitters Case Study

Exploratory Data Analysis:

For this case study, we are given the "Hitters" data set from the "ISLR" Package in R. The goal is to predict a player's salary during the 1986 — 1987 baseball season. This data set contains 322 observations of major baseball league players on 20 variables from the 1986 — 1987 seasons. From exploring the data, we notice that there exist numerous missing values "NA" for the response "Salary". We remove the rows with missing values, and define a new data frame without missing observaitons. The R code is shown below:

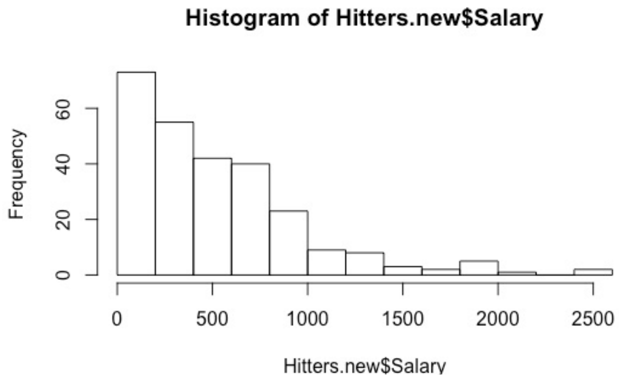
```
#re-define the data set by omitting all missing obs
Hitters.new = na.omit(Hitters)
attach(Hitters.new)
```

Then, we start out with Exploratory Data Analysis (EDA) on our dataset as follows:

First, we did the 5-number summary of the variables in the dataset:

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

From the output above, we notice that we have a mix of quantitative and qualitative variables in our dataset. We can also notice that many of the variables including the response variable (Salary) are skewed in nature. We can further cement the fact using the histogram below:



As we can see from the plot, the Salary variable is highly rightly skewed.

Next, we try to gauge whether there is multicollinearity in the dataset. First, we tried to calculate the VIF of every variable in the data to get the overall severity of the problem:

Multicollinearity:

```
> vifstep(x, th = 1000)
No variable from the 19 input variables has collinearity problem.

The linear correlation coefficients ranges between:
min correlation ( Errors ~ DivisionW ):  -0.0005569954
max correlation ( CHits ~ CAtBat ):  0.9950568

---------- VIFs of the remained variables --------
    Variables        VIF
1       AtBat  22.944366
2        Hits  30.281255
3       HmRun   7.758668
4        Runs  15.246418
5         RBI  11.921715
6       Walks   4.148712
7       Years   9.313280
8      CAtBat 251.561160
9       CHits 502.954289
10     CHmRun  46.488462
11      CRuns 162.520810
12       CRBI 131.965858
13     CWalks  19.744105
14    LeagueN   4.134115
15  DivisionW   1.075398
16    PutOuts   1.236317
17    Assists   2.709341
18     Errors   2.214543
19 NewLeagueN   4.099063
```

As we can see from the output, some of the variables have a really high VIF, much greater than VIF=10 which is usually considered the rule of thumb for removing variables.

We need to find the best first-order additive model for our dataset but since we have high multicollinearity in the data, it would be difficult to find a good model since multicollinearity increases the likelihood of rounding errors in the calculations of the β estimates, standard errors. It might cause non-significant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy is significant. It can also give us opposite signs (from what is expected) in the estimated parameters.

In order to get the best first-order model, we can either drop the highly correlated variables and move forward or we can take the existing variables and do stepwise regression. We cannot establish cause-and-effect relationship and also since our purpose for regression is only for estimation & prediction, we can do stepwise regression as it generally includes only one (or a small number) of a set of multicollinear independent variables will be included in the regression model as it tests the parameter associated with each variable in the presence of all the variables already in the model.

Stepwise Regression:

We tried stepwise regression procedure with backward, forward & with AIC, BIC. We also did best subsets regression comparison. We did all this to find out the best models given by these different procedures and see if there are any common variables or models we can find:

Procedure	Intercept	AtBat	Hits	Walks	CAtBat	CRuns	CRBI
Backward AIC	162.5354	-2.1687	6.9180	5.7732	-0.1301	1.4082	0.7743
Backward BIC	117.1520	-2.0339	6.8549	6.4407		0.7045	0.5273
Forward AIC	162.5354	-2.1687	6.9180	5.7732	-0.1301	1.4082	0.7743
Forward BIC	91.5118	-1.8686	7.6044	3.6976			0.6430

Minimum Cp	-	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Maximum Adj.R^2	-	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Procedure	DivisionW	CWalks	Assists	PutOuts	League
Backward AIC	-112.3801	-0.8308	0.2832	0.2974	
Backward BIC	-123.7798	-0.8066		0.2754	
Forward AIC	-112.3801	-0.8308	0.2832	0.2974	
Forward BIC	-122.9515			0.2643	
Minimum Cp	TRUE	TRUE	TRUE	TRUE	
Maximum Adj.R^2	TRUE	TRUE	TRUE	TRUE	TRUE

