EIA 3005 GRADUATION EXERCISE

PREDICTION OF FOOTBALL MATCH RESULTS USING STATISTICAL TECHNIQUES

MUHAMAD AFNAN DARWISY BIN HAMZAH

(17124877)

FACULTY OF ECONOMICS AND ADMINISTRATION

UNIVERSITY MALAYA

SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENT FOR A DEGREE

OF BACHELOR OF ECONOMICS

ACKNOWLEDGEMENT

The completion of this study could not have been accomplished without the contribution of data from 11v11. During the course of this inquiry, the author would also like to extend their gratitude to Dr. Ahmad Farid Bin Osman for offering insightful input and for freely sharing his broad expertise. In addition, the author would like to take this opportunity to thank his family and friends for their unwavering support during the entirety of this experience. Any errors are completely my fault, and they were not made with the intention of doing anything negative to the reputations of the people who were involved.

ABSTRACT

The use of observed and projected match statistics as inputs to forecast for the outcomes of football matches is the aim of investigation in this study. It has been proved that extremely accurate projections of the outcome of a match could be produced if the statistics of the match were available before the event began. If the pre-match statistics of a match can be predicted, and if those predictions are precise enough, then it is possible to make informative match forecasts. The Poisson distribution is used to demonstrate approaches to predict match statistics. Events in a Poisson distribution occur at random and independently. By utilizing previous match data and locating an ideal prediction approach based on a Poisson regression model that generates probabilities on the outcomes of assigned matches. This method is based on the use of historical match data. This study tries to investigate the potential match outcome of Manchester United against Liverpool in all matches in the English Premier League (1894-2022) and use it for prediction using the Poisson Regression in SPSS (Version 22).

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND OF STUDY

It is without a doubt one of the most significant revolutions in the industry all over the world that sports analytics, statistics, and the utilisation of data in its many forms have taken the sport sector seriously over the past several decades. However, association football has not been in the forefront of the data revolution that has been sweeping through the sporting world. Football, despite being the most popular sport on the planet, has not been able to match the analytical modelling quality of other major sports. This is even though football is the most popular sport on the planet.

The Poisson Distribution assumes that events occur at random and independently of one another. Activities take happen within a particular time frame. Modeling with a Poisson Distribution requires that the distribution be based on the total number of events.

The Poisson distribution has gained widespread acceptance as a fundamental modelling strategy for determining how the number of goals scored in sports that have two teams in direct competition with one another. In team sports such as football and basketball, where the two teams compete against one another during the game, it is reasonable to assume that the two result variables relate to one another. Additionally, as the goal of the two teams competing sports games is to score in the same sequence, the faster pace at which one team plays results in an increased number of scoring opportunities for both teams. (Karlis & Ntzoufras, 2003).

1.2 PROBLEM STATEMENT

As there are so many distinct approaches, scenarios, and game systems that may be used, football is a sport that offers a great deal of variety. It might be difficult to predict the outcome of certain football matches because there are so many different factors at play. The Poisson Distribution is based on the idea that occurrences take place at random and

are unrelated to one another. During a certain period, certain activities will take place. When modelling with a Poisson Distribution, the distribution should be based on the number of events that occur in football matches. It is possible that the statistic that determines how likely it is for a particular football club to come out on top of a given game situation will end up being rather significant in the context of the sport. Based on the considerations presented above, it is of the utmost importance to have statistical tools at one's disposal that can be utilised to make projections of the outcomes of these matches.

1.3 SIGNIFICANCE OF STUDY

Data analytics is becoming increasingly commonplace in all aspects of our lives, including businesses of all sizes, healthcare, the media, and sports. Up until a few years ago, it was believed that football would be able to withstand this trend. It is necessary to have a solid understanding of the probabilities associated with any event based on its history of occurrence. The probabilities are not simply plucked out of thin air when one looks at the prediction. A mathematical approach is used, whether it be for calculating a single game event or a set of occurrences, respectively. When it comes to prediction, the Poisson distribution is of utmost significance in sports like football and other similar activities that use a point system based on increments. It is a factor that is considered while calculating the likelihood of each possible score.

1.4 RESEARCH OBJECTIVES

The objective of this work is to develop a generalised linear model with Poisson regression that can be applied to the problem of predicting the results of matches between Manchester United and Liverpool played throughout the course of the season. By generating more accurate match outcome probabilities, we are going to test this model to determine how well it can accurately represent the outcomes of matches and how successful a team is over the course of a season.

1.5 RESEARCH QUESTIONS

This is the goal of this research to provide information in response to the following questions:

a) to what extent this prediction model is able to accurately portray match outcomes and success over the course of a season.
b) whether the accuracy of the model can be portrayed using a Poisson distribution.

1.6 SCOPE OF STUDY AND RESEARCH

This study attempts to investigate the potential match outcomes of Manchester United versus Liverpool in the English Premier League (1894-2022) by analysing the goals scored data in those matches and using it for prediction. The Poisson Regression in SPSS (Version 22) is supposedly used for this investigation. The generalised linear model was used to assess the statistical models that were suggested and analysed in this work. It is anticipated that this project will gather, calculate, and predict the outcomes of matches between those matches that will take place in the future.

