Multiple Regression Analysis of English Premier League Soccer Clubs: Number of Wins for the Season

Travis S Donoghue
Jonggoo Kang
Balakumaran Manoharan

Abstract:

Although many studies have been conducted to determine the outcomes of baseball and other sports such as football and sumo wrestling (Freakonomics, Stephen J. Dubner and Steven Levitt), there are few studies where the focus is exclusively on predicting how many games a team will win in a soccer season. The study you are about to read examines the influence of an individual team's statistics and compares it with other factors such as the number of passes, shots, goals, red cards, defense tackles, assists, etc. A multiple linear regression analysis using these factors are then performed resulting in significant positive results where the variables for the number of passes and shots being observed for all the models. Even further, the interaction of these two terms, shots and passes, contributed more to the regression analysis than any other factors. This implies that, the team that possesses the ball the most also generates significant scoring chances, ultimately influencing the outcome of the game. With possession and shots primarily taking place in the offensive jurisdiction of the game, the adage repeated time and time again for war and sporting events can be confirmed for soccer, that indeed "the best defense is a good offense."

Introduction:

In the recent boom of data mining and models, astounding outcomes of whole sports seasons have been predicted that no one could have ever expected. The practice of using data analytics in sports was made first famous in baseball with the Oakland Athletics (A's) and since then, data analytics has been used in every major-league baseball team since. If you do not know the story of the 2000-2003 Oakland A's, take a couple hours out of your busy schedule to watch the movie "Moneyball" which portrayed the story of the incredible feat. The feat could have only been accomplished using predictive analytics. So why soccer, why now? Out of all the major sports in the world, the sport that is by far the most popular with over 3.5 billion fans, but also utilizes the least number of predictive analytics techniques, is soccer. A 2010 New York Times article a referred to soccer as "the least statistical of all major sports". Luckily, a lot more attention has been made to soccer and the data associated with it. Just like any other sport, the data is collected and cataloged. Several companies have even sprouted with their sole purpose to collect statistics on soccer players (i.e. matchanalysis.com, whoscored.com, optasports.com, www.squawka.com, etc.) The data is available, in addition, records have been kept for decades only to be used for headlines and bar arguments…until now.

Data:

The data collected was acquired from Yahoo Sports which gathered data for the 2016 English Premier League throughout the first eight soccer games play for the season. Thirty-eight games will be played in all at which point the model will be repeated with the total data of the entire season. The soccer data we started with consists of a 23 columns and 400 rows, where each row represents a player. The columns include their statistics for the year, the team they play on, and what position they play. Some questions immediately came up when we began to

graph and analyze the data, a common occurrence during the process. Such as, should we add in the amount of times each team wins? Should we normalize the data? The answers to this were “Yes” and “Yes”. In response, the data was again cleaned in different variations and analyzed again. It added a significant amount of time to the process but only knowledge was gained. After normalizing the data, there still remained four separate data sets for each position: forward, midfield, defense, and goalie. The forward and midfield positions contained the same columns so it was decided to add a column to each dataset containing the players position. Once this column was created, we merged the two datasets. A step further was to repeat a similar routine and add the other positions to the dataset as well, making one large dataset. We were still unclear on if we should keep the datasets separate, so we decided to go down each path understanding the data could be filtered by position and team when needed. As mentioned, once the conjoined dataset was established, an obvious statistic was omitted which we quickly added; the amount of “Wins” each player’s team had. The “Wins” column would allow us to correlate the player’s individual statistics with the team’s performance generating a better response variable. Some other inquiries that rose to the surface included observing a large percentage of the data containing zero values. The observation was viewed prior to merging the datasets as not every player in the forward and midfield positions scored a goal, or generated an assist. However, after the data was merged, the amount of zero values drastically increased. The reason was obvious once a real thought was given, since a goalie will never have the opportunity to score a goal while a forward/midfielder/defender will also never have the opportunity to generate a “Save”. Therefore, when we generated graphs with stats for each individual player, we saw a drastic skew of information for zero values. In response to

the skew of zero values we again split the data and performed three different studies into our data. First, we graphed our data in separate buckets for each position. The goalies were analyzed with their individual statistics, then the defenders, the midfielders, and then the forwards. Secondly, we merged all the data sets and repeated the analysis. We then performed the first two steps again while removing all zero values in order to omit the skews seen in the graphs. After the extensive cleaning of the individual players data was done, it was ultimately decided that the data would be converted into team data. After it was all said and done, the final outcome of the data cleaned includes twenty rows where each team in the league represents each row. There are eighteen columns for each row that include the statistics of G (Goal), GA (Goal Assists), SHO (Shots), PAS (Pass), COR (Corners), FC (Foul Conceded), FS (Foul Surrendered), Y (Yellow Card), R (Red Card), Pen (Penalty), Mins (Minutes) gathered, again, from game 1 to game 8 as shown in Table 1.

