Statistical Analysis of Major League Baseball Salaries

A Project Report
Submitted in partial fulfillment of
the requirements for the course
CL 672 - Applied Multivariate Statistics in Chemical Engineering
by

Akhil Nasser
(140020117)

Department of Chemical Engineering
Indian Institute of Technology Bombay
Mumbai 400076 (India)

19 April 2018

Acceptance Certificate

Department of Chemical Engineering
Indian Institute of Technology, Bombay

The project report entitled "Statistical Analysis of Major League Baseball Salaries" submitted by Akhil Nasser (140020117) may be accepted for being evaluated.

Date: 19 April 2018

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I declare that I have properly and accurately acknowledged all sources used in the production of this report. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Akhil Nasser
(140020117)

Date: 19 April 2018

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Baseball and Statistics have a long history together. The use of Statistics in Baseball is predominantly due to the rich availability data from the sixties. In this report we will develop a model to predict the salaries of baseball pitchers in 1987 based on their previous performance statistics. The model will be developed, verified and be used for predicting salaries.

Chapter 2

Data Processing & Input Data Analysis

The Data for Baseball pitchers is widely publicized and available. Our First step is to clean the data i.e remove those observations for which certain data points of the players are not available. This led to the removal of 30 data points i.e players.

We next split the available data into two groups one is a modelling set and the other is a validation set. The modelling set is used to construct our model whose performance will be tested against the validation set. This is done in **R** with the help of the `DUPLEX algorithm` as part of the `prospectr` package. We keep 40 points for validation. The Algorithm sorts out the Data points in a way that both the modelling and validation sets have approximately equally distant points from a "mean". This is to say to prevent any skew in these sets due to composition of points with only low variability or high variability.

We are now going to work on the modelling Data set alone.

The modelling data consists of players with various performance figures. Our aim in this report is to develop a model for the Salaries of Players or more specifically pitchers in 1987. From now on whenever we refer to players we mean pitchers which is a class of player in baseball. The Salary of Player in 1987 in 1000's of dollars is referred to as the Outcome Variable as this is what we seek to model.

2.1 Analysis of Outcome Variable

We would ideally like our Outcome variable to be normal or close to normal of which symmetry is an indicator. The need for normality arises because of may of the subsequent test like Hypothesis tests require normality of data. A Histogram plot of Salary of Players in 1987 is displayed below.

As is clearly evidenced the histogram is non-symmetric and all the values are clustered to the left end of the plot. We further take a look at the Q-Q plot, shown below.

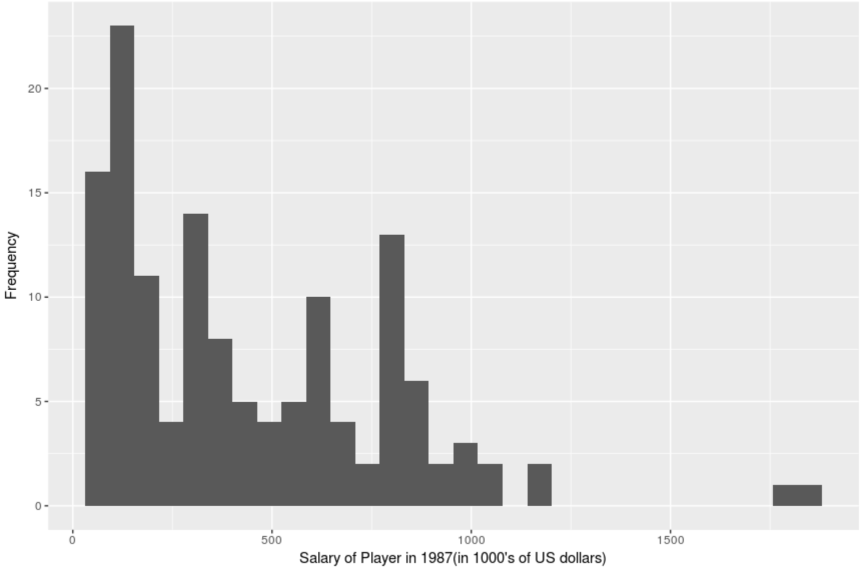


Figure 2.1: Histogram of Outcome Variable

We see a tail at the bottom that deviates away from the Normality line. This Variable is neither linear nor symmetric. In order to achieve these criteria we now undertake transformations of the Outcome Variable. These Transformations have to be chosen such that the clustered values at the lower end get spaced out. This can be achieved with Logarithm Transformations as well as n_{th} root Transformations.

1. **Ln Transformation**

   We take natural logarithm of all the Data points in Player Salary in 1987 and repeat the above procedure i.e plot the Histogram and Q-Q plots

   We observe that the Natural Logarithm Transformation results in a more symmetric distribution. The Q-Q plot which was "heavy tailed" towards a single side previously is now evened out to smaller tails on both sides.

2. **Square Root Transformation**

   We take square root of all the Data points in Player Salary in 1987 and repeat the above procedure i.e plot the Histogram and Q-Q plots

   We observe that the Square Root Transformation results in a more symmetric distribution. The Q-Q plot which was "heavy tailed" towards a single side previously is now evened out to smaller tails on both sides. On Comparison with the Natural Logarithm Transformation, the Q-Q plot looks similar but the Histogram Plot of the Logarithm Transformation is far more symmetric.

