**ITMD- 529 ADVANCED DATA ANALYICS**

**MIDTERM STATUS CHECK 3**

Contents

**1. Business Scenario**

**2. Business Objective**

**3. Data Overview**

**4. Explaining Variables**

**5. Exploratory Data Analysis – Results**

**5.1. QQ Plots**

**5.2 Histogram**

**5.3 Correlation**

**6. Regression Analysis – Results**

**6.1 Scatterplot**

**6.2 Calculation**

**7. Conclusion**

*Moneyball: The Art of Winning an Unfair Game* is a book by [Michael Lewis](https://en.wikipedia.org/wiki/Michael_Lewis), published in 2003, about the [Oakland Athletics](https://en.wikipedia.org/wiki/Oakland_Athletics) baseball team and its [general manager](https://en.wikipedia.org/wiki/General_manager_%28baseball%29) [Billy Beane](https://en.wikipedia.org/wiki/Billy_Beane). Its focus is the team's analytical, evidence-based, [sabermetric](https://en.wikipedia.org/wiki/Sabermetric) approach to assembling a competitive [baseball](https://en.wikipedia.org/wiki/Baseball) team, despite Oakland's disadvantaged revenue situation.

The book argues that the Oakland A's' front office took advantage of more analytical gauges of player performance to field a team that could better compete against richer competitors in [Major League Baseball](https://en.wikipedia.org/wiki/Major_League_Baseball) (MLB).

By re-evaluating the strategies that produce wins on the field, the [2002 Athletics](https://en.wikipedia.org/wiki/2002_Oakland_Athletics_season), with approximately US$44 million in salary, were competitive with larger market teams such as the [New York Yankees](https://en.wikipedia.org/wiki/New_York_Yankees), who spent over US$125 million in payroll that same season. Because of the team's smaller revenues, Oakland is forced to find players undervalued by the market, and their system for finding value in undervalued players has proven itself thus far. This approach brought the A's to the playoffs in 2002 and 2003.

What is this moneyball about ?

The moneyball is based on Michael Lewis's 2003 nonfiction book of the same name, an account of the Oakland Athletics baseball team's 2002 season and their general manager Billy Beane's attempts to assemble a competitive team.

1. ***Business Objective:***

The objective of our analysis is to find the baseball teams who will make it to the playoffs and by scoring how many more runs.

The dataset contains baseball matches played by the teams from the year 1962-2012. And the runs scored by the teams and whether they could make it to the playoffs. With the details provided in the dataset, we will predict how many games a team needs to win to make it to the playoffs. We'll then use linear regression to find how many more runs a team needs to score.

**Playoffs** - A series of games played to determine a championship.

1. ***Business Scenario:***

This will be addressed by building models which identifies the best teams to get into the playoffs. The study will be conducted based on the number of events they played and the matches they have won. We'll then use linear regression to predict the number of runs a team will score than their opponent to win that many games.We'll start by figuring out how many games a team needs to win to make the playoffs, by forming regression model and finding the values of the Run Difference which is needed to find the answer.

1. ***Data Overview:***

**Data sources**

|  |  |
| --- | --- |
| **No of rows** | 1233 |
| **Dependent variable (Y)** | wins |
| **Independent variables(X)** | Runs Scored, Runs Allowed, Run Difference |
| **Data source** | [Baseball-Reference.com](http://www.baseball-reference.com/) |
| **Time period of the dataset** | 1962-2002 |

The dataset will be called as the moneyball. Moneyball discusses how sports analytics changed baseball. Moneyball tells the story of the Oakland A's. The A's is a team near San Francisco, California.

They were once a rich team, but the team was purchased in 1995 by owners who enforced strict budget cuts. Despite this, they were improving over the years 1997 to 2001.That is why in his analysis we are focusing on data from the year 1962 – 2002 in order to predict the wins in the forecoming years.

***4.Explaining variables:***

|  |  |  |  |
| --- | --- | --- | --- |
| Team | Team | Factor |  |
| League | League | Factor | Categorical |
| Year | Year | integer | Continuous |
| RS | Runs Scored | integer | Continuous |
| RA | Runs Allowed | integer | Continuous |
| W | Wins | integer | Continuous |
| OBP | On Base percentage | numerical | Numerical |
| SLG | Slugging Percentage | numerical | Numerical |
| BA | Batting Average | numerical | Numerical |
| Playoffs | Playoffs | integer | Continuous |
| Rank season | Rank season | integer | Continuous |
| Rank Players | Rank Players | integer | Continuous |
| G | Ground | integer | Continuous |
| OOBP | Opponents on base percentage | num | Numerical |
| OSLG | Opponents Slugging percentage | num | Numerical |

Out of all these variables we are going to do regression analysis with Runs Allowed, Runs Scored, Run Difference. After the analysis is done, with the results we can see whether the above mentioned variables are going to impact the Wins, the team earns to make it to the playoffs.

