ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 4 No. 2 (2024) 349-358
https://publisher.uthm.edu.my/periodicals/index.php/ekst

The Poisson Regression and Quasi-Poisson Regression Analysis on FIFA World Cup Games

Kang Yue Teng[1], Norziha Che Him[1*]

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 86400 Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: norziha@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2024.04.02.038

Article Info

Received: 27 December 2023
Accepted: 11 January 2024
Available online: 12 December 2024

Keywords

FIFA World Cup, Football, Poisson Regression, Quasi-Poisson Regression, AIC, BIC

Abstract

The FIFA World Cup's surging popularity has attracted a diverse fan base, including passionate enthusiasts in Malaysia. This widespread interest has motivated researchers to delve into the details of the tournament, using it as a crucial platform for predictions, team performance evaluations, and an exploration of international football competition at its top. This study aims to predict number of goals in World Cup matches, employing Poisson and quasi-Poisson regression models. The optimal model is determined through a comprehensive assessment, considering AIC, BIC, standard error, and p-values. Utilizing a dataset from Kaggle, originally sourced from the official FIFA website, the findings consistently associate variables such as goal inside the penalty area, goal outside the penalty area, left inside channel, right channel, attempted defensive line breaks, completed defensive line breaks, yellow cards, passes, and own goals with a raised probability of number of goals. Significantly, the quasi-Poisson model exhibits a superior fit, as evidenced by its lower AIC value of -50.0853 and a reduced deviance value of 38.7935. Consequently, the quasi-Poisson regression model emerges as a more suitable choice than the Poisson regression model, particularly for addressing the overdispersion inherent in the data.

1. Introduction

Soccer or football is one of the most popular and widely followed sports globally. Since 1994, football has consistently maintained its status as the most popular sport [1]. Upon conducting a search, the term `soccer' generated at least 11,301 records [2]. Since its basic rules and minimal equipment needs, football, often addressed as the "king of sports", is adaptable to diverse settings such as grass fields, indoor gyms, streets, parks, and even beaches [3]. In the realm of football, the most iconic tournament is the FIFA World Cup. The importance of the World Cup in the sports community has led many organisations to develop various models for predicting match outcomes [4].

FIFA or the Fédération Internationale de Football Association, is a non-governmental organisation (NGO) headquartered in Switzerland. Serving as the international governing body for association football, beach football, and futsal, FIFA is responsible for various aspects of international soccer. This includes organising the quadrennial football World Cup competition and overseeing several other international tournaments [5]. Established on May 21, 1904, in Paris, FIFA stands as one of the world's oldest and largest NGOs. FIFA organises the World Cup every four years, except for 1942 and 1946 when the event was cancelled due to the Second World War. Since its start in 1930 in Uruguay, the FIFA World Cup has transformed into one of the most eagerly awaited and extensively viewed sporting events globally. The tournament typically spans approximately one

month and is hosted in a different country for each edition. The FIFA World Cup stands as a prominent global event, exerting a substantial social impact on both the host country and the global community.

Humans face limitations when it comes to handling vast amounts of data [6]. Given its widespread popularity, football engages many participants at both professional and unprofessional levels, leading to the generation of a large amount of data. An example of a large amount of data in football could be the extensive statistics and metrics collected during matches. This includes player performance data such as the number of goals, assists, shots on target, passes completed, tackles made, possession percentages, and more. Additionally, data on team statistics, injury reports, weather conditions during matches, and fan engagement metrics on social media platforms contribute to the vast amount of information generated in the realm of football. Consequently, there is a growing need to analyse this data. When specifically considering World Cup games, various aspects demand analysis, including assessing team performances, evaluating individual player performances, examining goal-scoring patterns, analysing defensive strengths and weaknesses, and other pertinent factors. Examining FIFA World Cup games is essential for developing tactical insights, assessing performance, deriving statistical information, scouting players, involving fans, catering to the media, and establishing a historical perspective. This analysis is pivotal in advancing the sport, providing valuable insights for teams and coaches, and enriching the overall World Cup experience for fans worldwide.