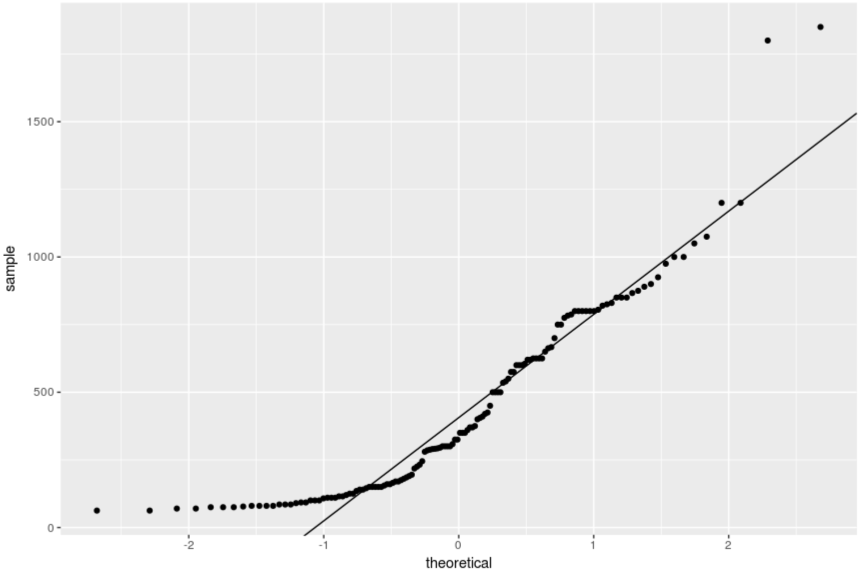
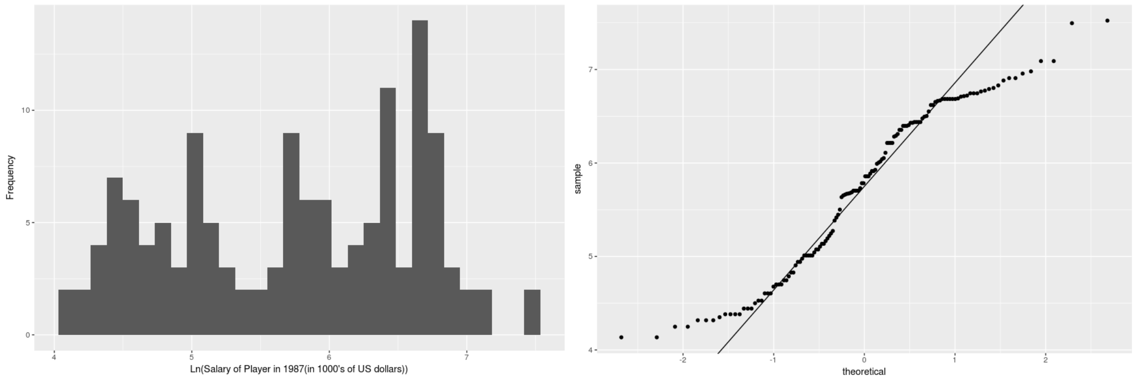


Figure 2.2: Q-Q plot for Outcome Variable



(a) Histogram of Ln(Outcome Variable)

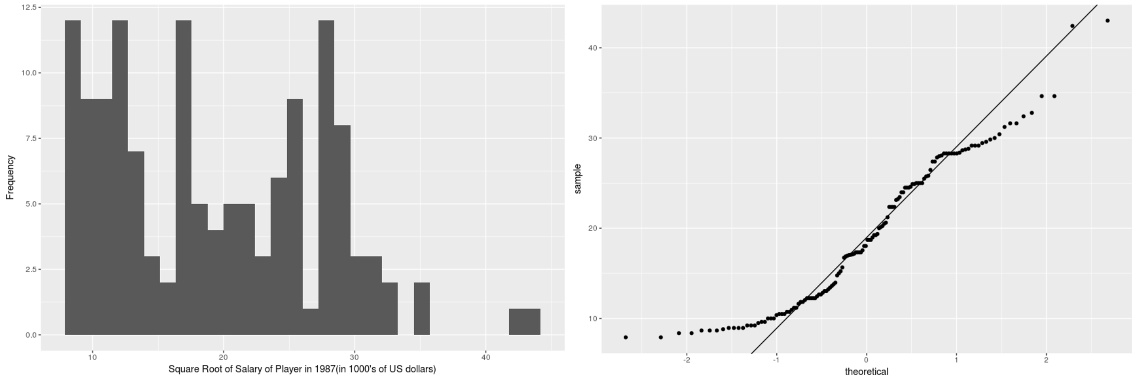
(b) Q-Q plot for Ln(Outcome Variable)

Figure 2.3: Symmetry and Normality Checks of Ln(Outcome Variable)

3. **Cube Root Transformation**

   We take cube root of all the Data points in Player Salary in 1987 and repeat the above procedure i.e plot the Histogram and Q-Q plots

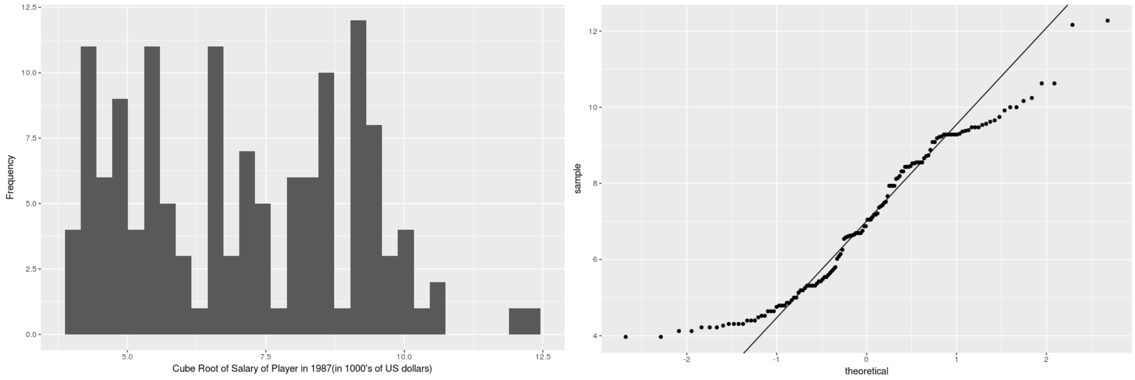
   We observe that the Cube Root Transformation results in a more symmetric distribution. The Q-Q plot which was "heavy tailed" towards a single side previously is now evened out to smaller tails on both sides. On Comparison with the Natural Logarithm Transformation, the Q-Q plot looks similar but the Histogram Plot of the Logarithm Transformation is far better as it has less extreme values away from the center. The Cube Root Transformation is definitely an improvement over the Square Root Transformation as is evidenced by the increasing symmetry of the Histogram Plot.



(a) Histogram of Sqrt(Outcome Variable)

(b) Q-Q plot for Sqrt(Outcome Variable)

Figure 2.4: Symmetry and Normality Checks of Sqrt(Outcome Variable)



(a) Histogram Plot

(b) Q-Q plot

Figure 2.5: Symmetry and Normality Checks of Cube Root of Outcome Variable

For Further Analysis, we are going to go ahead with the Natural Logarithm Transformation of Salary Data for reasons explained previously.

Chapter 3

Confidence Intervals & Hypothesis Testing

In this chapter we construct Confidence Ellipse, Simultaneous Confidence Interval and the Bonferroni Confidence Interval for means of variables of Interest. In the second part of the chapter we perform Multivariate Hypothesis Test for means.

3.1 Confidence Intervals

From the Given Data, we are supposed to develop a model for Salary of Player in 1987. This is as per general knowledge is most likely to depend on Number of Career Games, Number of Career Wins and Number of Years in baseball. We are interested in constructing the Confidence Ellipses, Simultaneous CI and Bonferroni CI.

3.1.1 Player Salary in 1987 with Number of Career Games

We would like to investigate this relation further. Player Salary implies the Salary after undergoing Logarithmic Transformation.