1.7 STRUCTURE OF REPORT

The findings of this study are presented in five separate chapters. The research problem is presented in Chapter 1, which also provides background information on the project. This chapter provides a concise explanation of how the Poisson distribution can be used to analyse football match results to make predictions. In addition, it specifies the parameters of the scope and the framework of the investigation.

In the second chapter, an in-depth analysis of the statistical aspects of football and the Poisson distribution is provided. It does so by compiling the findings of prior research that has been conducted on the broad topic of football statistics and data. The research investigates how well this prediction model can accurately forecast the outcomes of matches and the overall success of a season. Following that, the research investigates how the Poisson Distribution can be used to explain the likelihood that a certain outcome will occur in a football match.

The methodology of the study is broken down in Chapter 3. It describes the process for collecting the data, the size of the sample, the variables being studied, the analysis approach, and the research framework for this study. In addition, this chapter explains the factors of the study, as well as the data collection and analysis methodologies that were utilised.

In the fourth chapter, we conduct an in-depth analysis of the research data. The examination of the research data by comparing with actual recent results and the search for the limitations of the research are presented in this chapter, which is quite important. The outcomes of the study, as well as potential areas for development, are concluded in Chapter 5.

CHAPTER 2: LITERATURE REVIEW

2.1 PREDICTING AND EXPLAINING THE RESULT OF A FOOTBALL MATCH

Athletes, coaches, owners, and gamblers are beginning to understand the importance of quantitative sports analysis to get an advantage over their rivals. As a result, the field of quantitative sports analysis is experiencing tremendous growth. This has obviously given rise to people's need to obtain information that will assist them in making more informed decisions. According to Wheatcroft (2021), football is the most popular sport in the world. In the past, football's use of quantitative analysis has fallen behind that of other sports, although this is gradually starting to change. In recent decades, there has been a rise in interest in predicting the outcomes of football matches, which has contributed to an increase in the demand for analytical research study.

"You can plan, but you can't guarantee what will happen on a football ground," legendary German goalkeeper Manuel Neuer famously observed. This idea contributes to an explanation for why football is the most popular sport on the entire earth. On the other hand, Taha Yasseri (2021) claims that the outcomes of football matches are becoming increasingly predictable.

A significant amount of match data is recorded in the world's most renowned football leagues, which are played in modern times. While match statistics such as the number of shots, corners, and fouls committed by each team are not available for free, it is possible to make a purchase for data pertaining to the location and outcome of every event that occurred during the match. People that can understand data in a meaningful way will find that this opens a lot of doors for them. This work focuses on the probabilistic prediction of the outcomes of football matches, such as determining whether the game will end in a victory for the home team, a tie, or a victory for the away team. A probabilistic forecast of such an event is comprised of the estimated probabilities that have been placed on each of the three probable outcomes. The use of statistical models allows for the incorporation of information into probabilistic forecasts.

According to (Guillermo Martinez Arastey, 2019), the first proponent was a British Royal Air Force accountant named Charles Reep. After World War II, Reep started utilising a

pencil and paper to gather and analyse statistics on football matches. According to the findings of Reep's analysis, most goals are scored with fewer than three passes, highlighting the importance of pushing the ball forward as rapidly as possible. His theory, which became popularly known as the long ball, would have a big influence on English football for several years to come, especially in the 1980s. Always emphasising a straightforward aggressive approach to the game, Reep's playing career included stops at Brentford, Wolverhampton Wanderers, and Sheffield Wednesday. He also spent time with Wimbledon, Watford, and the Norwegian national team.

Certain match data, such as the number of shots or corners each team had, were made accessible in some way prior to the beginning of the game. In such a scenario, it is reasonable to assume that one would be able to use the knowledge to the formulation of accurate projections, and it has been shown that this presumption was correct. It should come as no surprise that none of this information would ever be divulged in advance. On the other hand, if data collected from prior games can be extrapolated to forecast the outcomes of the current game before it ever begins, and if the accuracy of those projections is high enough, those statistics can be utilised to generate accurate predictions for the upcoming match (Wheatcroft, 2021)

2.2 THE USE OF POISSON DISTRIBUTION IN PREDICTION

For a performance trait or event to have any significance, it must be connected to the accomplishments of the team, such as winning matches. In order to perform such an analysis, you will need to collect a great deal of data using a variety of sensors, cameras, and analytic systems. Additionally, you will need a clever strategy to invest in this large data, which are often linear and nonlinear in nature. In this regard, the use of the Poisson distribution for the identification, categorization, and prediction of performance in soccer has seen a significant increase in recent years (Hassan et al., 2020). It is now feasible to make predictions based on player statistics thanks to the analysis of data collected from previously recorded matches.

Poisson regression is a method for making predictions about "count data" dependent variables when given one or more independent variables to work with as inputs. The variable that we want to forecast is known as the dependent variable. We make our predictions about the value of the dependent variable based on the independent variables, which are factors.