From the above tables, we can see that Forward AIC, Backward AIC, Minimum Cp gives us the same model. The model with Maximum Adjusted R^2 is almost the same except for extra League variable. The only difference in the AIC models is the order in which the variables are entered in the model that we see in the R output. So, we go ahead with the backward AIC model since the variables enter in sequence and it's the same given by many other procedures as well:

```
> summary(BackAIC)

Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + Division + PutOuts + Assists, data = Hitters.new)

Residuals:
    Min      1Q  Median      3Q     Max
-939.11 -176.87  -34.08  130.90 1910.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.53544   66.90784   2.429 0.015830 *
AtBat        -2.16865    0.53630  -4.044 7.00e-05 ***
Hits          6.91802    1.64665   4.201 3.69e-05 ***
Walks         5.77322    1.58483   3.643 0.000327 ***
CAtBat       -0.13008    0.05550  -2.344 0.019858 *
CRuns         1.40825    0.39040   3.607 0.000373 ***
CRBI          0.77431    0.20961   3.694 0.000271 ***
CWalks       -0.83083    0.26359  -3.152 0.001818 **
DivisionW  -112.38006   39.21438  -2.866 0.004511 **
PutOuts       0.29737    0.07444   3.995 8.50e-05 ***
Assists       0.28317    0.15766   1.796 0.073673 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 311.8 on 252 degrees of freedom
Multiple R-squared:  0.5405,    Adjusted R-squared:  0.5223
F-statistic: 29.64 on 10 and 252 DF,  p-value: < 2.2e-16
```

```
> anova(BackAIC)
Analysis of Variance Table

Response: Salary
           Df   Sum Sq Mean Sq F value    Pr(>F)
AtBat       1  8309469 8309469 85.4674 < 2.2e-16 ***
Hits        1  2545894 2545894 26.1859 6.152e-07 ***
Walks       1  3850603 3850603 39.6056 1.369e-09 ***
CAtBat      1  8773884 8773884 90.2442 < 2.2e-16 ***
CRuns       1   808877  808877  8.3197 0.0042608 **
CRBI        1  1164332 1164332 11.9758 0.0006328 ***
CWalks      1   798475  798475  8.2127 0.0045109 **
Division    1   870346  870346  8.9520 0.0030467 **
PutOuts     1  1383182 1383182 14.2268 0.0002020 ***
Assists     1   313650  313650  3.2261 0.0736726 .
Residuals 252 24500402   97224
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the assists variable is not significant, so we try fitting the model without it. Following is the revised model we get:

```
> summary(BackAIC2)

Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CRuns + CRBI + CWalks +
    Division + PutOuts, data = Hitters.new)

Residuals:
    Min      1Q  Median      3Q     Max
-794.06 -171.94  -28.48  133.36 2017.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.15204   65.07016   1.800 0.072985 .
AtBat        -2.03392    0.52282  -3.890 0.000128 ***
Hits          6.85491    1.65215   4.149 4.56e-05 ***
Walks         6.44066    1.52212   4.231 3.25e-05 ***
CRuns         0.70454    0.24869   2.833 0.004981 **
CRBI          0.52732    0.18861   2.796 0.005572 **
CWalks       -0.80661    0.26395  -3.056 0.002483 **
DivisionW  -123.77984   39.28749  -3.151 0.001824 **
PutOuts       0.27539    0.07431   3.706 0.000259 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 314.7 on 254 degrees of freedom
Multiple R-squared:  0.5281,    Adjusted R-squared:  0.5133
F-statistic: 35.54 on 8 and 254 DF,  p-value: < 2.2e-16
```

```
> anova(BackAIC1)
Analysis of Variance Table

Response: Salary
           Df   Sum Sq Mean Sq F value    Pr(>F)
AtBat       1  8309469 8309469 84.7220 < 2.2e-16 ***
Hits        1  2545894 2545894 25.9575 6.831e-07 ***
Walks       1  3850603 3850603 39.2601 1.586e-09 ***
CAtBat      1  8773884 8773884 89.4571 < 2.2e-16 ***
CRuns       1   808877  808877  8.2472 0.0044274 **
CRBI        1  1164332 1164332 11.8713 0.0006672 ***
CWalks      1   798475  798475  8.1411 0.0046852 **
Division    1   870346  870346  8.8739 0.0031740 **
PutOuts     1  1383182 1383182 14.1027 0.0002148 ***
Residuals 253 24814051   98079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heet Shah
Shah_heet@bentley.edu

```
> summary(BackAIC1)

Call:
lm(formula = Salary ~ AtBat + Hits + Walks + CAtBat + CRuns +
    CRBI + CWalks + Division + PutOuts, data = Hitters.new)

Residuals:
    Min      1Q  Median      3Q     Max
-907.68 -169.60  -41.25  136.91 1986.56

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  146.24960   66.58161   2.197 0.028960 *
AtBat         -1.93677    0.52281  -3.705 0.000260 ***
Hits           6.65672    1.64741   4.041 7.07e-05 ***
Walks          5.55204    1.58697   3.499 0.000553 ***
CAtBat        -0.09954    0.05306  -1.876 0.061805 .
CRuns          1.25067    0.38208   3.273 0.001211 **
CRBI           0.66177    0.20090   3.294 0.001129 **
CWalks        -0.77798    0.26310  -2.957 0.003400 **
DivisionW   -115.34950   39.35150  -2.931 0.003685 **
PutOuts        0.27773    0.07396   3.755 0.000215 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 313.2 on 253 degrees of freedom
Multiple R-squared:  0.5346,	Adjusted R-squared:  0.5181
F-statistic: 32.29 on 9 and 253 DF,  p-value: < 2.2e-16
```