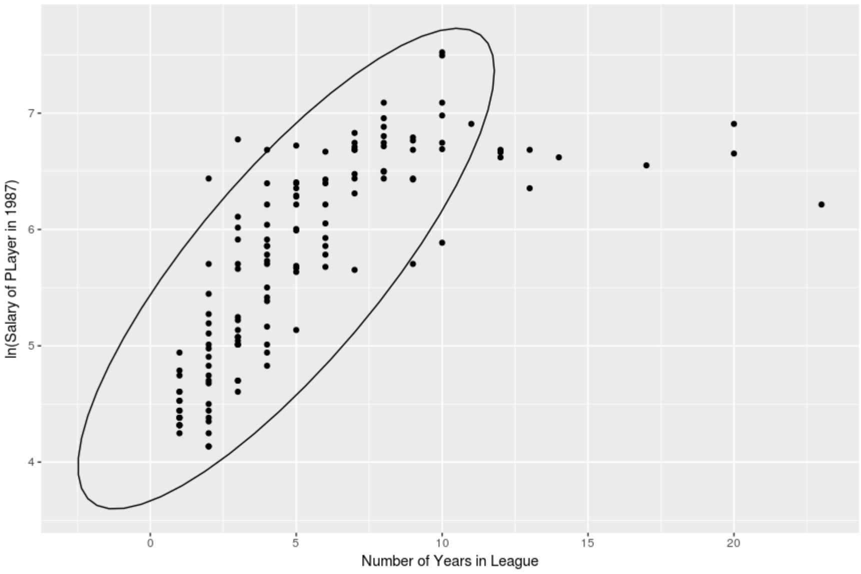


Figure 3.1: Confidence Ellipse between Number of Years in League and ln(Salary in 1987)

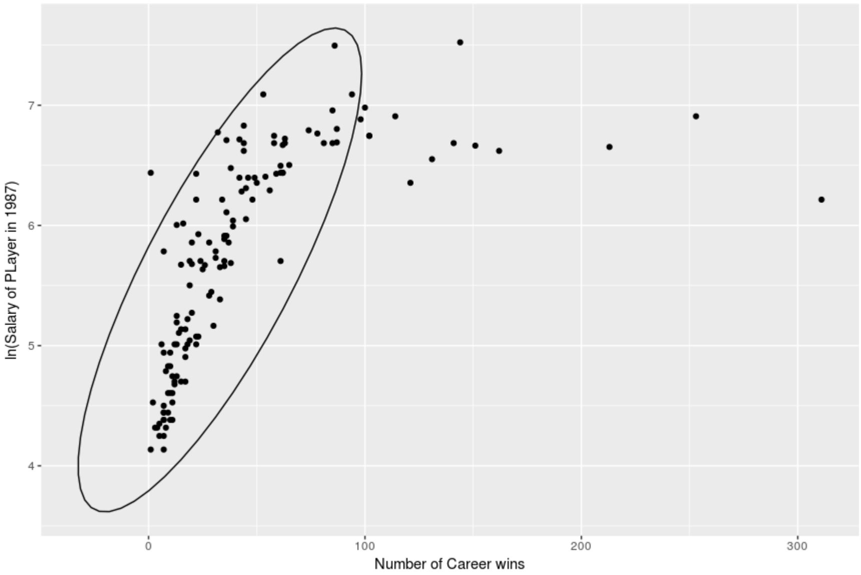


Figure 3.2: Confidence Ellipse between Number of Career wins and ln(Salary in 1987)

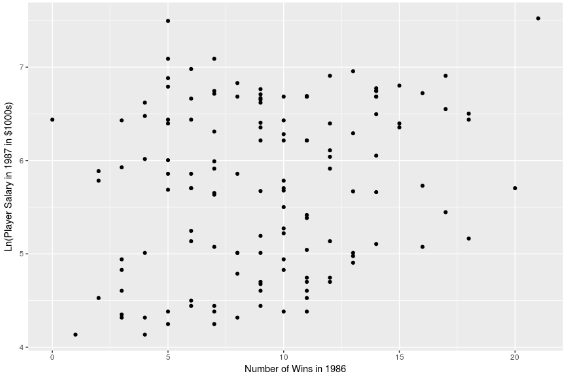
Chapter 4

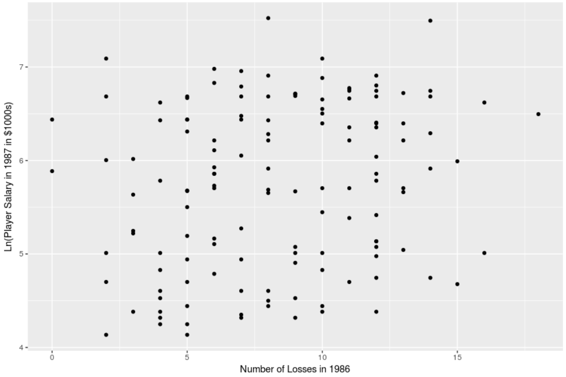
Model Development

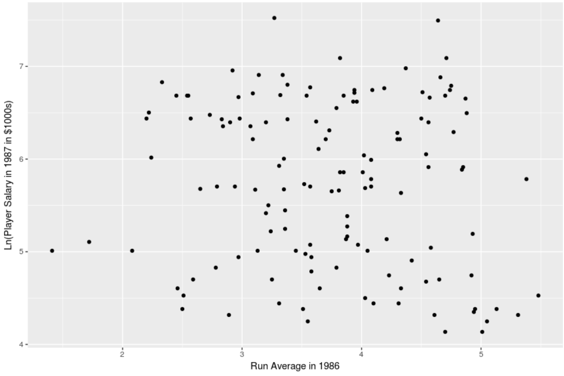
In this chapter, a linear regression model is developed that can be used to predict the Salary of Baseball pitchers in 1987 using the Data Given. The first section explains the reasons for the selection of the linear regression model. In the second section, we look at the histograms of the regressors (X_i) and scatter plots of regressors (X_i) with the outcome variable (Y) and develop inferences from them. The third section deals with the

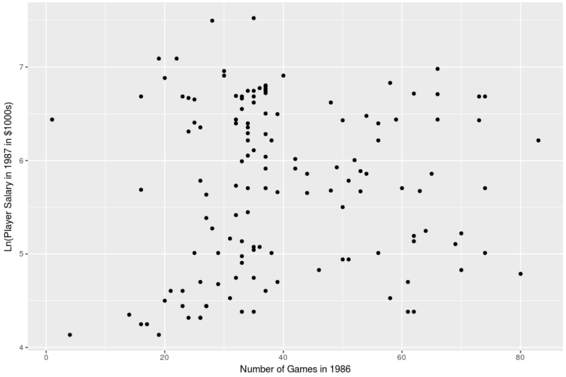
4.1 Regression Checklist

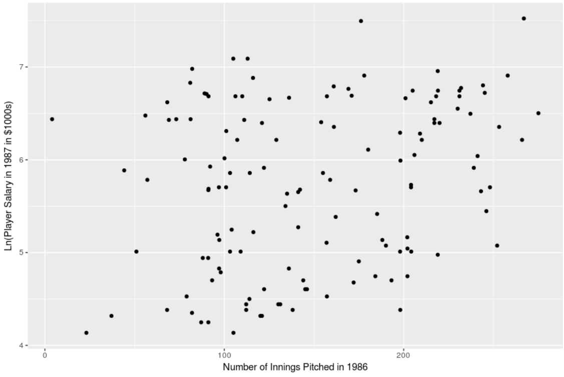
The Scatter plots of regressors vs. the Outcome Variable (ln(Salary)) is presented and inferences are made from these plots. We would want the scatter plots to have ideally a linear relationship or at worst a non conclusive relationship. All the Scatter Plots are available below.