The use of these performance indicators has become increasingly widespread because of the ease with which attributes and data from matches may be obtained through research. In recent years, the Poisson distribution has become an increasingly popular choice for modelling and estimating the qualities that lead a team to win or lose a match, as well as for predicting the outcome of a specific match. However, match data modelling has not been successful to this point in comprehending and evaluating the vast amount of data that is accessible while simultaneously allowing it to be simplified objectively. (Hassan et al., 2020) added that because it is difficult to determine the actual contribution value of complex variables on match results, transferring such match attributes into the training process has been overlooked in sport-related scientific literature up until this point. This is because it is difficult to determine the actual contribution value of complex variables.

CHAPTER 3: METHODOLOGY

3.1 RESEARCH FRAMEWORK

Through conduct this research, all Manchester United vs Liverpool league games from the English Premier League (236) from 1894 to 2022 were examined. 11v11.com, the official website of the Association of Football Statisticians, supplied the collected data set for the matches. This website is powered by a unique database of international football and the English Football League, including the FIFA World Cup and the FA Premier League, which have been running since their start (*International Football History and Statistics - 11v11*, 2022). We used secondary data collection as a strategy for gathering information. This method uses existing data, which takes less time than gathering new data and is more likely to be reliable and valid because it was obtained by a trained statistician or researcher. This included goal for and goal against Manchester United in those matches.

The dataset was subjected to a Poisson regression using SPSS (Version 22). All the goals in the dataset were processed using SPSS so that they could be examined in multiple tests. A careful examination of the data revealed no obvious flaws. The two independent variables (goal for and goal against) are crucial components in building a model that may be used to run a statistical test like Poisson regression. In this study, the match outcome (*win/lose/draw*) would be our dependent variables that should be predicted using both independent variables (*goal for and goal against*).

This is to ensure that the time series data being analyzed is non-linear so that the regression could continue. As a result, we use a scatterplot diagram to test for linearity. This indicates that the scatterplot's points closely resemble a straight line. If one variable increases at roughly the same pace as the other variables changes by one unit, the relationship is linear. Then, the Kolmogorov-Smirnov test was used in this investigation to determine if the data was normal or not. So, we could persuade ourselves that the data is from a Poisson distribution. The data analysis is continued in this study with the use of the Generalized Linear Model to compute the Poisson regression. To define how the dependent response variables are influenced by the explanatory variables, a linear model

is utilized. We'll then put the data to the test to see if there's any overdispersion. When the variance of the response variable is greater than the predicted value of this response variable, overdispersion occurs in Poisson regression.

3.2 SCATTERPLOT

The association between two quantitative factors that were examined for the same individuals can be graphically represented using a scatterplot. The values of one variable appear along the horizontal axis, and the values of the other variable appear along the vertical axis in this graph. Numerous research initiatives are correlational studies because they examine the possible correlations between variables. Before examining the relationship between two quantitative variables, it is always beneficial to generate a graphical representation of both variables. This type of graphic depiction is known as a scatterplot (Mindrila & Balentyne, 2017). This study examines the correlation between Manchester United's match outcome (win, loss, or tie) and their goals scored for and against Liverpool. Thus, we use scatterplot to check whether the dataset is linear or not.

3.3 KOLMOGOROV-SMIRNOV TEST

The Kolmogorov-Smirnov test is applied to data to determine whether or not a sample may be considered representative of a population that has a specific distribution (Chakravart, Laha, and Roy, 1967). One of the appealing aspects of this test is that the distribution of the K-S test statistic does not depend on the underlying cumulative distribution function that is being assessed. This is one of the reasons why this test is so popular. The fact that it is a test whose precision can be relied upon is still another advantage of using it (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test does have a few important disadvantages, the most notable of which are that it can only be used with continuous distributions and that it tends to be more sensitive in the middle of the distribution than it is at the tails of the distribution. In this part of the analysis, one sample Kolmogorov-Smirnov test is used to determine whether the dataset is normal, or Poisson distributed, and the significance value of the test's output is determined. It is quite uncommon for there to be such a large deviation in percentage, which shows that the goals scored throughout the matches did not follow a normal distribution across the total

population. Therefore, a significant deviation will have a low p-value. If the value of p is less than 0.05, we take the position that the null hypothesis cannot be true. If p is less than 0.05, then we do not have sufficient evidence to conclude that the variable in question follows a normal distribution in the population.

3.4 OVERDISPERSION

In Poisson regression, an issue that frequently arises is known as overdispersion. In Poisson regression, the term "overdispersion" refers to the situation in which the variance of the response variable is larger than the value that was anticipated for this response variable. To measure the accuracy of the calculation, we will recognize the ratio of the dependent variable's variance to its mean. The Poisson distribution assumes that the ratio is 1. (That is, the mean and the variance are both the same). As a result, we are able to recognise that even before we take into account any variables that can contribute to a better understanding of the data, there is still a possibility of either overdispersion or underdispersion. Overdispersion may result from the interdependence of observations.