From the new model without assists, we can see that even CAtBat is not significant, so we try fitting a new model without CAtbat:

```
> anova(BackAIC2)
Analysis of Variance Table

Response: Salary
           Df   Sum Sq Mean Sq F value    Pr(>F)
AtBat       1  8309469 8309469 83.8899 < 2.2e-16 ***
Hits        1  2545894 2545894 25.7026 7.682e-07 ***
Walks       1  3850603 3850603 38.8745 1.873e-09 ***
CRuns       1  9460012 9460012 95.5054 < 2.2e-16 ***
CRBI        1   757067  757067  7.6431 0.0061167 **
CWalks      1   868072  868072  8.7638 0.0033633 **
Division    1  1008416 1008416 10.1807 0.0015977 **
PutOuts     1  1360346 1360346 13.7336 0.0002586 ***
Residuals 254 25159234   99052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we arrive at a model which has all the variables that are significant and has an adjusted R2 of 0.5133.

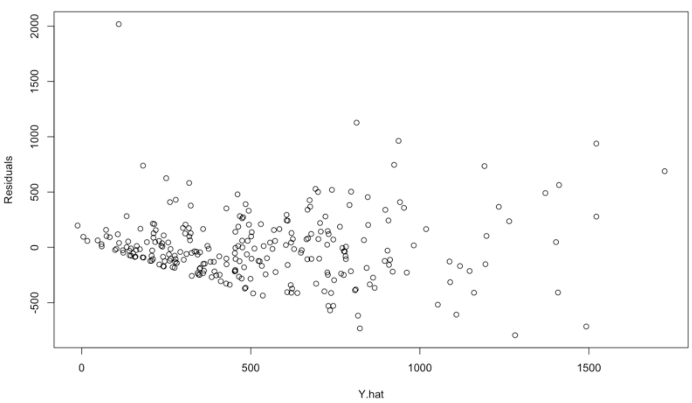
For Assists & CAtBat, I tried Partial F-test to confirm whether the variable needs to be removed & since the p-value was greater than 0.05, we removed the variables.

We check the correlation between the selected variables and we get the following matrix:

```
> x.final <- cbind(Salary, AtBat, Hits, Walks, CRuns, CRBI, CWalks, Division, PutOuts)
> cor(x.final)
             Salary       AtBat        Hits       Walks       CRuns        CRBI
Salary    1.0000000  0.39477094  0.43867474  0.44386726  0.56267771  0.56696569
AtBat     0.3947709  1.00000000  0.96396913  0.62444813  0.23727777  0.22139318
Hits      0.4386747  0.96396913  1.00000000  0.58731051  0.23889610  0.21938423
Walks     0.4438673  0.62444813  0.58731051  1.00000000  0.33297657  0.31269680
CRuns     0.5626777  0.23727777  0.23889610  0.33297657  1.00000000  0.94567701
CRBI      0.5669657  0.22139318  0.21938423  0.31269680  0.94567701  1.00000000
CWalks    0.4898220  0.13292568  0.12297073  0.42913990  0.92776846  0.88913701
Division -0.1925144 -0.05634123 -0.08326647 -0.07273229 -0.04681232 -0.02156384
PutOuts   0.3004804  0.30960746  0.29968754  0.28085548  0.05908718  0.09537515
              CWalks    Division     PutOuts
Salary    0.48982204 -0.19251440  0.30048036
AtBat     0.13292568 -0.05634123  0.30960746
Hits      0.12297073 -0.08326647  0.29968754
Walks     0.42913990 -0.07273229  0.28085548
CRuns     0.92776846 -0.04681232  0.05908718
CRBI      0.88913701 -0.02156384  0.09537515
CWalks    1.00000000 -0.05049393  0.05816016
Division -0.05049393  1.00000000 -0.02535144
PutOuts   0.05816016 -0.02535144  1.00000000
```

We see that there is still high correlation among a couple of variables (like AtBat & Hits – 96.39%, CRuns & RBI & CWalks, etc.)

We won't remove the variables at this stage because they might be significant later on when we find interaction or maybe higher order terms. Since we are using the model only for prediction & estimation purposes, there is no need to remove them.



Check for homoscedasticity:

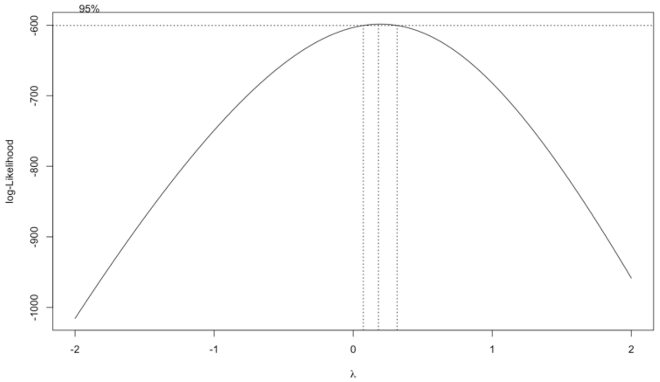
Next, we check the homoscedasticity of the model to see if the assumption is violated or not.

Here, we find that the variance is not constant and there is a cone like pattern in the residuals v/s y.hat plot.

So, we decided to go ahead with Box-cox transformation to find out which transformation would be appropriate to make the model more robust:

Heet Shah
Shah_heet@bentley.edu

Box-Cox Transformation:

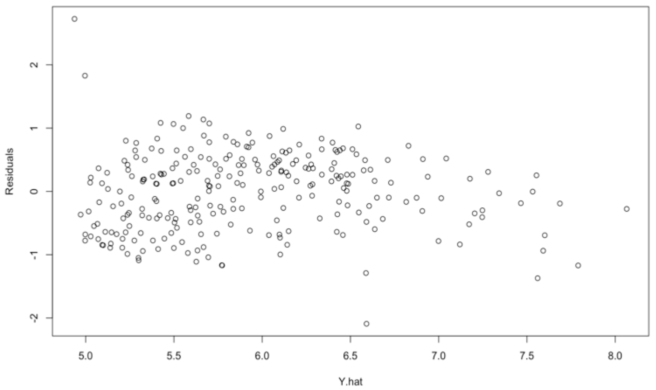


Here, from the graph on the side, we can see that the peak of the graph has a lambda value close to 0.

So, a log transformation would be more appropriate for the model.

We can try transforming the response variable and see how the model turns out.

After transforming the Salary variable, the residual plot shows no trend and so we can go forward with the new model to find variables that are significant.



We tried summary & anova on the new model to check whether the variables are still significant or not.