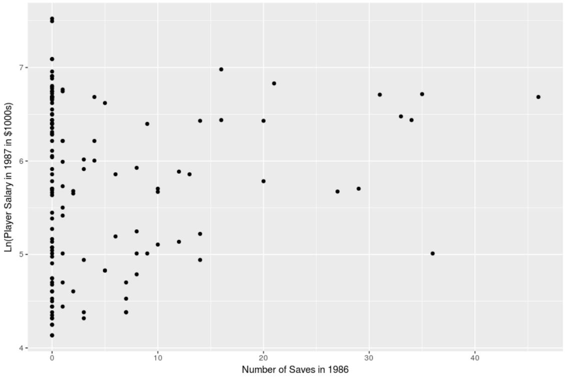


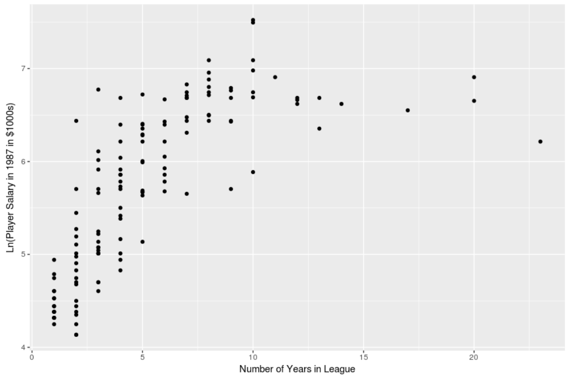


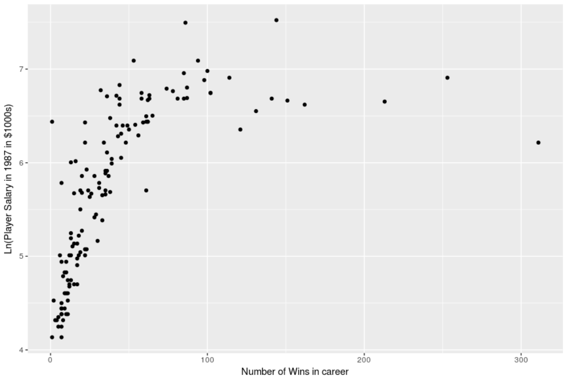


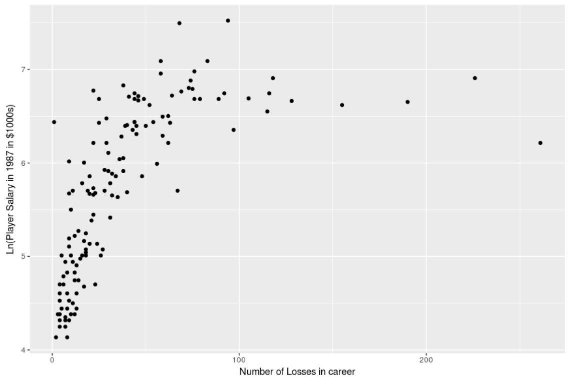


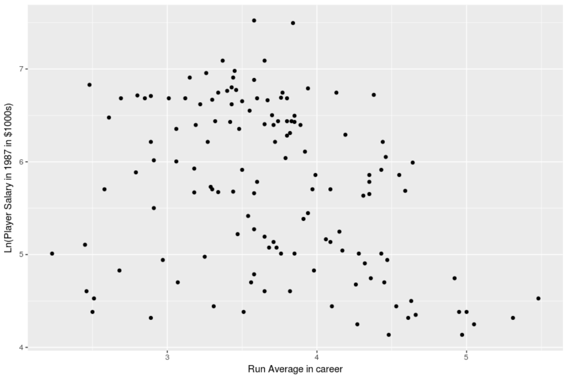


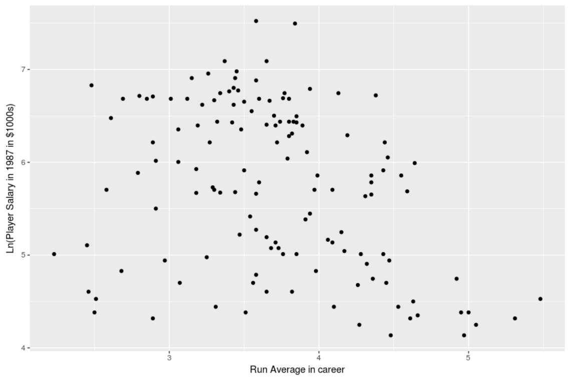


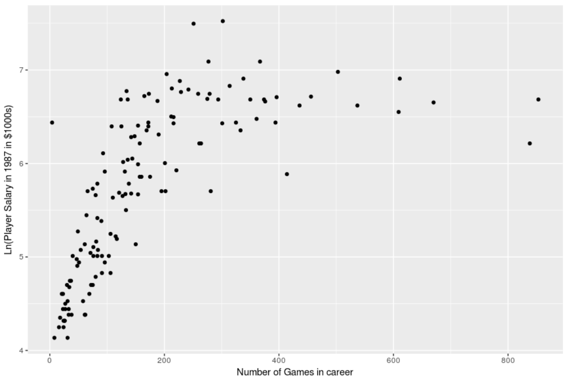

Ln(Player Salary in 1987 in $1000s)
4
5
6
7
0
5
10
15
20
Number of Years in League

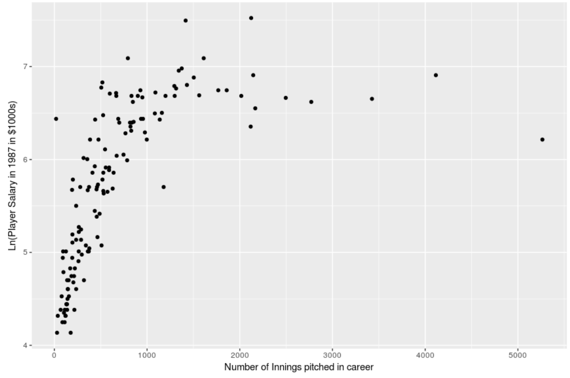

Ln(Player Salary in 1987 in $1000s)
4
5
6
7
0
100
200
300
Number of Wins in career