3.5 POISSON REGRESSION

Modeling count data requires an extended linear model, and the Poisson regression model is one such approach. This Poisson distribution is utilised to develop this regression model by allowing the rate parameter to depend on the variables that are being used to explain the data. This is accomplished through the usage of the regression model. The Poisson regression model specifies that the dependent variable Y, given independent variables (*goal_for, goal_against*) $x_1, x_2, \ldots, x_k$, follows a Poisson distribution with the probability function,

$P\ (Y = y|x_1, x_2,\ldots, x_k) = \frac{\lambda^y e^{-\lambda}}{y!}$, $y = 0, 1, 2, \ldots,$

where the rate $\lambda = Exp\ (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$, or, equivalently, $ln\ \lambda = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.

When performing Poisson Regression, it is common practice to think of Y as a rate. Poisson models are non-linear; therefore, coefficients do not have a straightforward linear interpretation (Pearson, n.d.). So, we use the log form of the model, which multiplies coefficients to help with interpretation by giving incidence rate ratios. Positive coefficients suggest a higher rate, and negative coefficients indicate a lower rate. In SPSS (Version 22), we use the Poisson Loglinear model in Generalized Linear Models to calculate the possible outcome for our results. The $Exp\ (\widehat{\beta_1})\cdot 100\%$ represents the estimated percent ratio in mean response for the level $x_1 = 1$ and that for the reference level, provided the other X variables are unchanged. These percentages were considered as the expected outcome for goals for and goals against value (probability). Therefore, it will be our model as potential outcome for the upcoming match for Manchester United against Liverpool.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 DATA ANALYSIS

Table 4.1: Descriptive statistics for both independent variables (*goal_for, goal_against)*

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
goal against	236	1.38	1.371	1.878
Valid N (listwise)	236			

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
goal for	236	1.36	1.228	1.508
Valid N (listwise)	236			

According to the statistical test above, mean, standard deviation and variance for both independent variables (*goal_for*) are 1.36, 1.228 and 1.508, while for (*goal_against*) are 1.38, 1.371 and 1.878 respectively. The standard deviation of both data is relatively distributed near their respected mean value.

Figure 4.1: Scatterplot between *goal_against and win/lose/draw*

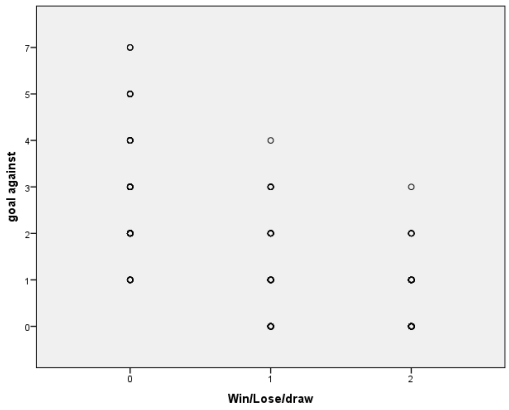
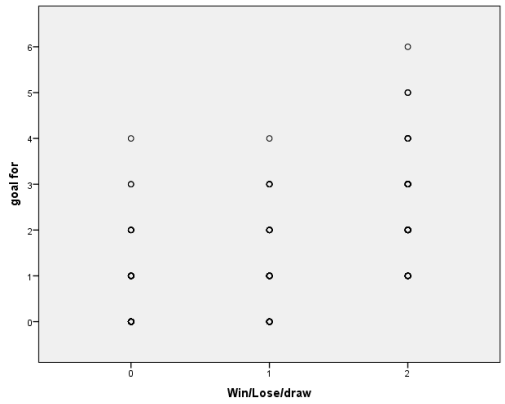


Figure 4.2: Scatterplot between *goal_for and win/lose/draw*



The scatterplot is used in this statistical test to determine whether the data is linear or not. It shows that the data points are spread out evenly in these graphs. This means that there is no trend to the data; thus, the data is not linear.

Table 4.2: One-Sample Kolmogorov-Smirnov for normal distribution test

One-Sample Kolmogorov-Smirnov Test

		goal for	goal against
N		236	236
Normal Parameters[a,b]	Mean	1.36	1.38
	Std. Deviation	1.228	1.371
Most Extreme Differences	Absolute	.209	.227
	Positive	.209	.227
	Negative	-.134	-.158
Test Statistic		.209	.227
Asymp. Sig. (2-tailed)		.000[c]	.000[c]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

Table 4.3: One-Sample Kolmogorov-Smirnov for Poisson distribution test

One-Sample Kolmogorov-Smirnov Test 2

		goal for	goal against
N		236	236
Poisson Parameter[a,b]	Mean	1.36	1.38
Most Extreme Differences	Absolute	.036	.044
	Positive	.036	.044
	Negative	-.017	-.021
Kolmogorov-Smirnov Z		.549	.681
Asymp. Sig. (2-tailed)		.924	.743

a. Test distribution is Poisson.

b. Calculated from data.

The One-Sample Kolmogorov-Smirnov Test is used for test normality and Poisson distribution of those both independent variables. For normality test, the significance value (p) for both independent variables are 0.000 and 0.000. As a rule of thumb, we reject the null hypothesis if $p < 0.05$. So, if $p < 0.05$, we did not believe that our variable follows a normal distribution in our sample. While the test for Poisson distribution is conducted and it shows the significance value (p) for both independent variables are 0.924 and 0.743. So, we accept the null hypothesis since $p > 0.05$ and believes that both independent variable follows the Poisson distribution.