Poisson regression models are frequently used in the examination of count data. A FIFA soccer match lasts for a total of 90 minutes, divided into two halves of 45 minutes each, separated by a brief break. The occurrence of goal scoring during a match is unpredictable, happening randomly throughout the game's duration. The Poisson distribution is well-suited for modelling the number of events within a fixed interval of time or space, making it applicable to represent the count of goals scored during the 90 minutes of play. In the soccer context, a major example is the low probability of every shot on goal resulting in a score. Numerous conditions must be met to score a goal in a soccer match, expressing a rare occurrence mathematically. Consequently, the count of rare events, such as the number of goals scored in a match, is depicted as the outcome of discrete trials [7]. The Poisson distribution is designed for modelling count data, making it fitting for scenarios involving the prediction of the number of occurrences of an event within a fixed interval or area, such as the count of goals scored in a football match.

Poisson regression is designed for modelling count data, but its reliance on the assumption of equidispersion may not always hold in real-world application data. An alternative to address this issue is the use of Quasi-Poisson regression. Quasi-Poisson regression is preferred in many cases because it is readily accessible in statistical software and can be extended from traditional Poisson regression situations [8]. This regression model proves to be a feasible option when dealing with count data marked by a phenomenon called overdispersion, wherein the variance surpasses the mean. By ensuring robust standard errors, Quasi-Poisson regression facilitates reliable statistical inferences. It provides an adaptable and practical approach for modelling real-world count data, overcoming the limitations associated with traditional Poisson regression [9]. Confronted with overdispersion, utilizing the Poisson regression model can lead to skewed parameter estimates and inaccurate interpretations [10].

Predicting the outcome of upcoming football matches is a common practice among enthusiasts, with these predictions often exhibiting a bias toward the team they favour. Even acknowledged experts in the field recognize the difficult challenge associated with forecasting game outcomes. Forecasting football results has occurred as an attractive theoretical challenge, given its inherent complexities. This issue stems from various factors that could influence the outcomes of football matches, encompassing teamwork, individual capabilities, weather conditions, home advantage, and other variables [11]. Developing a predictive system for football matches holds immense importance beyond academia, as it carries significant economic value [12].

In the context of this study, several objectives guide our exploration of FIFA games. Firstly, the aim is to investigate the correlation between the number of goals scored in FIFA matches and selected explanatory variables. Following this, the study seeks to estimate the expected number of goals in football matches using both Poisson and Quasi-Poisson regression models. Another objective is to conduct a thorough performance comparison between the Poisson and Quasi-Poisson regression models. Ultimately, the study aims to determine the superior model, choosing between Poisson and Quasi-Poisson regression, for predicting the number of goals accurately in FIFA World Cup games. These objectives collectively contribute to the advancement of predictive modelling within the domain of football outcomes.

2. Materials and Methods

2.1 Data sources and data set

This study used an extensive data related to the World Cup 2022. The dataset originated from the Kaggle website, encompassing comprehensive statistics for all matches held during the FIFA World Cup 2022. The dataset comprises complete records of the tournament, encompassing 64 matches with 32 teams, and is

geospatially limited to Qatar. The information, obtained from the official FIFA website, includes match scores, participating teams, and various football statistics. These statistics cover a range of metrics from goal-related aspects such as attempts, assists, and goals inside/outside the penalty area to defensive actions such as conceded goals and completed defensive line breaks. Additionally, the dataset encompasses player positioning strategies (in-behind, in-between, and in-front offers to receive) and team defensive efforts, such as forced turnovers and defensive pressure applied. The data collection, conducted through web scraping, spans from November 20th, 2022, to December 18th, 2022, and the subsequent analysis was carried out by using R Studio.

2.2 Methods

2.2.1 Descriptive analysis

Descriptive analytics entails the examination of both current and past data to recognise patterns and correlations. Through descriptive analysis, one examines a situation or phenomenon, discerns patterns in the data, and offers insights into the who, what, where, when, and to what degree aspects [13]. Descriptive statistics encompass various aspects of data, including the types of variables (nominal, ordinal, interval, and ratio), as well as measures related to frequency, central tendency, dispersion/variation, and position [14].