***5. Exploratory Data Analysis and Results:***

This data set includes an observation for every team and year pair from 1962 to 2012 for all seasons with 162 games. We have 15 variables in our data set, including Runs Scored, RS, Runs Allowed, RA, and Wins, W.

We will name our new data frame moneyball, and we will use the subset function to take a subset of baseball and only take the observations for which Year is less than 2002. If we take a look at the structure of moneyball, we can see that we have 902 observations of 15 variables.

we want to build a linear regression equation to predict wins using the difference between runs scored and runs allowed. We will create a new variable called moneyball$RD, for run difference, and set it equal to moneyball$RS, runs scored, minus moneyball$RA, runs allowed

head(moneyball)

tail(moneyball)

names(moneyball)

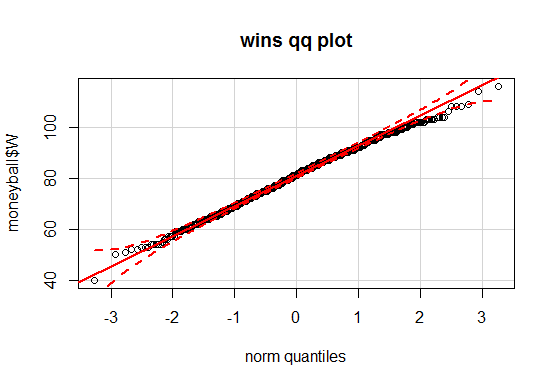
str(moneyball)

summary(moneyball)

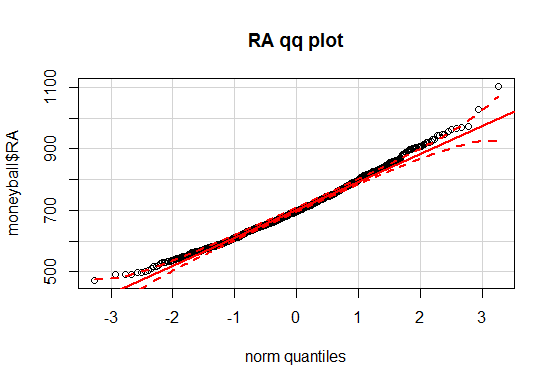
***5.1 QQ Plots:***

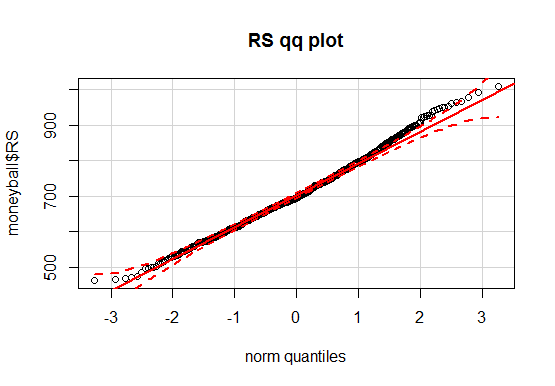
[Quantile-Quantile (Q-Q) plots](http://en.wikipedia.org/wiki/Qq_plot), are very useful plots for assessing how closely a data set fits a particular [distribution](http://en.wikipedia.org/wiki/Probability_distribution). If the points lie closely to the identity line, then the data has normal distribution.