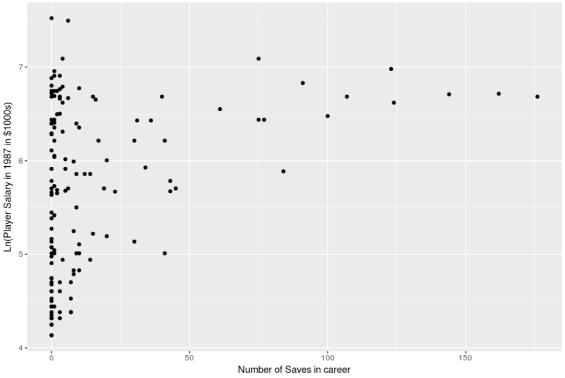

Ln(Player Salary in 1987 in $1000s)
4
5
6
7
0
100
200
Number of Losses in career


Ln(Player Salary in 1987 in $1000s)
4
5
6
7
3
4
5
Run Average in career

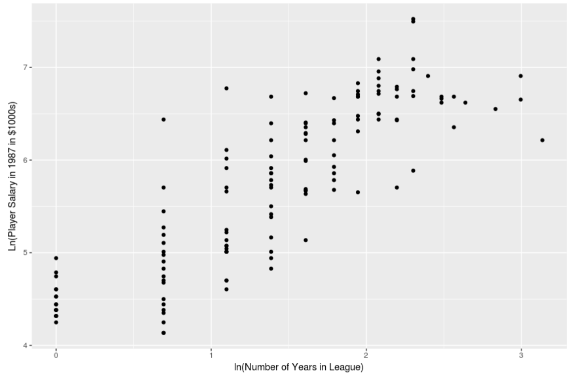

Ln(Player Salary in 1987 in $1000s)
4
5
6
7
3
4
5
Run Average in career

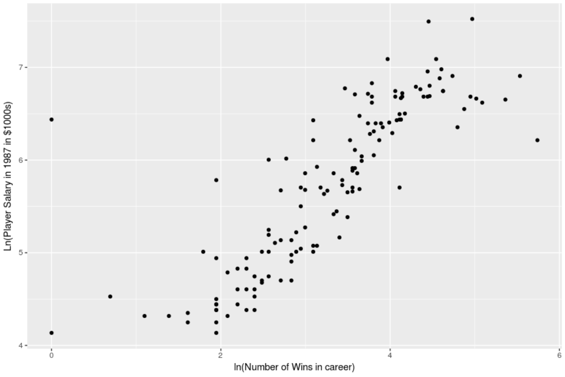

Ln(Player Salary in 1987 in $1000s)
4
5
6
7
0
200
400
600
800
Number of Games in career

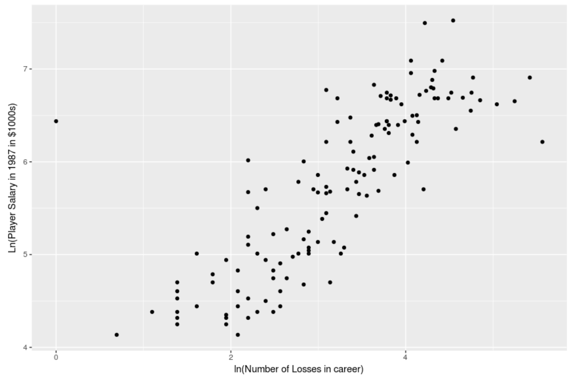


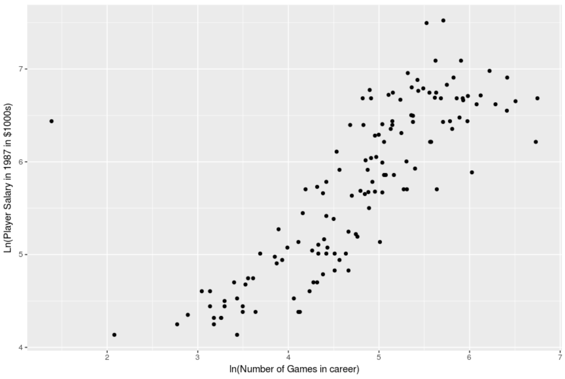


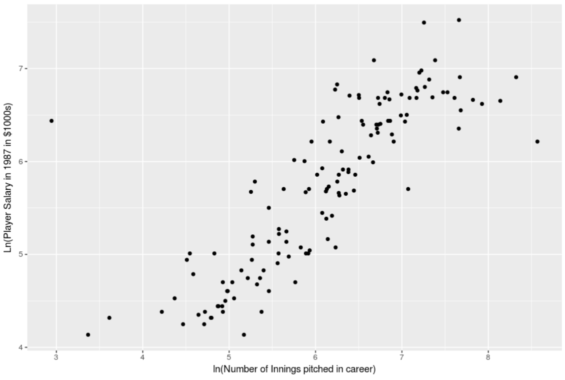
We observe that the figures 7,8,9,12 and 13 exhibit a sort of a linear to a flat transition plot. This is a significant trend. In fact it looks like a Logarithmic curve. We undertake specific Transformations of these variables by taking Natural Logarithms. The Revised Scatter Plots are Given Below.











These plots in fact give a more linear relationship with the Outcome Variable and hence we are using these going forward.

4.2 Linear Regression Model

4.2.1 Model Selection based on Adjusted R squared values

In this Method we choose that model which gives us the maximum Adjusted R squared value among all other possible models of consideration. We use Adjusted R square over R squared as it prevents over-parametrization and thus helps us to get a parsimonious model. This is achieved using the `ols_step_best_subset` function in the `olssr` package in **R** that computes the best model among all models based on certain criteria like Adjusted R squared. The `olssr` package can also be used to compute all possible Regression models using the given number of Regressors. After simulation in **R** , we obtain that the following model has the highest Adjusted R squared value. Do note that all of the regressors do have coefficients, but for the current scope of understanding we have not written them.

ln(Salary of Player in 1000$) = Earned Run Average in 1986 + ln(Number of Years in Major League) + ln(Number of Wins in Career) + Earned Run Average in Career

The Adjusted R squared value turned out to be 0.777.

4.2.2 Model Selection based on Mallow's Cp value

In this Method we choose that model which gives us the minimum Cp value among all other possible models of consideration. This is achieved using the `ols_step_best_subset` function in the `olssr` package in **R** that computes the best model among all models based on certain criteria like Adjusted R squared. The `olssr` package can also be used to compute all possible Regression models using the given number of Regressors. After simulation in **R** , we obtain that the following model has the lowest Cp value.

ln(Salary of Player in 1000$) = ln(Number of Years in Major League) + ln(Number of Wins in Career) + Earned Run Average in Career

The Cp value turned out to be -2.197.