Table 1: Summary of Model 2

Team	Goals	Assists	Shots	Passes	Corner Kicks	Fouls Conceded	Fouls Surrendered	Yellow Cards	Red Cards	Goalie Saves	Goals Conceded	Goal Kicks	Shut Outs	Tackles	Clears	Pentatlty Kicks	Wins
Arsenal	18	12	81	2710	44	74	77	7	1	23	9	61	3	58	130	2	6
Bournemouth	12	9	64	1705	44	82	100	4	1	27	12	72	2	47	169	1	3
Burnley	6	5	48	1383	26	76	67	3	0	39	12	62	2	52	202	1	2
Chelsea	15	11	102	2485	54	79	102	14	0	10	9	58	3	47	119	1	5
Crystal Palace	11	8	86	1672	54	90	97	13	0	20	9	63	0	52	154	0	3
Everton	11	8	85	1958	42	86	92	11	0	17	6	52	2	70	179	0	4
Hull City	8	5	58	2186	31	74	83	8	1	38	20	73	1	52	107	1	2
Leicester City	7	5	57	1892	37	93	59	3	0	30	14	37	3	46	128	1	2
Liverpool	18	12	91	3100	52	87	88	10	0	13	10	70	1	43	99	4	5
Manchester City	18	13	96	2510	53	73	104	8	1	19	8	47	1	75	141	3	6
Manchester United	13	8	98	2439	45	104	82	16	0	21	8	55	3	71	159	1	4
Middlesbrough	6	6	50	2131	33	101	77	12	0	16	12	57	1	80	146	0	1
South Hampton	9	4	107	2355	35	87	75	7	0	8	7	58	3	53	146	1	3
Stoke City	7	1	62	1586	44	86	112	11	0	30	16	67	1	81	125	1	1
Sunderland	6	2	47	1203	37	94	84	8	1	35	15	49	0	66	159	1	0
Swansea City	8	3	81	1782	31	93	79	4	0	28	15	103	1	72	165	1	1
Tottenham Hotspur	12	10	101	2178	65	108	98	6	0	22	4	71	4	54	176	0	5
Watford	13	9	61	1388	31	121	93	15	1	20	13	72	1	68	164	1	3
West Bromwich Albion	9	7	70	1231	41	107	67	9	0	27	8	39	2	54	122	1	2
West Ham United	9	7	79	1817	32	83	96	12	0	28	17	63	2	53	133	1	2

Underfitting & Overfitting:

One major fault in the data came when it was condensed from the large data set made up of individual players to the smaller data set of teams seen in Table 1. The fault is what is known as underfitting & overfitting. Underfitting is when the model does not fit the data well enough, and overfitting occurs when the model fits the data too well. We are more concerned of overfitting since the explanatory variables sizes are small. For example, for the Goal (G) statistic, the highest goal is 18 and lowest goal is 6. In other words, underfitting and overfitting occurs when the model is extremely simple model compared to the quantity of data contained.

Data Analysis:

Before beginning on the algorithm model to determine the amount of wins each team will conclude, the data needs to be cleaned a bit further. Luckily, the data does not require any normalizing at this point and is already somewhat clean from the preparations before. Some minor improvements were conducted in the open source program R removing three columns. The columns removed were column 1, which represented names of each team. Column 14 represented losses and column 15 represented draws. The columns were removed because they are irrelevant to completing the model. Draws and losses are not needed as the explanatory variable will be Wins. The team names are categorical variables any not needed as the statistics of each team are only useful in completing the model. The code for cleaning the data in the R program is listed below for reference.

```
> dat <- Offense.Project.Data[-c(1,14,15)]
```

As the data currently stands, it is cleaned in rows and columns useful to begin the data analysis. The data analysis portion of the process is equally as important as gathering and cleaning the

data. This is the portion of the process when you really start to gain an understanding of your data. There are plenty of ways to begin. Some of the most uses tools are building histograms, five number summaries, box plots, bar charts, scatter plots, and correlation matrices. The goal in the process is observe trends, discover outliers, and simply put, get to know your data. Prior to beginning with modelling, gaining a better understanding of the data is key. It would be disappointing to build, appears to be, a successful model only to conclude the model contains high correlations of variables that has caused multicollinarity. Therefore, a test for multicollinarity will be conducted first. Multicollinarity is an unfavorable correlation between two independent explanatory variables. One of the more favorable analytical tools used to understand the relationship between different variables is the use of a correlation matrix. A correlation matrix takes each variable in each column and provides a value to the relationship of each variable against each of the other variables. The value ranges between -1 and +1 where +1 demonstrations the variables are perfectly positively correlated. An example of positively correlated variables would be sales dollars and sales tax; when dollars go up, so does the sales tax. A value of -1 means perfect negative correlation, or in other words, when one goes up the other goes down. An example of negatively correlated variables would be items sold and inventory. Every time you sell an item you remove it from the inventory. A value close to 0 means either no relation or the relation isn't linear. The correlation matrix is an excellent tool in the beginning stages of analyzing data. The correlation matrix tells a lot of information in a short amount of time. It is also used to determine multicollinearity. Multicollinearity can create opposing signs (+/-) and can create rounding errors. During this exercise, a correlation matrix was created. Most correlation matrices are shown with numbers ranging from -1 to +1,

however, in order to better appeal to the way our minds function, a correlation matrix with color and shape was created. The scale is shown to the right of the matrix classifying as the correlation increases to +1 the color becomes a darker shade of blue while the circle shape becomes larger. As the correlation decreases to -1 the color becomes a darker shade of red while the circle shape becomes smaller. As mentioned before, the data represents the overall data of each individual player summed up to the overall team level. For example, if two individual goalies on the same team each had one save each, this would equate to two saves for the team. The correlation matrix provides a great representation of exactly which stats correlate at the team level. Some of the values are obvious as the amount of saves a goalie makes shouldn't have any correlation to the amount of shots the team takes. The stats occur on the complete opposite sides of the field. What is very interesting is to realize how correlated some of the values are. We didn't suspect the amount of Corner Kicks (COR) related to the number of Fouls Surrendered (FS), but it appears to. It is also rewarding to visually see the hypothesis ring true as the number of Saves (SAV) a goalie makes does actually correlate to the amount of Goals Conceded (GC). This makes sense because the chances of being scored on increase with the increasing saves the goalie must make. Eventually the goalie will not be able to make the save, and thus, the goal will be conceded. See Table 2 for the correlation table.