```
> anova(TFinalmodel)
Analysis of Variance Table

Response: S_log
           Df Sum Sq Mean Sq  F value    Pr(>F)    
AtBat       1 35.668  35.668  92.2655 < 2.2e-16 ***
Hits        1  7.199   7.199  18.6217 2.285e-05 ***
Walks       1 11.501  11.501  29.7520 1.164e-07 ***
CRuns       1 49.328  49.328 127.6006 < 2.2e-16 ***
CRBI        1  0.682   0.682   1.7636   0.18537    
CWalks      1  1.230   1.230   3.1819   0.07565 .  
Division    1  1.759   1.759   4.5492   0.03389 *  
PutOuts     1  1.597   1.597   4.1316   0.04313 *  
Residuals 254 98.191   0.387                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(TFinalmodel)

Call:
lm(formula = S_log ~ AtBat + Hits + Walks + CRuns + CRBI + CWalks + 
    Division + PutOuts, data = Hitters.new)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.09129 -0.43034  0.09065  0.42576  2.72597 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.9473135  0.1285489  38.486  < 2e-16 ***
AtBat       -0.0027403  0.0010328  -2.653 0.008478 ** 
Hits         0.0114523  0.0032639   3.509 0.000532 ***
Walks        0.0084330  0.0030070   2.804 0.005430 ** 
CRuns        0.0016064  0.0004913   3.270 0.001225 ** 
CRBI         0.0005194  0.0003726   1.394 0.164588    
CWalks      -0.0009635  0.0005215  -1.848 0.065806 .  
DivisionW   -0.1638332  0.0776141  -2.111 0.035761 *  
PutOuts      0.0002984  0.0001468   2.033 0.043129 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6218 on 254 degrees of freedom
Multiple R-squared:  0.526,	Adjusted R-squared:  0.5111 
F-statistic: 35.23 on 8 and 254 DF,  p-value: < 2.2e-16
```

From the output, we notice that the variables CRBI & CWalks are not significant in the model.

We tested a new model without these variables using Partial F-test and we find that the variables are indeed not significant and we should remove them.

So, finally we come to a model with 6 variables in the model. Following is the summary & the anova output of the final model:

Heet Shah
Shah_heet@bentley.edu

```
> summary(TFinalmodel1)

Call:
lm(formula = S_log ~ AtBat + Hits + Walks + CRuns + PutOuts +
    Division, data = Hitters.new)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21242 -0.46775  0.09048  0.41589  2.76425

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.9025524  0.1274481  38.467  < 2e-16 ***
AtBat       -0.0024877  0.0010311  -2.413 0.016542 *
Hits         0.0120917  0.0032517   3.719 0.000246 ***
Walks        0.0049485  0.0023775   2.081 0.038390 *
CRuns        0.0014015  0.0001241  11.290  < 2e-16 ***
PutOuts      0.0003214  0.0001463   2.197 0.028891 *
DivisionW   -0.1534399  0.0778305  -1.971 0.049748 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6251 on 256 degrees of freedom
Multiple R-squared:  0.517,	Adjusted R-squared:  0.5057
F-statistic: 45.68 on 6 and 256 DF,  p-value: < 2.2e-16
```

```
> anova(TFinalmodel1)
Analysis of Variance Table

Response: S_log
           Df  Sum Sq Mean Sq  F value    Pr(>F)
AtBat       1  35.668  35.668  91.2677 < 2.2e-16 ***
Hits        1   7.199   7.199  18.4203 2.515e-05 ***
Walks       1  11.501  11.501  29.4303 1.342e-07 ***
CRuns       1  49.328  49.328 126.2208 < 2.2e-16 ***
PutOuts     1   1.894   1.894   4.8456   0.02861 *
Division    1   1.519   1.519   3.8867   0.04975 *
Residuals 256 100.046   0.391
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now this is the first order model we have arrived at finally which has variables which are significant and it complies with the constant variance assumption.