4.2.3 Model Selection based on Akaike information criterion

In this Method we choose that model which gives us the minimum AIC value among all other possible models of consideration. This means that compared to all the other models it is the best. This is achieved using the `ols_step_best_subset` function in the `olssr` package in **R** that computes the best model among all models based on certain criteria like Adjusted R squared. The `olssr` package can also be used to compute all possible Regression models using the given number of Regressors. After simulation in **R** , we obtain that the following model has the lowest AIC value.

Ln(Salary of Player in 1000$) = ln(Number of Years in Major League) + ln(Number of Wins in Career) + Earned Run Average in Career

The AIC value turned out to be 154.7938.

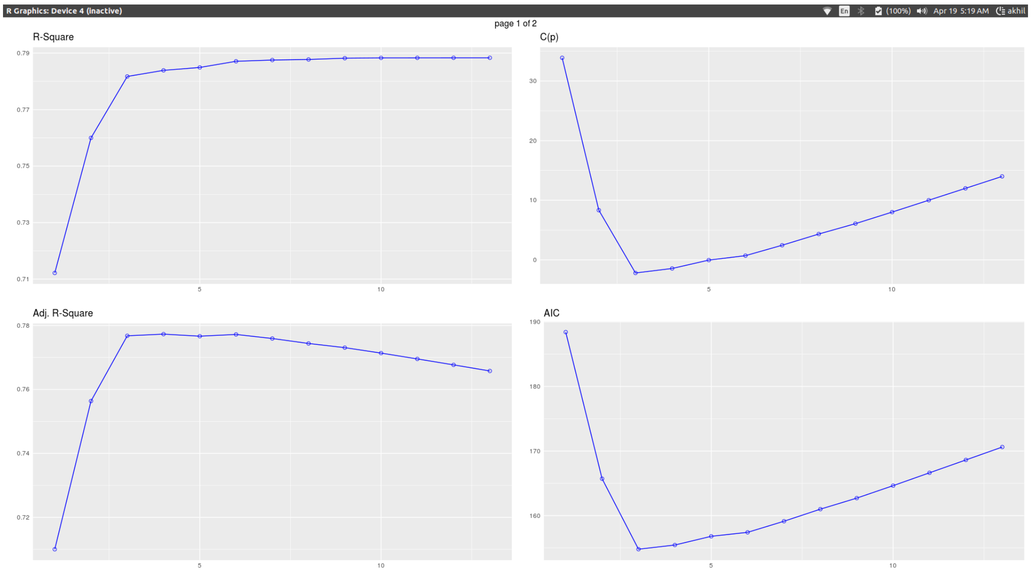


Figure 4.11: Selection of Best Model based on Different Criteria

4.2.4 Model Selection based on Stepwise Forward Regression

Since we are more or less starting with a limited idea on the structure of the model, our first case is to consider all the regressors and then judge on the basis of the statistics the relative importance among regressors to the final model. This is what is done by the function `ols_step_forward_` in the previously mentioned package.

ln(Salary of Player in 1000\$) = Earned Run Average in 1986 + ln(Number of Years in Major League) + ln(Number of Wins in Career) + Earned Run Average in Career

The Adjusted R squared value turned out to be 0.784.

Since all the methods agree that only 3 to 4 variables are required, we are for the sake of a slight improvement going for the four variable model.

4.2.5 Principal Component Analysis

In Principal Component Analysis we seek to undertake linear transformations of the data variables in such a way that the resulting variables are Orthonormal to each other. This transformation allows us to isolate those variables that are major contributors to the variability in the Outcome Variable based on their variance. Since they are all orthonormal this is simply the eigenvalues. In R, we use the $PCA()$ command as part of the package $'FactoMineR'$ to perform Principal Component Analysis. The variances of the input data re first normalized and then calculations are performed. The results are presented below:

Table 4.1: Principal Component Analysis - Eigenvalues

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	4.950651014	38.08193088	38.08193088
Dim.2	3.603630618	27.72023552	65.8021664
Dim.3	1.718940194	13.22261687	79.02478327
Dim.4	1.089032259	8.377171222	87.4019545
Dim.5	0.4936692827	3.797456021	91.19941052
Dim.6	0.4132235103	3.178642387	94.3780529
Dim.7	0.3205147784	2.465498295	96.8435512
Dim.8	0.2382944163	1.833033972	98.67658517
Dim.9	0.0872560717	0.6712005513	99.34778572
Dim.10	0.0562020805	0.4323236959	99.78010942
Dim.11	0.0159787641	0.1229135698	99.90302299
Dim.12	0.0101364668	0.0779728214	99.98099581
Dim.13	0.0024705449	0.0190041917	100

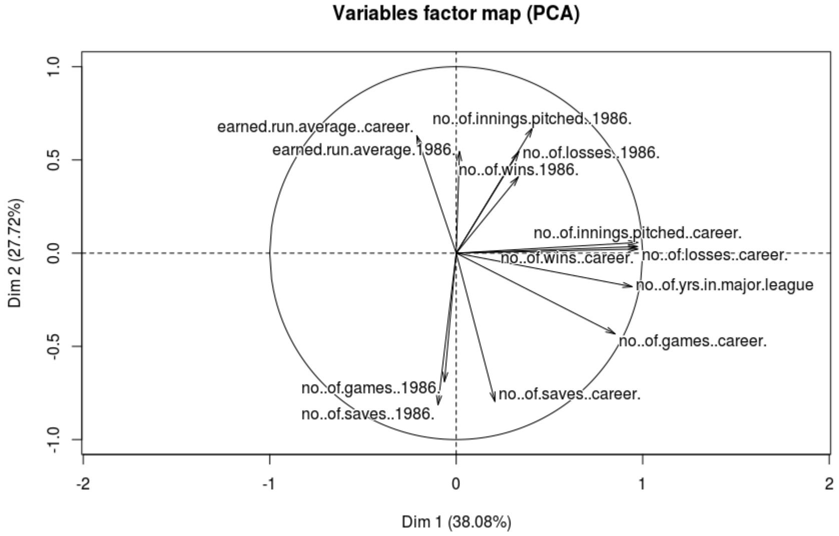


Figure 4.12: Parameters of Final Regression Model

From the table we observe that the first seven new transformed variables account for 96% of all variation. Also since the data has been normalized the eigenvalues all add up to 14 which is exactly the sum of the variance of the normalized data.

We have plotted this information below.