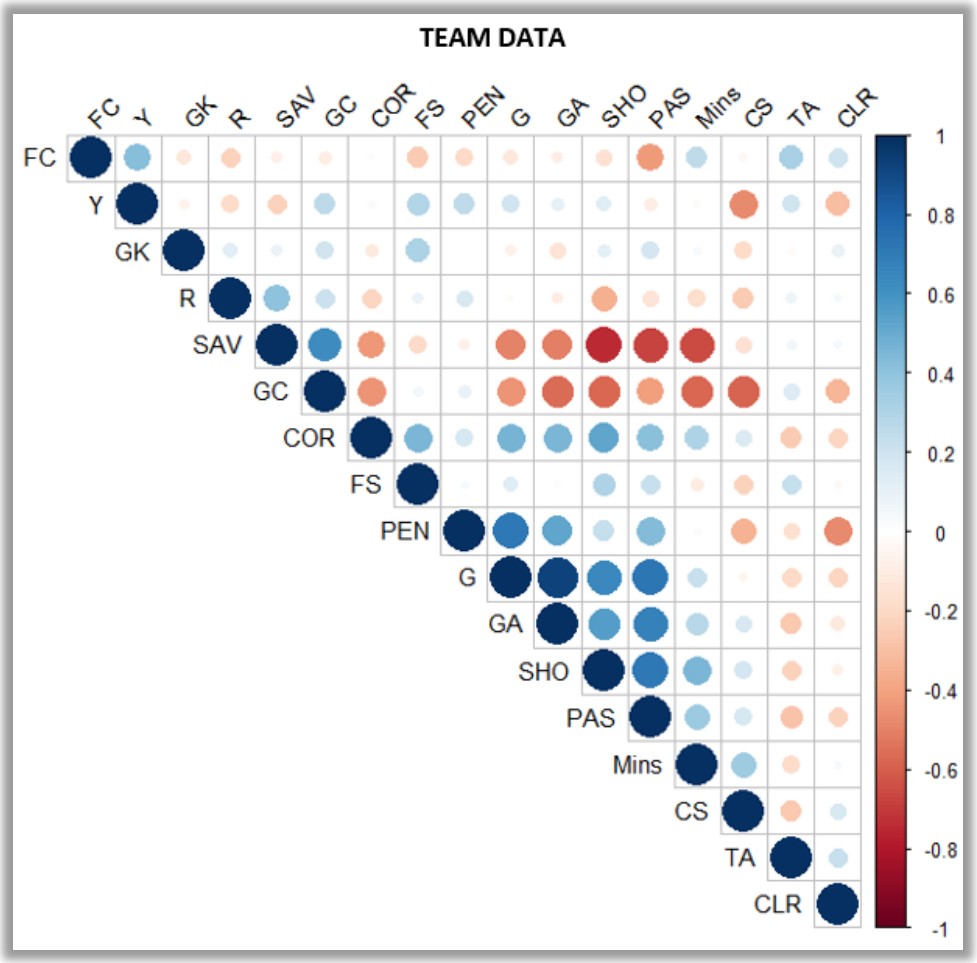


Table 2: *Visual Correlation Table*

Another useful tool when analyzing the data is to use box plots which are derived from the five number summaries. The box plots provide the visualization for the five number summaries, and are useful when determining the min, max, mean, and any outliers associated with the data. This was useful for understanding the range and outliers of the data. The boxplot will provide the true five number summary for the final cleaned data product. Box plots were created for each explanatory variable. An example of the stat for “Passes” is shown in Figure 1.

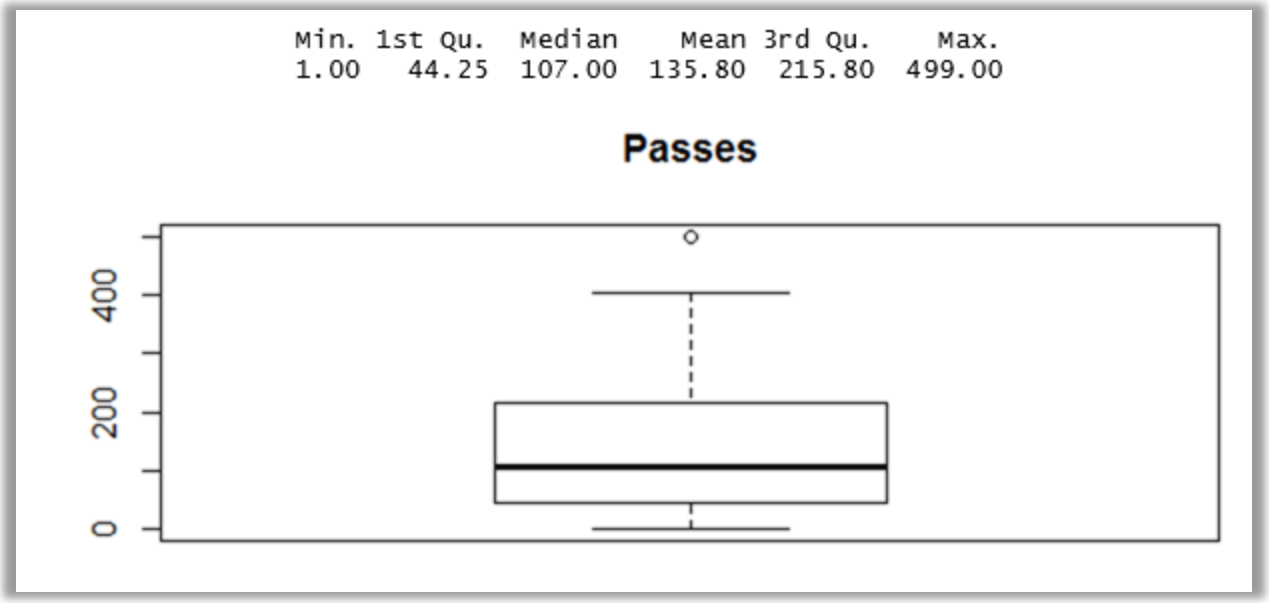


Figure 1: *Box Plot for the Variable Passes*

Other ways to analyze the data and gather further insight in the data is to utilize histograms and bar graphs. Although histograms and bar graphs are not shown here, they were utilized for each variable. The histograms and bar graphs are not shown as the value was minimal. Histograms were used for the residual errors of the model and will be touched upon later in the paper.