4.2 GENERALIZED LINEAR MODELS

Table 4.4: Model Information of the model.

Model Information

Dependent Variable	Win/Lose/draw
Probability Distribution	Poisson
Link Function	Log

The displayed Model Information table shows that the dependent variable is "Win/Lose/Draw," that the probability distribution is "Poisson," and that the link function is the natural logarithm (i.e., "Log").

Table 4.5: Continuous Variable Information of the model.

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Win/Lose/draw	236	0	2	1.03	.848
Covariate	goal for	236	0	6	1.36	1.228
	goal against	236	0	7	1.38	1.371

The Continuous Variable Information table does a basic check of the data to see if there are any issues, but it is not as helpful as other descriptive statistics that you may run independently before you run the Poisson regression. However, there is a way to gain an understanding of whether there might be overdispersion in the analysis by considering the ratio of the variance to the mean for the dependent variable (*win/lose/draw*). The mean is 1.03 and the variance is 0.7191 (0.848^2), which is a ratio of $1.03 \div 0.7191 = 0.6981$. The Poisson distribution is based on the assumption of a ratio of 1. (i.e., the mean and variance are equal). Therefore, we can see that there is a minor amount of underdispersion before any independent variables are included.

Table 4.6: Goodness of Fit of the model.

Goodness of Fit[a]

	Value	df	Value/df
Deviance	96.874	233	.416
Scaled Deviance	96.874	233	
Pearson Chi-Square	72.520	233	.311
Scaled Pearson Chi-Square	72.520	233	
Log Likelihood[b]	-230.440		
Akaike's Information Criterion (AIC)	466.880		
Finite Sample Corrected AIC (AICC)	466.983		
Bayesian Information Criterion (BIC)	477.271		
Consistent AIC (CAIC)	480.271		

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst[a]

a. Information criteria are in smaller-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

The table titled "Goodness of Fit" gives a number of different metrics that can be utilised in order to evaluate how well the model matches the data. However, the value that can be found in the "Value/df" column for the "Pearson Chi-Square" row is called attention to because it is the one that is used to determine whether or not the data are evenly distributed. This value is 0.311. Equidispersion is indicated when the value is 1, whereas overdispersion is indicated when the value is greater than 1, and underdispersion is indicated when the value is less than 1. The value displayed in the table demonstrates that the data have an inadequate amount of dispersion. It indicates that there was a smaller amount of fluctuation in the data compared to what was projected.

Table 4.7: Omnibus test of the model.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
132.908	2	.000

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst[a]

a. Compares the fitted model against the intercept-only model.

A likelihood ratio test is displayed in the table for the Omnibus Test to determine whether all of the independent variables together make the model more accurate than the intercept-only model (i.e., with no independent variables added). A statistically significant overall model is shown by a p-value of 0.000 for the independent variables that were assessed.

Table 4.8: Tests of model effects.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.168	1	.682
goalfor	60.314	1	.000
goalagainst	63.056	1	.000

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst

The Tests of Model Effects table, in its current format, provides insight into the degree to which each of the independent variables has an impact on the model by calculating each variable's significance value. The fact that both independent variables (goalfor and goalagainst) that were tested were assigned a p-value of 0.000 (that is, $p = 0.000$) indicates that the entire model is statistically significant. Both independent variables that were evaluated were goals for and goals against.

Table 4.9: Parameter estimates of the model.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	.048	.1170	-.181	.277	.168	1	.682	1.049	.834	1.320
goalfor	.342	.0441	.256	.429	60.314	1	.000	1.408	1.292	1.535
goalagainst	-.583	.0734	-.726	-.439	63.056	1	.000	.558	.484	.645
(Scale)	1[a]									

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst

a. Fixed at the displayed value.

The goal for variable (i.e., the "*goalfor*" row). The exponentiated value is 1.408. Then we can compute it in $Exp\ (\widehat{\beta_1})\cdot 100\%$ which give us the value 40.8%. This means that the match outcome that will win ("*win/lose/draw*") will be 1.408 times or 40.8% greater for each extra goal whenever Manchester United scores. However, the goal against variable (i.e., the "goalagainst" row). The exponentiated value is 0.558. Then we can compute it in $Exp\ (\widehat{\beta_1})\cdot 100\%$ which give us the value 55.8%. This means that the match outcome that will win ("*win/lose/draw*") will be 0.558 times or 55.8% lesser for each extra goal whenever Manchester United concedes.