**For Wins:**

****

**For RA, RS and RD**

****

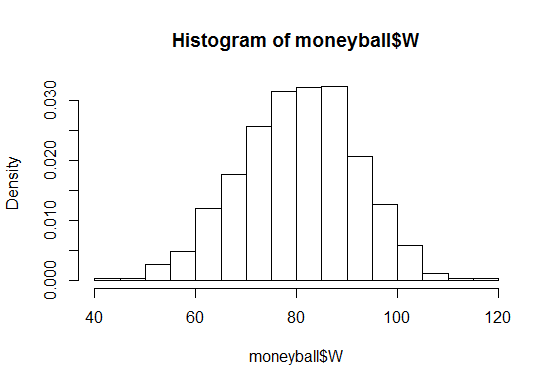
****

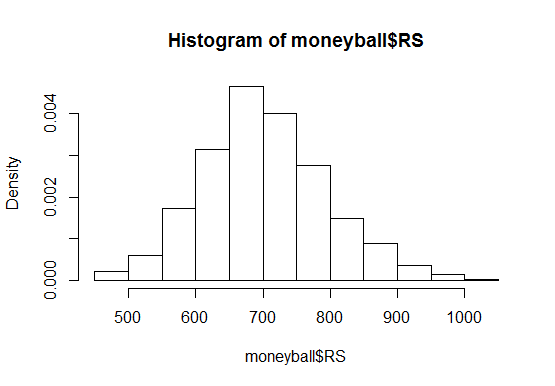
According to the close observation of the qq plots of all independent and dependent variables, almost all the variables have normal distribution. The dotted lines are lying close to the identity line which conforms that the variables have normal distribution.

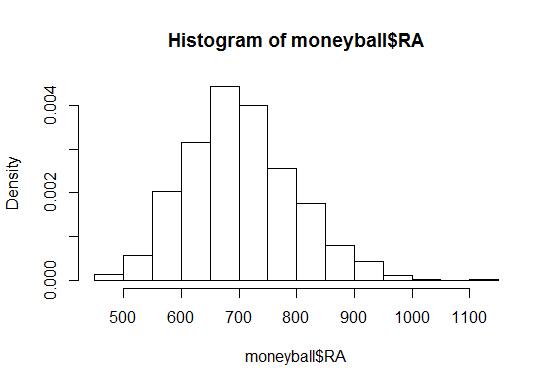
***5.2 Histogram:***

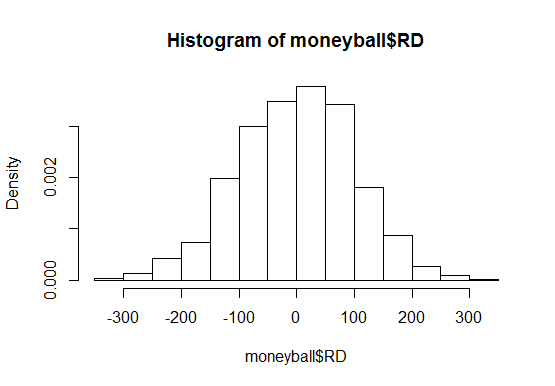
The histogram tells us the data distribution and which range the data lies. The histogram will show the maximum data value and the data range of the respective variable. From the below graph, it is clear that most of the Wins data lies from 75 to 95.

The maximum of RS data range is from 650 to 700. Likewise we can clearly see where most of the data lies in all these variables.This graph shows us that historically, if a team won 95 or more games, or is on the right side of this line, then they almost all made it to the playoffs. But if a team only won, say, 85 or more games, then they were likely to make it to the playoffs but it wasn't as certain. And if a team won, say, 100 or more games, than they definitely made it to the playoffs. So this plot shows us that historically, if a team won 95 or more games, then they had a strong chance of making it to the playoffs.

****

****

****

****

***5.3 Correlation:***

**cor(moneyball$RD, moneyball$W) -** 0.938515

A strong uphill (positive) linear relationship. With this correlation value, we can find how much significance an independent value will give to a dependent variable. The Run Difference variable will provide a strong relationship to the wins for each team.

**cor(moneyball$RS, moneyball$W) -** 0.5073824

A moderate uphill (positive) relationship. The Runs scored will have a moderate relationship with he variable wins.

**cor(moneyball$RA, moneyball$W -** -0.5077718

A moderate downhill (negative) relationship

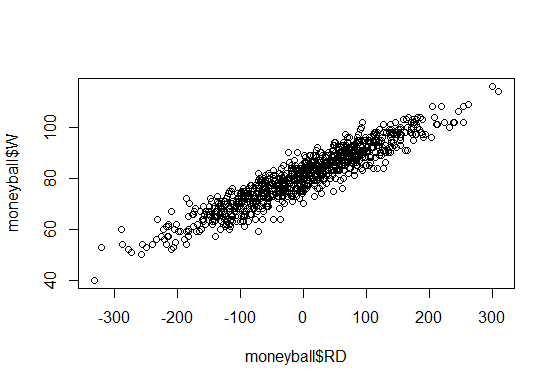
|  |  |  |
| --- | --- | --- |
| **RD** | **W** | 0.938515 |
| **RS** | **W** | 0.5073824 |
| **RA** | **W** | -0.5077718 |

***6.Regression Analysis Results:***

***6.1 Scatterplot :***

We have created a scatter plot between Run Difference and Wins. On the x-axis, we have put RD Run Difference and on the y-axis, we have put W

This scatter plot shows us that there's a very strong linear relationship between these two variables, which is a good sign for our linear regression equation. We can create our linear regression model, using the lm function to predict W, or Wins, using the independent variable RD using our dataset.