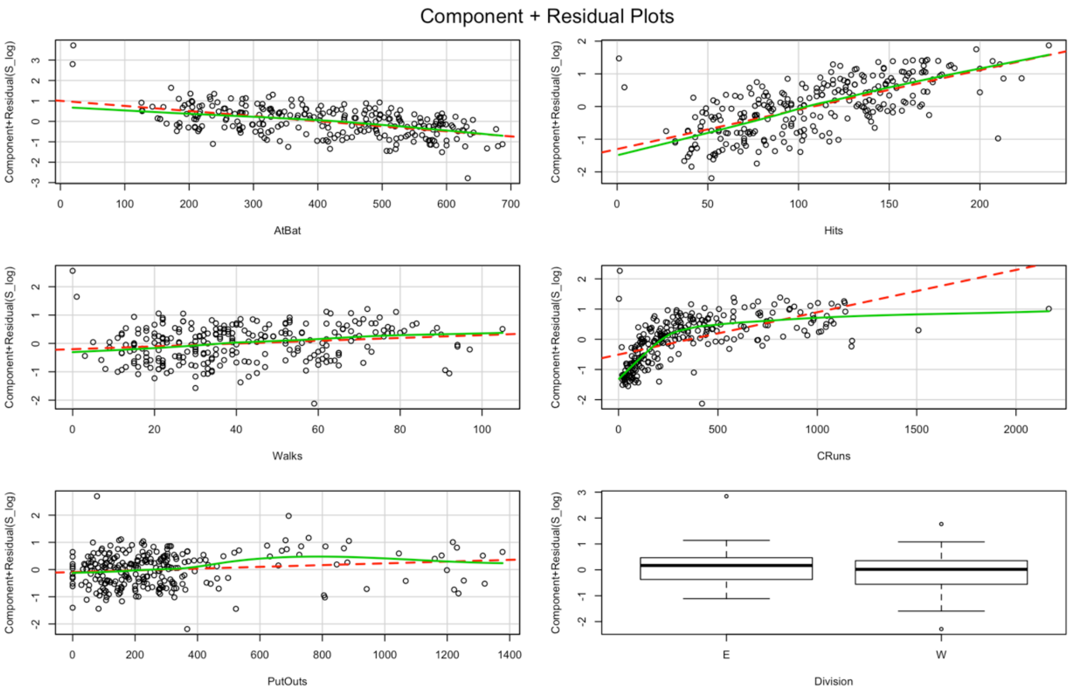
Now, we go ahead and check the correlation among these variables to gauge for multicollinearity:

```
> cor(x.final1)
               AtBat        Hits       Walks       CRuns     PutOuts    Division
AtBat     1.00000000  0.96396913  0.62444813  0.23727777  0.30960746 -0.05634123
Hits      0.96396913  1.00000000  0.58731051  0.23889610  0.29968754 -0.08326647
Walks     0.62444813  0.58731051  1.00000000  0.33297657  0.28085548 -0.07273229
CRuns     0.23727777  0.23889610  0.33297657  1.00000000  0.05908718 -0.04681232
PutOuts   0.30960746  0.29968754  0.28085548  0.05908718  1.00000000 -0.02535144
Division -0.05634123 -0.08326647 -0.07273229 -0.04681232 -0.02535144  1.00000000
```

From the table, we can see that there is still high correlation among Atbat & Hits. Atbat is the number of times a player came to bat in 1986 & Hits are the number of hits the player made. So, it is natural for the variables to be correlated since one cannot happen without the other. We can try adding an interaction term with these variables to check if it is significant.

Partial Residual Plot:

We plotted partial residual plot of all the predictors in the model to check for any untoward influence of Xj on response y after effects of other independent variables accounted for.



From the partial residual plot, we can see the most of the variables stick close to the least square line but variables like CRuns & PutOuts show curvilinear trend which we can investigate further by adding interaction or higher order terms.

The terms CRuns shows a concave down, increasing trend which looks a lot like a logarithm function so we can try adding a log of CRuns to see if we can improve the model fit.

New variable added (Years):

After taking into account the variables in the current model, one of the variables missing in the model which I feel should be included is Years. Years is the number of years the player has played in major leagues. It is quite logical that a player who has played a lot of leagues & has experience will earn more salary. I tried inserting the value and found the value to be significant in the new model tested using Partial F-test.

Final model:

After a lot of combinations of interaction terms & higher order terms, I came to the following model as my final model:

```
> TFinalmodel2 <- lm(formula = S_log ~ AtBat + Hits + log(CRuns) +
+                       PutOuts + Hits*log(CRuns) + AtBat*log(CRuns)
+                     + Hitsq*AtBat + Years + Yearsq, data = Hitters.new)
> summary(TFinalmodel2)

Call:
lm(formula = S_log ~ AtBat + Hits + log(CRuns) + PutOuts + Hits *
    log(CRuns) + AtBat * log(CRuns) + Hitsq * AtBat + Years +
    Yearsq, data = Hitters.new)

Residuals:
     Min       1Q   Median       3Q      Max
-2.46421 -0.21816 -0.00106  0.25211  1.54609

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.770e+00  3.071e-01  18.788  < 2e-16 ***
AtBat            -1.461e-02  4.324e-03  -3.379 0.000843 ***
Hits             -2.571e-02  1.770e-02  -1.453 0.147520
log(CRuns)        1.981e-01  7.547e-02   2.625 0.009204 **
PutOuts           3.399e-04  1.019e-04   3.334 0.000984 ***
Hitsq             5.735e-04  9.457e-05   6.064 4.83e-09 ***
Years             9.405e-02  3.145e-02   2.990 0.003065 **
Yearsq           -6.943e-03  1.223e-03  -5.675 3.80e-08 ***
Hits:log(CRuns)  -9.037e-03  2.926e-03  -3.088 0.002239 **
AtBat:log(CRuns)  3.860e-03  7.881e-04   4.898 1.74e-06 ***
AtBat:Hitsq      -5.151e-07  8.549e-08  -6.026 5.94e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4305 on 252 degrees of freedom
Multiple R-squared:  0.7745,    Adjusted R-squared:  0.7656
F-statistic: 86.57 on 10 and 252 DF,  p-value: < 2.2e-16
```