2.2.2 Spearman's Rank Correlation

Named after Charles Spearman, the Spearman rank correlation coefficient, denoted by rs, is a nonparametric measure for assessing the strength and direction of a monotonic relationship between two variables. It is designed to handle ties between the variables but may be slightly less precise compared to the previously computed Pearson coefficient [15]. The formula for Spearman's rank correlation coefficient is as equation 1,

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

Spearman rank correlation, denoted by ρ, spans a scale from -1 to 1. A ρ value of 1 signifies a perfect monotonic positive relationship, indicating that as one variable increases, the other consistently increases monotonically. Conversely, a ρ value of -1 indicates a perfect monotonic negative relationship, where as one variable increases, the other consistently decreases monotonically. A ρ value of 0 suggests no monotonic relationship between the variables.

2.2.3 Poisson regression

Poisson regression is a statistical technique used to model count data, and it is based on the Poisson distribution, which was first introduced by Poisson in 1837. The Poisson density function is used to calculate the probability of a specific number of events occurring within a fixed interval of time, based on the average rate at which those events typically occur. The formula for the Poisson density function as shown in equation 2 and the mean and variance of Poisson distribution is equal to λ (see equation 3).

$$f\left(y|\lambda\right) = \frac{\lambda^y e^{-\lambda}}{y!},\ y = 0,1,2,\ldots,\infty \tag{2}$$

$$E(x) = V(x) = \lambda \tag{3}$$

To predict the number of goals, the initial step involves constructing a Poisson model. This foundational model serves as the basis for a comprehensive approach. Through an iterative process, the model is refined until a predefined stopping criterion is met, specifically a significance threshold set at a p-value below 0.1. This iterative refinement ensures the development of a robust and reliable predictive model. Ultimately, the final model, shaped through this iterative procedure, is employed to make predictions for the number of goals.

2.2.4 Overdispersion test

Count data examined using a Poisson assumption or proportion data analysed under a binomial assumption frequently display signs of overdispersion [16]. Various statistical tests can be used to formally assess overdispersion. Common tests include the likelihood ratio test, Pearson chi-squared test, and deviance goodness-of-fit test. Statistical software suites like R, coupled with specialized libraries such as qcc, offer comprehensive tools for both testing and mitigating overdispersion in statistical models. The result of the test will include a p-value. If the p-value is less than 0.05, it may suggest evidence of overdispersion, indicating that a Poisson model may not be the best fit for the data.

2.2.5 Quasi-Poisson regression

The Quasi-Poisson distribution serves as an extension of the Poisson distribution, proving valuable when the assumptions of the Poisson model are not fully met. It is commonly employed in Generalized Linear Models (GLM) to analyse count data exhibiting overdispersion, which is the observed variance surpasses expectations based on a Poisson distribution. In practical situations, equidispersion is frequently absent in data. Consequently, the Quasi-Poisson model emerges as a viable alternative to tackle such situations [8].

This model begins by estimating parameters through quasi-likelihood, distinguishing itself from likelihood estimation, which requires the response variable to obey to a specific distribution. Furthermore, the quasi-Poisson regression model assesses the dispersion parameter's value, a consideration that is overlooked in Poisson regression.

$$E(y) = \mu \tag{4}$$

$$V(y) = \varnothing\mu \tag{5}$$

Equation 4 reveals that the mean resembles that of a Poisson distribution, but its variance includes an additional parameter. This extra parameter is the overdispersion parameter, which accounts for additional variation in the response, resulting in a variance function that is now a linear function of the mean.

The extended quasi-Poisson model refers to a quasi-Poisson model that incorporates random effects. [17] defined this model in 1986, and it can be expressed as in equation 6:

$$f(y \mid v) = \varnothing^{-\frac{1}{2}} \exp[-\frac{\mu}{\varnothing}] \frac{\exp[(\frac{1}{\varnothing} - 1)y] y^{y}}{y!} (\frac{\mu}{y})^{\frac{y}{\varnothing}} \tag{6}$$

To forecast the number of goals, the process initiates with the construction of a quasi-Poisson model. This foundational model forms the basis for a comprehensive approach. Through an iterative procedure, adjustments are made to transition the model into a quasi-Poisson framework, continuing until the predefined stopping criterion is achieved—specifically, a significance threshold set at a p-value below 0.1. This iterative refinement ensures the establishment of a robust and dependable predictive model. Ultimately, the finalized quasi-Poisson model, shaped through this iterative approach, is utilized to predict the number of goals.