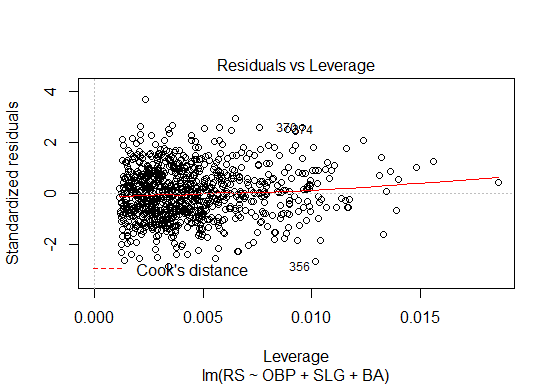


The below scatterplot actually gives us the answer we were looking for.

That is how many games does a team need to win in the regular season to make it to the playoffs?

The variable Run Difference has linear relationship with the Wins variable. From this we can say that the variable is significant enough to impact the Wins.





When we take a close look at the scatterplot, it is very clear from the graph that the teams who have made it to the playoffs have won atleast 95 games. This gives us a clue to find out how many more runs a team has to score in order to win. We will find out by forming the regression equation and finding the value for the run difference variable.

|  |  |  |
| --- | --- | --- |
|  | **Intercept** | **RD** |
| **Coefficient** | **80.881375** | **0.105766** |
| **Estimate Std. Error** | **0.131157** | **0.001297** |
| **t value** | **616.67** | **81.55** |
| **P value** | **<2e-16 \*\*\*** | **<2e-16 \*\*\*** |

|  |  |
| --- | --- |
| **Residual standard error** | **3.939** |
| **Multiple R-squared** | **0.8808** |
| **Adjusted R-squared** | **0.8807** |
| **F-statistic** | **6651** |

Residuals:

Min 1Q Median 3Q Max

-14.2662 -2.6509 0.1234 2.9364 11.6570

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 80.881375 0.131157 616.67 <2e-16 \*\*\*

RD 0.105766 0.001297 81.55 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.939 on 900 degrees of freedom

Multiple R-squared: 0.8808, Adjusted R-squared: 0.8807

F-statistic: 6651 on 1 and 900 DF, p-value: < 2.2e-16

[**F-test of overall significance**](http://blog.minitab.com/blog/adventures-in-statistics/what-is-the-f-test-of-overall-significance-in-regression-analysis) **determines whether this relationship is statistically significant.**

**F statistics - if the means between two variables are significantly different.**

**An *R*2 of 1 indicates that the regression line perfectly fits the data.**

In this regression result it shows that RD is very significant with three stars, and the **R-squared of our model is 0.88**. We have a strong model to predict wins using the difference between runs scored and runs allowed

***6.2 Calculation:***

Our Coefficients table tells us that our regression equation is Wins equals the intercept term, 80.8814, plus the coefficient for run difference, 0.1058, times RD, or Run Difference. We want Wins to be greater than or equal to 95 so that the team will make it to the playoffs. This will be true only if our regression equation is greater than or equal to 95. This is equal to 133.4. So this tells us that if the run difference of a team is greater than or equal to 133.4, as we predict that the team will win at least 95 games

W = 80.8814 + RD\*0.1058 >= 95

RD >= 133.4

Using this model we found that a team needs to win at least 95 games. By doing calculation with the values obtained from the regression analysis, we predicted the team needs to score at least 135 more runs than they allow to win the game.

***7. Conclusion:***

The basic question in the objective we had were

**“How many games does a team need to win in the regular season to make it to the playoffs?”**

**Ans : 95**

**“And how many more runs a team needs to score “**

**Ans : 134**

So by using linear regression and data, we were able to predict how many games a team has to win 95 games to get into the playoffs.

It is clear that our model was able to find us the answers for the questions in order to predict statistics to make it to the playoffs.