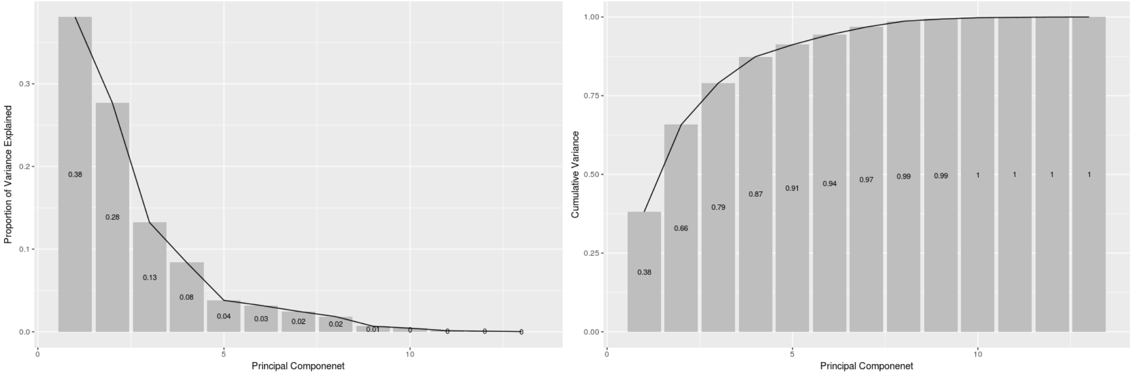


Figure 4.13: Scree and Cumulative Variance Plots

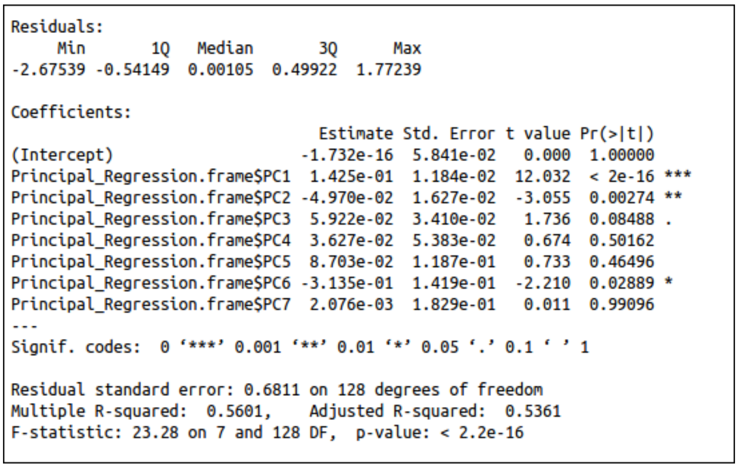
From Literature there are two ways of going forward,

1. We take those transformed variables in our model that have eigenvalues i.e variance greater than 1 as this means that these variables explain the variability far better than the original variables who had a variance equal to one (after normalization)

2. The other is based on how much of the variability do we want to explain in our model which leads us to how many variables to include in the model

We take the second route and are happy with 96% variability explained. This leads us to the use of the first seven variables.

Upon Performing Linear Regression using the first seven principal components we obtain the following result.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.67539 -0.54149  0.00105  0.49922  1.77239 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -1.732e-16  5.841e-02   0.000  1.00000    
Principal_Regression.frame$PC1  1.425e-01  1.184e-02  12.032  < 2e-16 ***
Principal_Regression.frame$PC2 -4.970e-02  1.627e-02  -3.055  0.00274 ** 
Principal_Regression.frame$PC3  5.922e-02  3.410e-02   1.736  0.08488 .  
Principal_Regression.frame$PC4  3.627e-02  5.383e-02   0.674  0.50162    
Principal_Regression.frame$PC5  8.703e-02  1.187e-01   0.733  0.46496    
Principal_Regression.frame$PC6 -3.135e-01  1.419e-01  -2.210  0.02889 *  
Principal_Regression.frame$PC7  2.076e-03  1.829e-01   0.011  0.99096    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6811 on 128 degrees of freedom
Multiple R-squared:  0.5601,	Adjusted R-squared:  0.5361 
F-statistic: 23.28 on 7 and 128 DF,  p-value: < 2.2e-16
```

Figure 4.14: Parameters of Principal Component Regression Model

As we had already obtained models with far superior Adjusted R squared value we will not be considering this model.

4.3 Residual Analysis

We perform Residual Analysis on our model to get an insight into whether the model violates any of the assumptions of Linear Regression and to identify outliers.

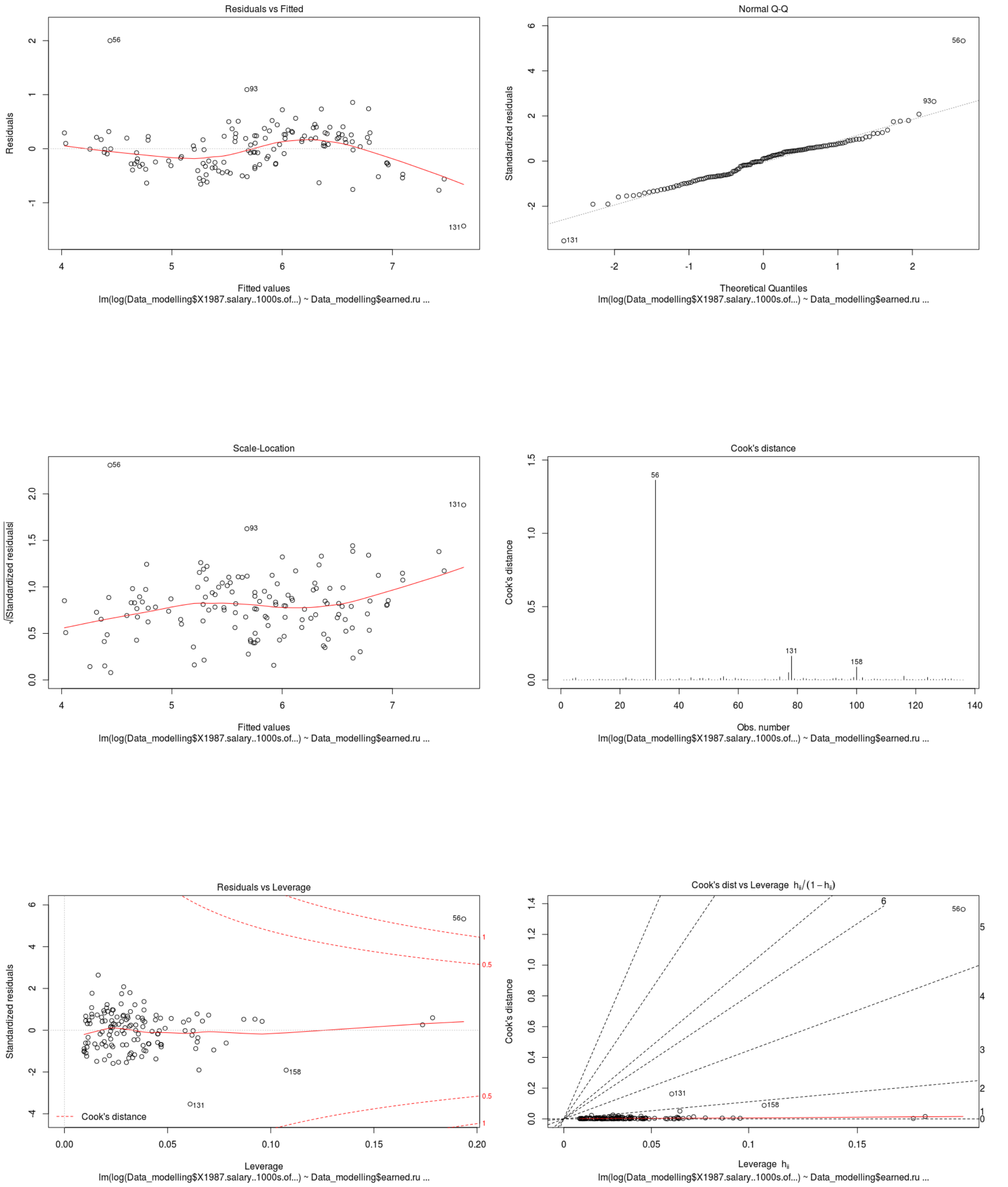
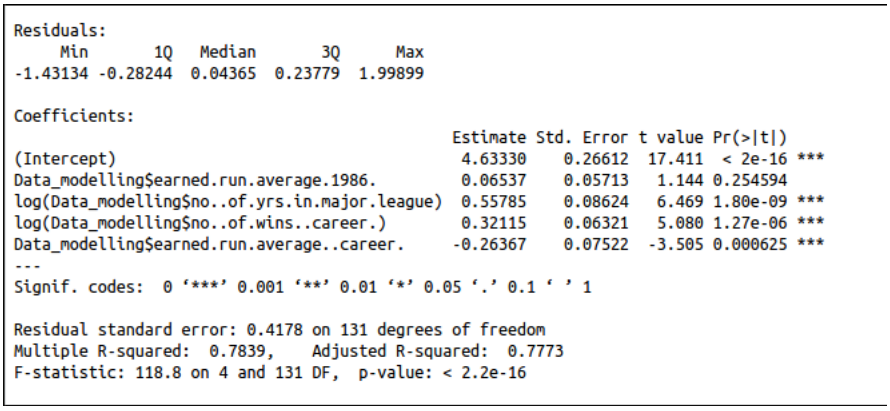


Figure 4.17: Residual Analysis Plots

As is clear from the first plot, the points 56, 93, 131 have the largest residuals. We also obtain a near band like structure indicating that we have not missed out any other significant regressor. Upon Inspection of the Q-Q plot, we get a near normal distribution which is what we wanted. Points 56,131 continue to be outliers while point 93 is comparatively close. The standardized residual plot reveals that points 56,93,131 continue to be outliers. The Cook's distance plot confirms that point 56 is very influential in determining the regression coefficients while the other two points are not significant as they are below 0.5 .

The model is in accordance with the assumptions of Linear Regression Methodology.

The final model is:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.43134 -0.28244  0.04365  0.23779  1.99899 

Coefficients:
                                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                     4.63330    0.26612  17.411  < 2e-16 ***
Data_modelling$earned.run.average.1986.         0.06537    0.05713   1.144 0.254594    
log(Data_modelling$no..of.yrs.in.major.league)  0.55785    0.08624   6.469 1.80e-09 ***
log(Data_modelling$no..of.wins..career.)        0.32115    0.06321   5.080 1.27e-06 ***
Data_modelling$earned.run.average..career.     -0.26367    0.07522  -3.505 0.000625 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4178 on 131 degrees of freedom
Multiple R-squared:  0.7839,	Adjusted R-squared:  0.7773 
F-statistic: 118.8 on 4 and 131 DF,  p-value: < 2.2e-16
```