Heet Shah
Shah_heet@bentley.edu

```
> anova(TFinalmodel2)
Analysis of Variance Table

Response: S_log
                 Df Sum Sq Mean Sq  F value    Pr(>F)    
AtBat             1 35.668  35.668 192.4520 < 2.2e-16 ***
Hits              1  7.199   7.199  38.8420 1.919e-09 ***
log(CRuns)        1 73.334  73.334 395.6865 < 2.2e-16 ***
PutOuts           1  4.178   4.178  22.5413 3.453e-06 ***
Hitsq             1  2.686   2.686  14.4933 0.0001767 ***
Years             1  0.187   0.187   1.0105 0.3157511    
Yearsq            1 12.411  12.411  66.9678 1.374e-14 ***
Hits:log(CRuns)   1 15.687  15.687  84.6414 < 2.2e-16 ***
AtBat:log(CRuns)  1  2.371   2.371  12.7933 0.0004170 ***
AtBat:Hitsq       1  6.729   6.729  36.3087 5.943e-09 ***
Residuals       252 46.704   0.185                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the model, we can notice that the value of adjusted R^2 has increased to 76.56% so the selected model explains 76.56% of sample variation in response variable ln(Salary) after accounting for sample size & no. of predictors in the model.

We can test the overall utility/ model adequacy using the global F-test and from the output, we can see that the value of F-statistic is 86.57 where the p-value is close to 0 so we reject H_0 so therefore, the model is useful in predicting the value of ln(y).

The value of regression coefficients of all predictor variables have changed due to log transformation of the response variable. We can find the out actual salary value taking anti-log of the prediction we get from the model. The endpoints of the prediction interval are similarly transformed back to the original scale, and the interval will retain its meaning. In repeated use, the intervals will contain the observed y-value $100(1-\alpha)$ % of the time.

The value of beta-coefficients of PutOuts is the expected change in log of y with respect to a one-unit increase in value of PutOuts holding all other variables constant.

The value of beta coefficients of AtBat, Hits, log(CRuns), Years would be difficult since we have interaction terms in the model with these variables. With interaction terms, the effect of 1 predictor will depend on the value of other predictors, even if they are held constant.

The intercept of the model is 0.577 which is predicted value of salary when all the other x-variables in the model are equal to 0.

For the variable **PutOuts**, we can say that for a one-unit increase in **PutOuts**, we expect to see about a 0.03% increase in Salary, since exp(0.0003399) = 1.0003.

Now, let's focus on the effect of **CRuns**. But, we have an interaction term in the model with log(CRuns) & Hits, AtBat so we can say that effect of 1 predictor affects another so with an added interaction term, the effect of log of CRuns is different for different AtBat score & Hits score.

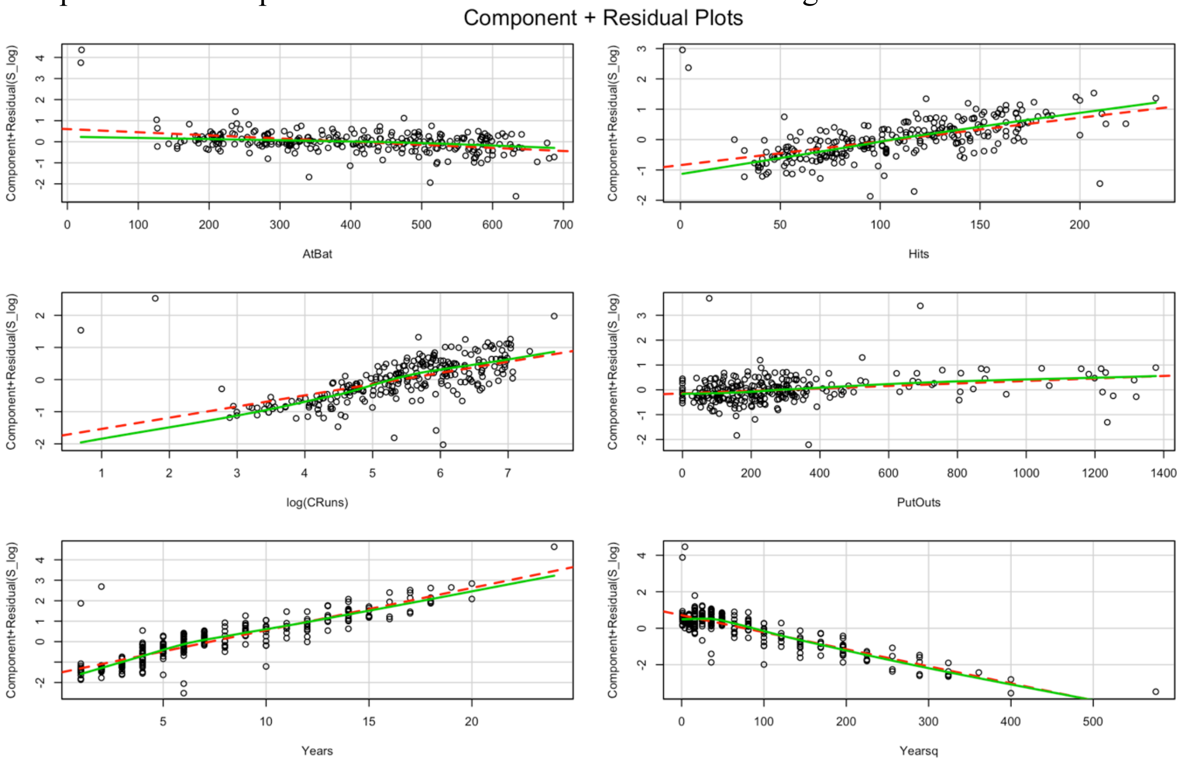
We have 2 quadratic terms which are Hits & Years as well as they were quite significant in the model after looking at the residual plots and partial residual plots.

Heet Shah
Shah_heet@bentley.edu

Model Diagnostics:

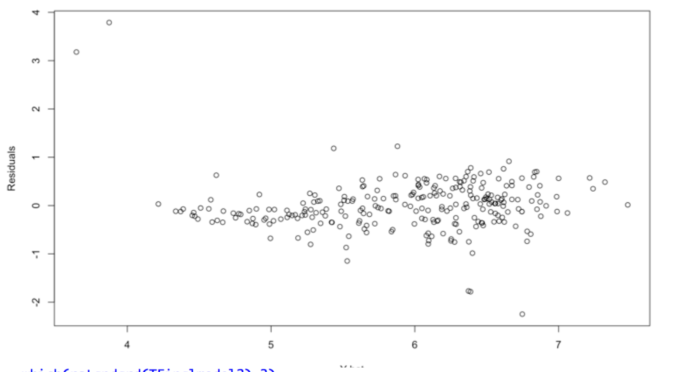
1) Homoscedasticity

The partial residual plot for our final model looks like the one given below:



All the variables seem to follow a linear trend now and since we added interaction & higher order terms, log transformed some of the x & y variables, the assumption of constant variance is maintained as we can see below:

Outlying & Influential Observations:




```
> which(rstandard(TFinalmodel2)>3)
 -Mike Schmidt -Terry Kennedy
          173            241
> which(abs(rstandard(TFinalmodel2))>3)
-Jeffrey Leonard    -Mike Schmidt   -Steve Balboni        -Steve Sax   -Terry Kennedy
             120              173              220               230              241
> which(hatvalues(TFinalmodel2)>2*(9/263))
  -Bill Buckner   -Don Mattingly  -Graig Nettles    -Mike Schmidt      -Pete Rose
             21               62              92              173             189
-Reggie Jackson    -Steve Garvey  -Terry Kennedy     -Ted Simmons     -Wade Boggs
            201              226             241              249             256
> which(cooks.distance(TFinalmodel2)>0.5)
 -Mike Schmidt -Terry Kennedy
          173            241
> Hitters.new1 <- Hitters.new[c(-173,-241),]
> TFinalmodel3 <- lm(formula = S_log ~ AtBat + Hits + log(CRuns) +
+                      PutOuts + Hits*log(CRuns) + AtBat*log(CRuns)
+                    + Hitsq*AtBat + Years + Yearsq, data = Hitters.new1)
```

We notice that there seem to be few outliers in the model so we tested them against standardized residuals, leverage & cook's distance of each of the observations to see if any of them are influential.

We find that there were quite a few observations that were flagged by the model that exceed the rule of thumb for standardized residuals & also leverage that shows outlying observations in the predictor dimension.

But, we can't really know if the observations are influential until we find the cook's distance for them and when we found out that there are 2 observations that are affecting the model fit, so we removed those observations and tried fitting the model again to see if there is any great deal of difference in the statistical inference of the model.

```
> summary(TFinalmodel3)

Call:
lm(formula = S_log ~ AtBat + Hits + log(CRuns) + PutOuts + Hits *
    log(CRuns) + AtBat * log(CRuns) + Hitsq * AtBat + Years +
    Yearsq, data = Hitters.new1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32953 -0.21276  0.01028  0.22102  1.08123