There are some instances where a bar graph can provide a quick visual of a couple parameters, but the ideal tool to use for compare and contrast are scatter plots and correlation matrices. As correlation matrices were touched upon, scatter plots will not be addressed. Using knowledge from the sport of soccer, it is understood that some stats of the players and teams correlate. An example of this would be the amount of saves a goalie makes vs the amount of goals conceded. The goalie being the last line of defense on the team. It can be understood that the more saves a goalie makes in the game, the more chances the opposing team has to score, which then likely translates to the more goals conceded by the goalie. This hypothesis, when

graphed into a scatter plot, can be visualized. See figure 2, for a scatter plot displaying the amount of "Saves" on the x-axis verse the amount of "Goals Conceded" on the y-axis. The trend is clear which holds our hypothesis true. The more saves a goalie makes, the more likely that a goal will be conceded. Other scatter plots that showed trends were "Goals" vs "Passes" and "Assists" vs "Passes". The largest trend was shown to be "Assists" vs "Goals", seen in Figure 3, which makes complete sense. In order to score a goal, someone needs to pass you the ball.

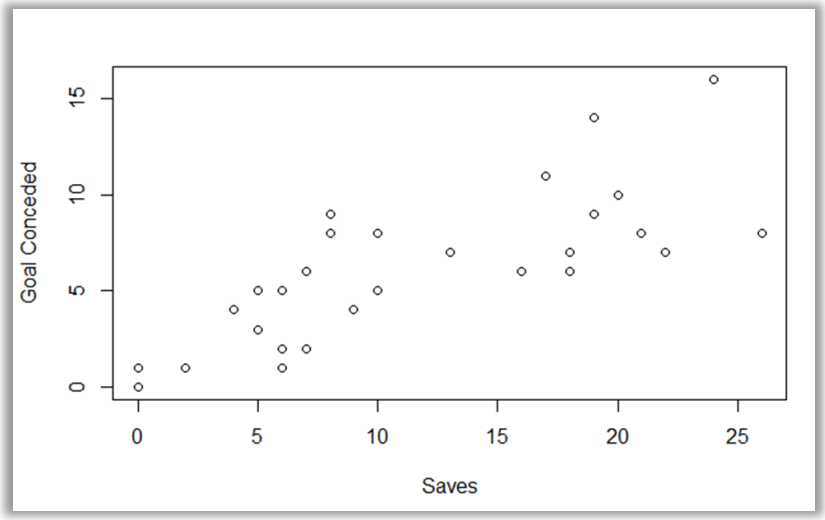


Figure 2: Scatter Plot of Saves vs Goals Conceded

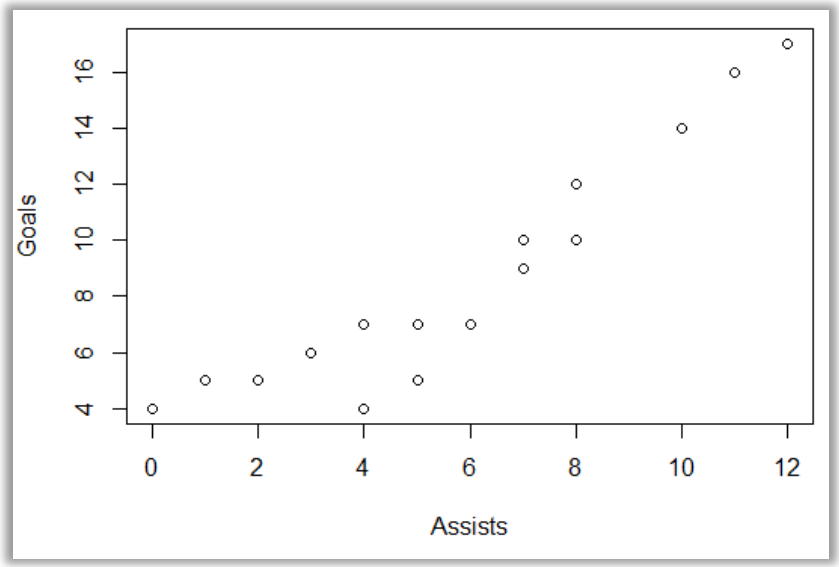


Figure 3: Scatter Plot of Goals vs Assists

The knowledge of correlated variables were used to better understand the data but also which variables would be ideal when completing the model. Two final variables are seen graphed linearly in Figure 4 and Figure 5.