Figure 4.18: Parameters of Final Regression Model

Chapter 5

Model Validation & Prediction

In this chapter, we will validate the model developed in the previous chapter using the Validation Set of Data that we had set apart in the initial stage.

The Model from Linear Regression is:

*ln(Salary of Player in 1000$) = 4.6333 + 0.06547*Earned Run Average in 1986 + 0.55785*ln(Number of Years in Major League) + 0.32115*ln(Number of Wins in Career) + (-0.26367)*Earned Run Average in Career*

We will now use this model to calculate the Predicted ln(Salary in 1987) and then compute the residual and residual squared values and then finally calculate R^2 and **R** and compare the values with the one obtained from Model Development. From Model,

R = 0.4178 on 131 degrees of freedom

Radj = 0.7773

From Validation,

R = 5.2 on 39 degrees of freedom

Radj = 30.93

There is a huge difference in the values as the Q-Q plot of log(Salary) of this data is far from normal and consists of many outliers as evidenced below.

Hence, while this model explained the modelling data quite well it failed when it came to the validation set. The model cannot be accepted.

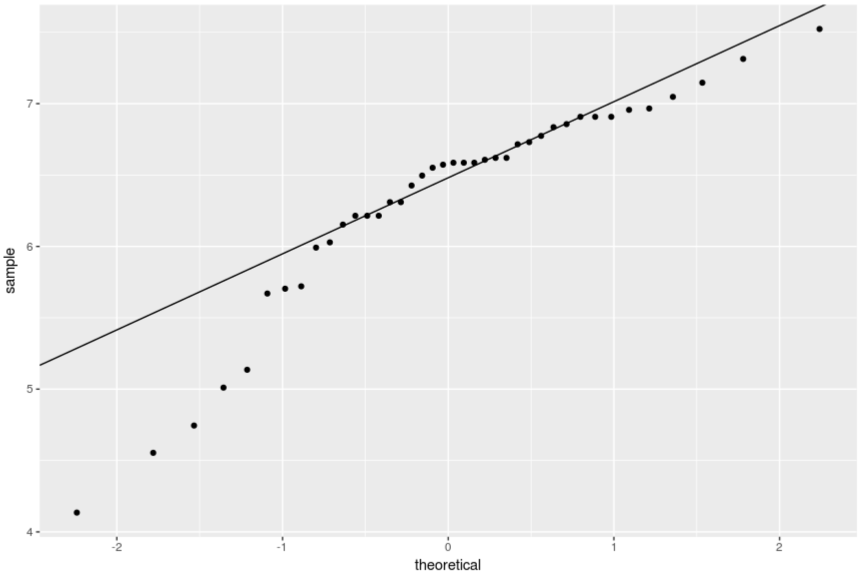


Figure 5.1: Q-Q plot of ln(Validation Data Salary)

References

[1] Hoaglin, D. C., and P. F. Velleman, âĂIJA critical look at some analyses of major league baseball salariesâĂİ, The American Statistician, Vol. 49, No. 3, pages 277- 285, August 1995

[2] Statistical tools for high-throughput data analysis - Website

[3] R-bloggers - Website

[4] ggplot2 Library Reference

[5] CL672 - R codes