4.3 DISCUSSIONS

Figure 4.3: Count of Manchester United Win/Lose/Draw vs Liverpool

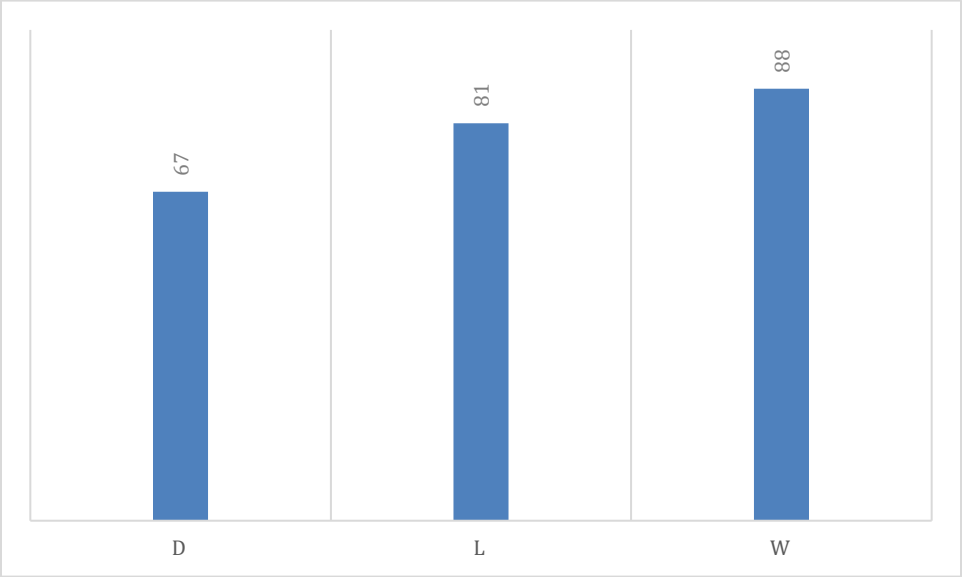


Figure 4.4: Average goal for Manchester United vs Liverpool when they win, lose, or draw the match.

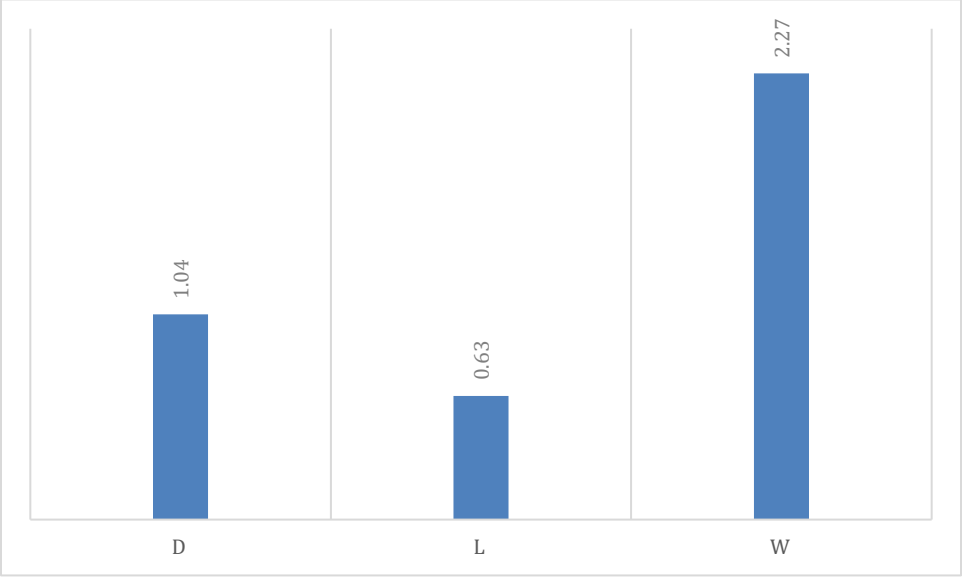
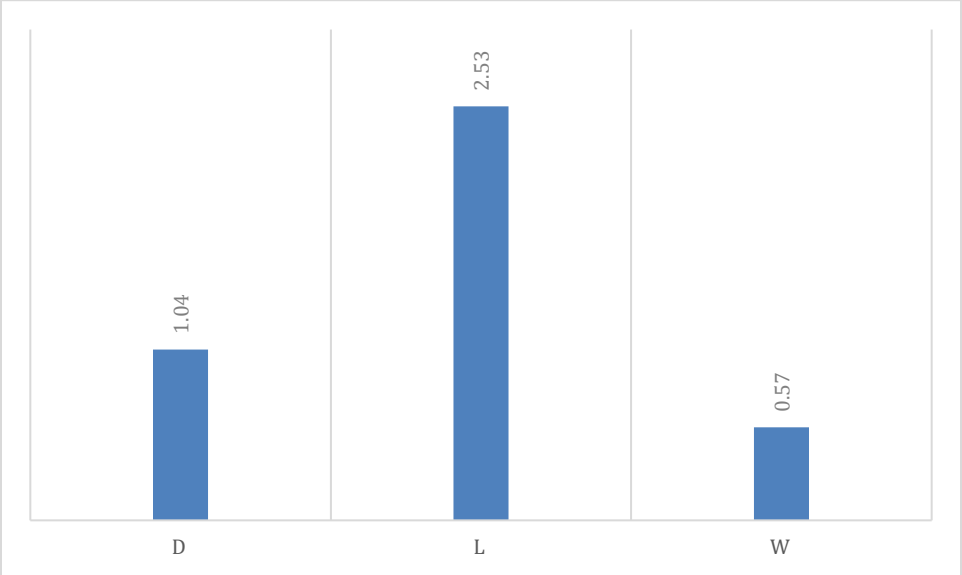


Figure 4.5: Average goal against Manchester United vs Liverpool when they win, lose, or draw the match.