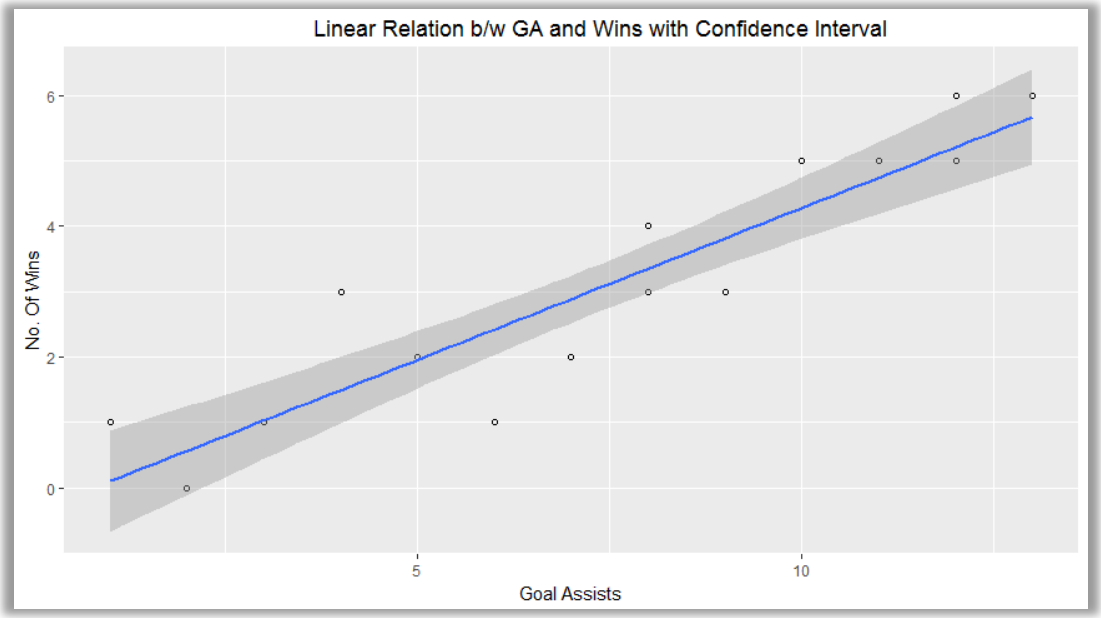


Figure 4: Linear Relation of Goal Assists vs No. of Wins

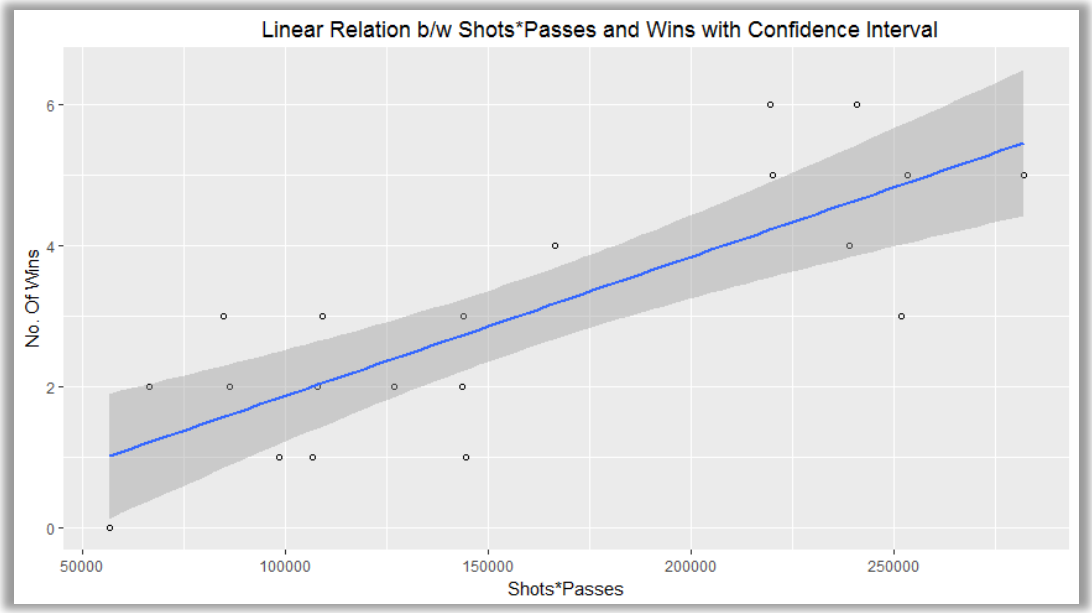


Figure 5: Linear Relation of Shots*Passes vs No. Of Wins

Methods:

Response Variable

The response variable in the model will be Wins. As identified earlier, the goal of the outcome is to predict how many times each team will win. Therefore, the column for the number of Wins each team has will be used as the response variable. The other variables are going to be explanatory variables.

Explanatory Variable

Results:

After exhausting all other plausible regression analysis techniques, a simple elegant model was chosen from a total of three models that were ultimately considered in the end. The first model considered in the end was Model 1 seen below. With an adjusted R-Square of .8072, F-Test of 1.488e^-06, a promising model was born. The term SPC represented an interaction term of (Shots*Passes*Corners) along with the term for red cards (R) T-Tests were both

significantly below 0. The third term representing Goals Conceded (GC) was just below the .05 significance level we were striding for. Two immediate concerns arose before removing Model 1 from the running. First, the beta for red cards (R) was positive. Understanding the domain, when a red card is given to a player on a team, that player is ejected from the game and the team is then forced to play with one less player for the remainder of the game. The red card beta would be perceived to be negative, however only minimally positive, it was indeed still positive. Further data was researched on the circumstances that occur when a red card is given begging to find more clarity in questions such as: 1) How many minutes are left in a game when a red card is given? 2) Are positive adjustments made to the team? 3) Do team member's player harder in replace of the player that was ejected? The questions among others were unable to be answered providing the main reason for removing Model 1. The second reason for removing Model 1 is because of the complexity of the term SPC (Shots*Passes*Corners). In practice, a model should be simplified rather than over complicated with interactions of a term finally begging the removal of Model 1.

```
lm(formula = Wins ~ SPC + R + GC, data = AllData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94794 -0.33973 -0.04164  0.57378  1.13283

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.057e+00  9.368e-01   2.196  0.04318 *
SPC          2.867e-07  5.374e-08   5.335  6.7e-05 ***
R            1.166e+00  3.886e-01   2.999  0.00849 **
GC          -1.250e-01  5.797e-02  -2.156  0.04665 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7671 on 16 degrees of freedom
Multiple R-squared:  0.8377,    Adjusted R-squared:  0.8072
F-statistic: 27.52 on 3 and 16 DF,  p-value: 1.488e-06
```