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.587e+00  4.283e-01   8.376 3.97e-15 ***
AtBat           -8.785e-03  4.090e-03  -2.148 0.032691 *
Hits            -5.183e-03  1.659e-02  -0.312 0.755021
log(CRuns)       4.511e-01  7.992e-02   5.644 4.49e-08 ***
PutOuts          3.670e-04  9.423e-05   3.895 0.000126 ***
Hitsq            2.471e-04  9.888e-05   2.499 0.013083 *
Years            9.207e-02  2.893e-02   3.183 0.001643 **
Yearsq          -7.042e-03  1.125e-03  -6.260 1.66e-09 ***
Hits:log(CRuns) -4.273e-03  2.783e-03  -1.535 0.125949
AtBat:log(CRuns) 2.022e-03  7.755e-04   2.607 0.009677 **
AtBat:Hitsq     -2.320e-07  8.857e-08  -2.619 0.009349 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3958 on 250 degrees of freedom
Multiple R-squared:  0.8073,	Adjusted R-squared:  0.7996
F-statistic: 104.8 on 10 and 250 DF,  p-value: < 2.2e-16
```

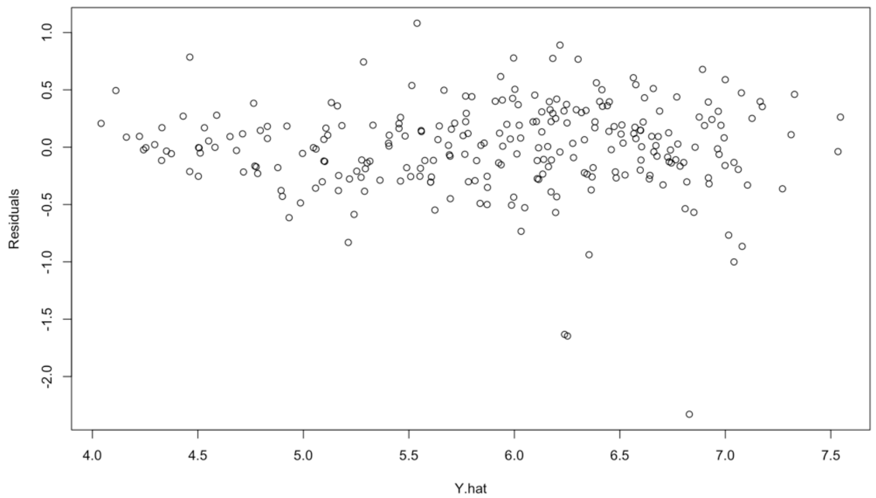
```
> anova(TFinalmodel3)
Analysis of Variance Table

Response: S_log
                  Df Sum Sq Mean Sq  F value    Pr(>F)
AtBat              1 43.213  43.213 275.8173 < 2.2e-16 ***
Hits               1  6.214   6.214  39.6609 1.350e-09 ***
log(CRuns)         1 95.495  95.495 609.5114 < 2.2e-16 ***
PutOuts            1  2.370   2.370  15.1280 0.0001288 ***
Hitsq              1  0.048   0.048   0.3035 0.5821799
Years              1  3.323   3.323  21.2095 6.560e-06 ***
Yearsq             1  8.982   8.982  57.3267 7.119e-13 ***
Hits:log(CRuns)    1  2.916   2.916  18.6095 2.311e-05 ***
AtBat:log(CRuns)   1  0.507   0.507   3.2388 0.0731190 .
AtBat:Hitsq        1  1.075   1.075   6.8609 0.0093487 **
Residuals        250 39.169   0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the model is still significant and the overall model is still useful looking at the value of F-stat and the corresponding p-value.

The value of adjusted R^2 has improved a lot as we can from the that it has risen to 79.96%. Some of the first order variables are not significant after adding interaction terms but their interaction terms are significant, this might just be the fallacy of multicollinearity present in the model.

We again plot the residuals v/s y.hat to see if the plot has improved:

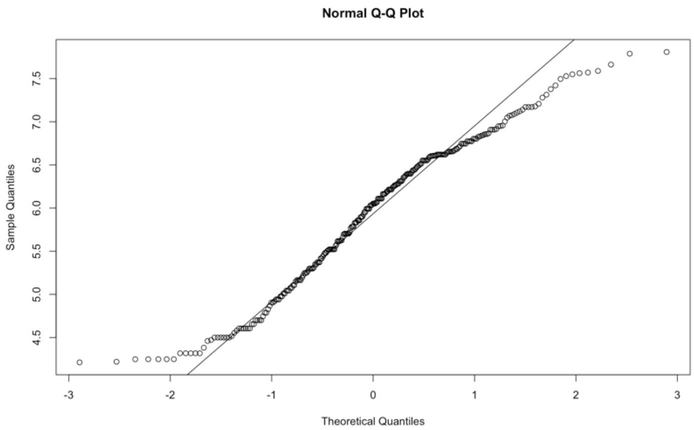


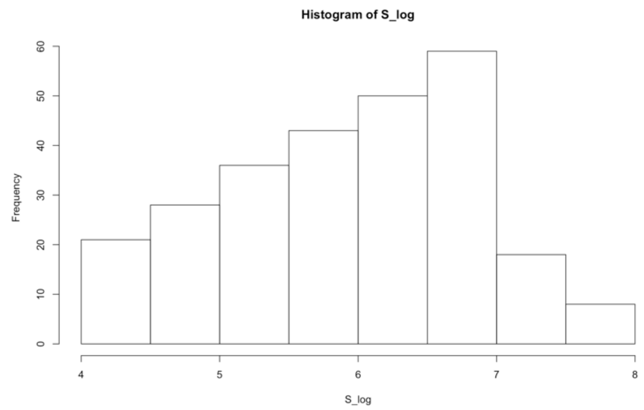
We notice that the plot is much better after the removal of those 2 observations and there is no trend observed in the plot.

2) **Normality Assumption:**

Lastly, we check the normality assumption of the response variable using histogram and Q-Q plot:

Heet Shah
Shah_heet@bentley.edu





From the histogram & Q-Q plot, we can see that the response variable is majorly normal except for a little skewness on the right side.

The histogram is also quite normal shaped.

Little or minor departures from the assumptions of normality are fine as long as the model fit is not affected by a significant effect & we can interpret the model well.

The skewness in the plot might just be due to the fact that the real world data is skewed and removing a lot of outliers just to make the normal wouldn't give a very good estimable model.