As football is a sport with a relatively low scoring average, our capacity to assess the possibility of a goal being scored is critical for providing vital context when attempting to forecast or explain the outcome of a football game. People who follow the sport of football on a regular basis won't be surprised to learn that the quality of the various teams is not always comparable; hence, different sides are likely to be expected to score a varying number of goals throughout the course of a match. Despite this, there was once a time when providing a satisfactory response to the question of whether football was a game of talent or chance was challenging. For our cases, we can see the count of win that Manchester United had, it shows that they are more likely to win against Liverpool throughout many seasons. Manchester United have average goals scored of 2.27 goal for every game they won against Liverpool. Manchester United have average goals conceded of 0.57 goal for every game they won against Liverpool. we can see the count of loss that Manchester United had, it shows that they are moderately lose against Liverpool throughout many seasons. Manchester United have average goals scored of 0.63 goal for every game they lose against Liverpool. Manchester United have average goals conceded of 2.53 goal for every game they won against Liverpool. In upcoming matches, as we used our regression model to predict the match outcome. We predicted that next match of Manchester United vs Liverpool using our probability model with 1.408 and 0.558. This means that the match outcome that will win will be 40.8% greater for each extra goal

whenever Manchester United scores or 55.8% lesser for each extra goal whenever Manchester United concedes as we calculated. Yet, the recent actual result match between Manchester United vs Liverpool that took place on 20 April 2022 (*Liverpool v Man Utd, 2021/22* | *Premier League*, 2022) was reasonable. The results shown that Manchester United loss to Liverpool by 0 to 4. It satisfies our probability results that every goal conceded by Manchester United, the game is 55.8% likely to be a loss to them.

However, there are several limitations in this model as the actual football match results likely tend to be influenced by the form and the morale in the individual of a team. The morale and form of each player in a team is one of the factors that could bring a positive result to the team. Given the potentially excessive degradation that the audience's behaviour may present during a competition, it is an external aspect that has the potential to affect the performance of an athlete while they are competing. An audience that is encouraging and cheering for the squad can be credited as being one of the most common positive effects on the team's motivation (Calleja et al., 2022). Having a good crowd and audiences in certain matches could boost the morale's player. The player will be comfortable and in a good atmosphere to maximize their full potential in a pitch and this will affect the outcome of a match. Other than that, bad officials could bring chaos in the pitch. That would be to imply, the officials in charge of determining who the winner is in a sporting event should strive to avoid making as many errors and fabrications as they can. In order to increase the likelihood that this will in fact be the case, referees are installed to ensure that the competition is conducted in a fair manner (MacMahon et al., 2015). The amount of effect they can potentially have over the result of a soccer game is enormous. As an illustration, at least two finals of the World Cup were determined by judgments made by the referees that were met with controversy (Tamir & Bar-eli, 2021). These are likely to affect the match results that limits our research.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

This study has shown the generalized linear model with Poisson regression that has been used to predict the outcome of Manchester United vs Liverpool matches throughout the season. A Poisson regression was run to predict the match outcome of Manchester United vs Liverpool based on (goal for and goal against) of Manchester United from 1894-2022 vs Liverpool. For every extra goal for Manchester United, 1.408 (95% CI, 1.292 to 1.535) times or 40.8% more likely to win the game, a statistically significant result, p = .000. While, for every extra goal against Manchester United, 0.558 (95% CI, 0.484 to 0.645) times or 55.8% less likely to win the game, a statistically significant result, p = .000. Predicting the outcomes of football games accurately is notoriously tough, and our approach is no exception. There is an infinite number of factors that may be considered, such as possession, free kicks, headers, chances created, and dribbles, and they will all have some effect on the final score of a match (Rathke, 2017). Our model is not reliable enough to rely on to predict the match results of a football match. We used the Poisson regression In Generalized Linear Model to find the probability of winning the matches by goals for Manchester United against Liverpool. We suggest for future improvement on expected match outcome research that try to use other models for predicting the match results. There is research on predict match outcome by comparing on expected goals (xG) values by shot and pass types (Caley, 2013).

CHAPTER 6: REFERENCES

Caley, M. (2013, November 16). Shot Matrix III: The Incredible Through-Ball. Cartilage Free Captain; Cartilage Free Captain. https://cartilagefreecaptain.sbnation.com/2013/11/16/5111212/shot-matrix-iii-the-incredible-through-ball

Calleja, Paul & Muscat, Adele & Decelis, Andrew. (2022). The Effects of Audience Behaviour on Football Players' Performance. 10.22103/JNSSM.2022.18890.1055.

Chakravarti, Laha, and Roy, (1967). *Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons*, pp. 392-394.