Figure 6: Summary of Model 1

Model 2 showed an 8.07-point improvement in the adjusted R-square along with improvements in the F-Test as well as the overall T-tests. The explanatory terms chosen were Goals (G), interaction terms of Shots*Passes (SHOP), and penalty kicks (PEN). Before the testing of Model 2 began, Model 2 was scrapped for one main reason. With the penalty spot 36ft from the goal, the average ball kicked in a penalty traveling at 70mph, with only the goal keeper able to stop the ball within goal posts that span 24 ft wide and 8 ft tall, a penalty kicks (PEN) results with an 85% chance of scoring a goal. The simple understanding of more goals scored equating to a higher likelihood of winning the game, the beta for penalty kicks (PEN) should be positive, however, in Model 2 the beta is negative. For this reason, the model was not used.

```
Call:
lm(formula = Wins ~ G + SHOP + PEN, data = All.Team.Project.Data1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9099 -0.3391 -0.2031  0.3102  1.3377

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.604e+00  3.956e-01  -4.054 0.000921 ***
G            3.739e-01  5.696e-02   6.564 6.52e-06 ***
SHOP         6.412e-06  2.827e-06   2.268 0.037521 *
PEN         -4.029e-01  1.778e-01  -2.266 0.037648 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.585 on 16 degrees of freedom
Multiple R-squared:  0.9056,    Adjusted R-squared:  0.8879
F-statistic: 51.16 on 3 and 16 DF,  p-value: 2.017e-08
```

Figure 7: Summary of Model 2

Model 3 was the final model which utilized the least amount of terms of the three but also showing the significant improvements over the other two. Uses an interaction term learned from Model 1, Shots*Passes (SHOP) along with Assists (GA), all traits of Model 3 were an improvement over the other two models. The model terms of Shots and Passes were also included in the model because of the common knowledge of regression model building that "you keep the children" of interactive active terms. The model also distinctively made sense when understanding passes generate shots, that in turn generates assists when a goal is scored. Again, the more goals scored, the more likely the team is going to win the game. The only hesitations that were received with the model was multicollinearity which will later be put to rest, as well as the absence of defensive and goalie statistical terms. Initially, the final model to be used was hypothesized to contain positive betas from the offense as well as negative betas from the defense and goalies. Model 3, in the end, relied solely on the offensive statistics to

predict the response variable of winning the game. The summary of Model 3 can be seen in Figure 8.

```
Call:
lm(formula = Wins ~ GA + SHOP, data = AllData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08645 -0.23368  0.00746  0.30182  0.79282

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.724e-01  3.285e-01  -2.960 0.008776 **
GA           3.384e-01  4.743e-02   7.136 1.67e-06 ***
SHOP         9.651e-06  2.265e-06   4.262 0.000527 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5505 on 17 degrees of freedom
Multiple R-squared:  0.9112,    Adjusted R-squared:  0.9007
F-statistic: 87.19 on 2 and 17 DF,  p-value: 1.155e-09
```

Figure 8: Summary of Model 3

Model Validation:

Once the confidence in the Model 3 seen in Figure 8 was accomplished, validating the model was the next step in the process. Two validation tools were used when validating the model; N-Fold Cross Validation and Jack Knifing.

N-Fold Cross Validation:

Pondering if R-Squared could mislead the interpretation of the quality of regression model 3, N-Fold Cross Validation was conducted. In layman's terms, N-Fold Cross Validation is when 80% of the data is used for training the model and 20% of the data is used for testing the model. The tests are repeated depending on the amount of N-Folds that will be conducted. The process is completed until 100% of the data has been tested through each fold. The open

source program, R, was also used to complete the N-Fold Cross Validation technique. As can be seen from the code in Figure 9, the mean of squared error (ms) is 0.332, a number satisfactory to conclude the model successful. The mean of squared error (ms) is a measure of the prediction error inherent to the model.

```
fold 1
Observations in test set: 2
                8    16
Predicted   1.761  1.44
cvpred      1.783  1.50
Wins        2.000  1.00
CV residual 0.217 -0.50

Sum of squares = 0.3    Mean square = 0.15    n = 2

fold 2
Observations in test set: 3
                9   17     20
Predicted    5.81 4.53  2.782
cvpred       6.02 4.68  2.847
Wins         5.00 5.00  2.000
CV residual -1.02 0.32 -0.847

Sum of squares = 1.86    Mean square = 0.62    n = 3

fold 3
Observations in test set: 3
                  1     11     19
Predicted   5.207 4.0418  2.228
cvpred      5.073 3.9685  2.216
Wins        6.000 4.0000  2.000
CV residual 0.927 0.0315 -0.216

Sum of squares = 0.91    Mean square = 0.3    n = 3

fold 4
Observations in test set: 3
                   5    10    18
Predicted    3.1227 5.753 2.891
cvpred       3.0921 5.681 2.838
Wins         3.0000 6.000 3.000
CV residual -0.0921 0.319 0.162

Sum of squares = 0.14    Mean square = 0.05    n = 3

fold 5
Observations in test set: 3
                6    12    13
Predicted   3.341  2.09 2.813
cvpred      3.346  2.15 2.616
Wins        4.000  1.00 3.000
CV residual 0.654 -1.15 0.384

Sum of squares = 1.9    Mean square = 0.63    n = 3

fold 6
Observations in test set: 3
                2     7     14
Predicted    3.127 1.943 0.3150
cvpred       3.171 1.831 0.0721
Wins         3.000 2.000 1.0000
CV residual -0.171 0.169 0.9279

Sum of squares = 0.92    Mean square = 0.31    n = 3

fold 7
Observations in test set: 3
                3     4     15
Predicted   1.360  5.196  0.250
cvpred      1.274  5.251  0.166
Wins        2.000  5.000  0.000
CV residual 0.726 -0.251 -0.166

Sum of squares = 0.62    Mean square = 0.21    n = 3

Overall (Sum over all 3 folds)
   ms
0.332
```