Guillermo Martinez Arastey. (2019, November 27). *Sport Performance Analysis*. Sport Performance Analysis. https://www.sportperformanceanalysis.com/article/history-of-performance-analysis-the-controversial-pioneer-charles-reep

Hassan, A., Akl, A.-R., Hassan, I., & Sunderland, C. (2020). Predicting Wins, Losses, and Attributes' Sensitivities in the Soccer World Cup 2018 Using Neural Network Analysis. *Sensors*, *20*(11), 3213. https://doi.org/10.3390/s20113213

How to perform a Poisson Regression Analysis in SPSS Statistics | Laerd Statistics. (2018). Laerd.com. https://statistics.laerd.com/spss-tutorials/poisson-regression-using-spss-statistics.php

Ibrahim Kovan. (2021, November 10). *What is Poisson Distribution? When to use it? How to predict football match results?* | *Towards Data Science*. Medium; Towards Data Science. https://towardsdatascience.com/predicting-football-match-result-using-poisson-distribution-ac72afbe36e0

Karlis, D., & Ntzoufras, I. (2003). Analysis of Sports Data by Using Bivariate Poisson Models. *Journal of the Royal Statistical Society. Series D (the Statistician)*, *52*(3), 381–393. https://www.jstor.org/stable/pdf/4128211.pdf?refreqid=excelsior%3A3e436609fce77ee71d6c0dfe78e28625&ab_segments=&origin=

Liverpool v Man Utd, 2021/22 | Premier League. (2022). Premierleague.com; https://www.premierleague.com/match/66635

MacMahon, C., Mascarenhas, D., Plessner, H., Pizzera, A., Oudejans, R., & Raab, M. (2014). *Sports officials and officiating: Science and practice*. Routledge.

Manchester United football club: record v Liverpool. (2012). 11v11.com. https://www.11v11.com/teams/manchester-united/tab/opposingTeams/opposition/Liverpool/

Pearson, F. (n.d.). *Poisson Regression*. http://www.netph.sgul.ac.uk/training-materials/advanced-epidemiology-and-modelling-course/Poisson%20Regression_AEC.pdf

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise, 12(Proc2). https://doi.org/10.14198/jhse.2017.12.proc2.05

Tamir, I., & Bar-eli, M. (2021). The Moral Gatekeeper: Soccer and Technology, the Case of Video Assistant Referee (VAR). *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.613469

Wheatcroft, E. (2021). Forecasting football matches by predicting match statistics. *Journal of Sports Analytics*, *7*(2), 77–97. https://doi.org/10.3233/jsa-200462

CHAPTER 7: APPENDICES

Descriptive statistics for both independent variables (*goal_for, goal_against)*

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
goal against	236	1.38	1.371	1.878
Valid N (listwise)	236			

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
goal for	236	1.36	1.228	1.508
Valid N (listwise)	236			

One-Sample Kolmogorov-Smirnov for normal and Poisson distribution test.

One-Sample Kolmogorov-Smirnov Test

		goal for	goal against
N		236	236
Normal Parameters[a,b]	Mean	1.36	1.38
	Std. Deviation	1.228	1.371
Most Extreme Differences	Absolute	.209	.227
	Positive	.209	.227
	Negative	-.134	-.158
Test Statistic		.209	.227
Asymp. Sig. (2-tailed)		.000[c]	.000[c]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

One-Sample Kolmogorov-Smirnov Test 2

		goal for	goal against
N		236	236
Poisson Parameter[a,b]	Mean	1.36	1.38
Most Extreme Differences	Absolute	.036	.044
	Positive	.036	.044
	Negative	-.017	-.021
Kolmogorov-Smirnov Z		.549	.681
Asymp. Sig. (2-tailed)		.924	.743

a. Test distribution is Poisson.

b. Calculated from data.

Generalized linear model selection for Poisson regression.

Model Information

Dependent Variable	Win/Lose/draw
Probability Distribution	Poisson
Link Function	Log

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Win/Lose/draw	236	0	2	1.03	.848
Covariate	goal for	236	0	6	1.36	1.228
	goal against	236	0	7	1.38	1.371

Goodness of Fit[a]

	Value	df	Value/df
Deviance	96.874	233	.416
Scaled Deviance	96.874	233	
Pearson Chi-Square	72.520	233	.311
Scaled Pearson Chi-Square	72.520	233	
Log Likelihood[b]	-230.440		
Akaike's Information Criterion (AIC)	466.880		
Finite Sample Corrected AIC (AICC)	466.983		
Bayesian Information Criterion (BIC)	477.271		
Consistent AIC (CAIC)	480.271		

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst[a]

a. Information criteria are in smaller-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
132.908	2	.000

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst[a]

a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.168	1	.682
goalfor	60.314	1	.000
goalagainst	63.056	1	.000

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	.048	.1170	-.181	.277	.168	1	.682	1.049	.834	1.320
goalfor	.342	.0441	.256	.429	60.314	1	.000	1.408	1.292	1.535
goalagainst	-.583	.0734	-.726	-.439	63.056	1	.000	.558	.484	.645
(Scale)	1[a]									

Dependent Variable: Win/Lose/draw
Model: (Intercept), goalfor, goalagainst

a. Fixed at the displayed value.