Figure 9: R code used for N-Fold Cross Validation

Jackknifing:

Jackknifing, like N-Fold Cross Validation, is a tool used to validate the model. However, jackknifing is much different than N-Fold Cross Validation. The jackknife estimator of a parameter is found by leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations. Jackknifing can also be called "Leave-One-Out" because of the process of leaving out each observation of the dataset. Given a sample of size N, the jackknife estimate is found by aggregating the estimates of each N-1 estimate in the sample. A code in Figure 6, created in R generates the model to predict Wins based on Goal Assists(GA), Shots (SHO), Passes (PAS) and an interaction term SHOP (Shots * Passes). The result of each test is then stored in a list and used to retrieve the Adjusted R-Squared as described earlier.

```
JKFMds <- list()
for (i in 1:length(JKGA)){
  iWins = c(JKWins[[i]])
  iGA = c(JKGA[[i]])
  iSHOP = c(JKSHOP[[i]])

  JKFMd <- lm(iWins ~ iGA + iSHOP)
  print(i)
  print(summary(JKFMd))
  JKFMds[[i]] = summary(JKFMd)
}
```

```
for (i in 1:length(JKFMds)){
  print(JKFMds[[i]]$adj.r.squared)
}

JKFMdlsAdjRSq = c(0.8977617,0.9004378,0.9075947,0.8932024,0.900384,0.9072754,0.898294,0.8994571,0.9098845
                  ,0.8823413,0.8982656,0.9190063,0.9012315,0.9051006,0.8823615,0.896995,0.8972597,0.9003697
                  ,0.899406,0.9109677)
```

Figure 10: *R code used for Jackknifing*

Residuals:

Plotting the residuals in a scatter plot is a final method used during the building of the model process. When plotting the residuals in a scatter plot vs the explanatory variable, the goal is to try to observe a trend. If a trend is observed, the residuals is said to be heteroscedastic, where the model should be homoscedastic. Types of heteroscedastic trends are binomial, multiplicative, and Poisson. From looking at the residual plot in Figure 11, there are not any visible trends. From the absence of trends, the model is said to be complete and ready for analysis. If a trend did exist, an explanatory variable used would need to transformed depending on the type of heteroscedasticity.

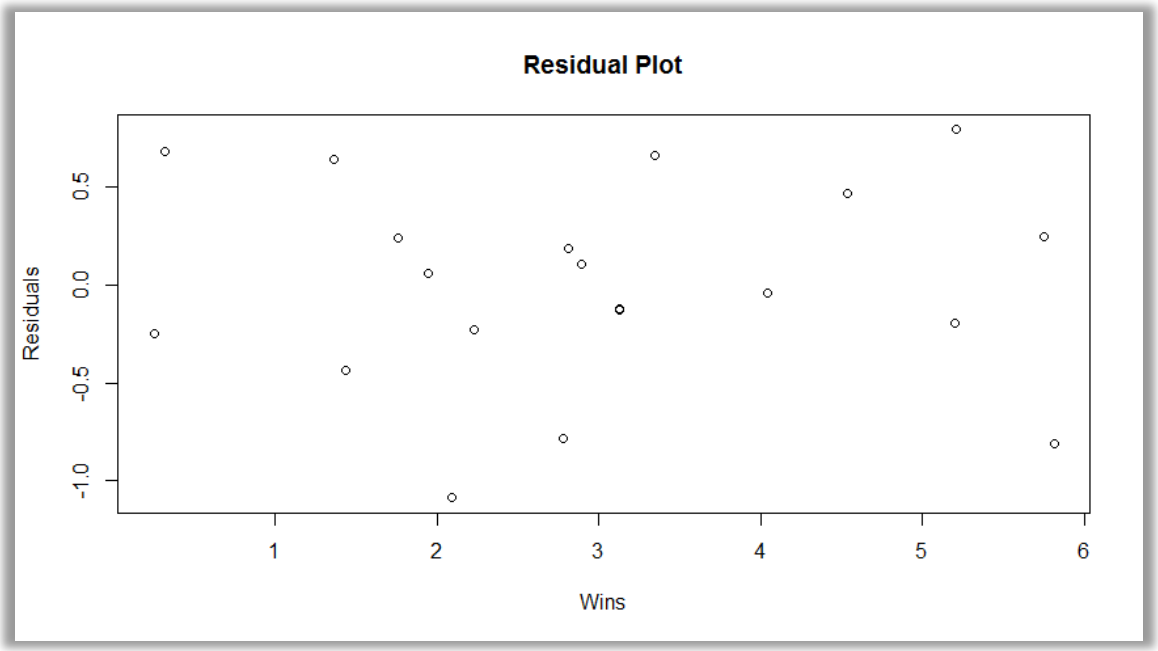
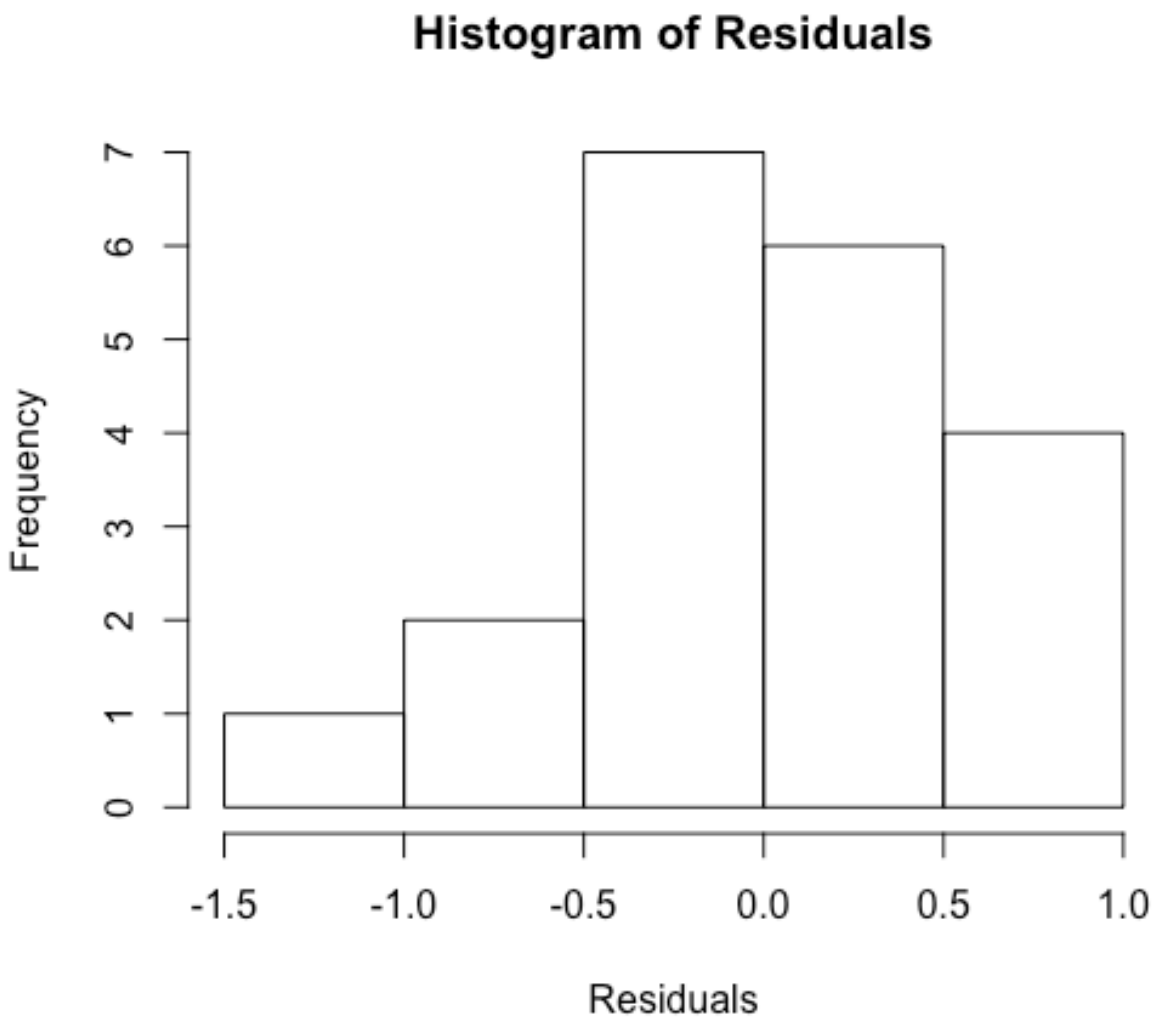


Figure 11: Residual Plot


Histogram of Residuals
Frequency
0
1
2
3
4
5
6
7
-1.5
-1.0
-0.5
0.0
0.5
1.0
Residuals

```
> cv.lm(dat1,model8,m=7)
Analysis of Variance Table

Response: Wins
          Df Sum Sq Mean Sq F value  Pr(>F)
GA         1   47.3    47.3   156.2 5.4e-10 ***
SHOP       1    5.5     5.5    18.2 0.00053 ***
Residuals 17    5.2     0.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: ANOVA of Model 3

Conclusion:

The success of a soccer team is wildly dependent on the number of passes, shots, and assists the team has for each game. Although it is known the defense also contributes, the total number of Wins a team has is reliant on the three variables described throughout the paper: passes, shots, and assists. Most of these variables are directly contributed to the offensive portion of the team; midfielders and offensive, with minor instances coming from the defense. In more detail, the final model is: -0.09724 + 0.03384*GA + 9.651e-6 SHO*PAS with Adjusted R-squared: 0.9007 and p-value: 1.155e-09. This model has been chosen over two other models based on F-Test, T-Test, better Adjusted-R-Squared value, stepAIC and a "no pattern residual plot". We used N-Fold Cross Validation and Jack Knifing to evaluate the model. It turns out that it is accurate enough to 90% of the data in predicting the wins. To add further, the Mean Square Error (ms) from N-Fold Test is less significant with 0.332. Results from Jack Knifing seems to be satisfying with 0.008 variance over the mean value of 90 (Adjusted R-Squared). The model is parsimonious, the validation of the model has been completed, the only thing left to conclude is next time you decide to watch a game of soccer, realize that the outcome of the game is contributed to mostly the offensive side of the ball. The team with the most passes and shots will get the most assists concluding a victory for the team!