2.2.6 Selection K-Covariance

In the context of generalised linear models, the AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and deviance (D) are often used as a statistical measure for the evaluation and selection of the best models. A lowest value of an information criterion (IC) of the given form indicates a better fit or model performance [18]. Meanwhile, [19] propose the utilisation of appropriate indicators AIC and BIC to effectively strike a balance between the model's complexity and accuracy (as shown in equations 6, 7 and 8).

$$AIC = -2\ln(likelihood) + 2k \tag{7}$$

$$BIC = -2\ln(likelihood) + k\log(n) \tag{8}$$

$$D = 2[ln(\hat{L}) - ln(L)] \tag{9}$$

Table 1 : Explanation of information criterion

Information Criterion	Explanation
AIC	The AIC value is frequently used as a predictive error estimator.
BIC	The BIC is an adaptation of the AIC. A lower BIC value is indicative of a superior model fit, making it desirable to minimize the BIC value to obtain an optimal model.
Deviance	The D which is a significant indicator that deserves careful examination, is defined as the difference between the log likelihoods of the fitted model and the saturated model.

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 349-358

3. Results and Discussion

3.1 Exploratory Data Analysis

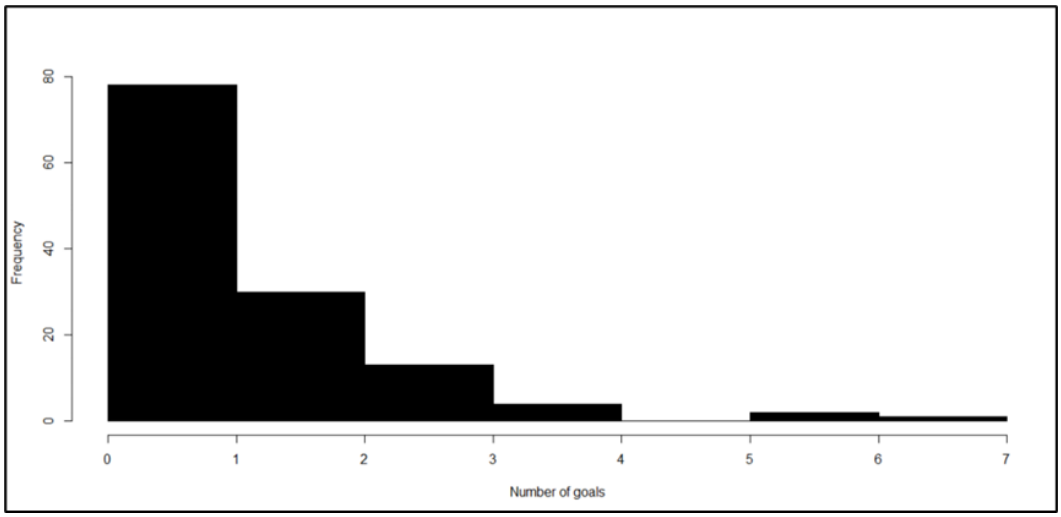

Fig. 1 Histogram depicting the count of goals scored

From Fig. 1, there are 40 occurrences where the number of goals scored was 0. In 38 instances, the teams scored 1 goal. The count decreases to 30 for 2 goals. For 3 goals, there were 13 occurrences. Teams scored 4 goals in 4 instances. A higher number of goals, specifically 6, were recorded in 2 occurrences. England and Portugal are the two countries that have scored 6 goals. England had registered a dominant 6-2 victory over Iran. In another impressive performance, Goncalo Ramos, who replaced Cristiano Ronaldo, scored a hat trick as Portugal advanced to the World Cup quarterfinals with a resounding 6-1 win over Switzerland. Finally, there was 1 instance where 7 goals were scored. Spain is the only country that achieving a score of 7 in a match. Spain began their World Cup journey impressively, securing a loud 7-0 victory against Costa Rica. Ferran Torres scored twice, and Gavi made history in a commanding performance.

The Spearman correlation matrix is a matrix of correlation coefficients that assesses the strength and direction of monotonic relationships between variables. This matrix proves particularly useful in analysing data with heavy-tailed distributions.

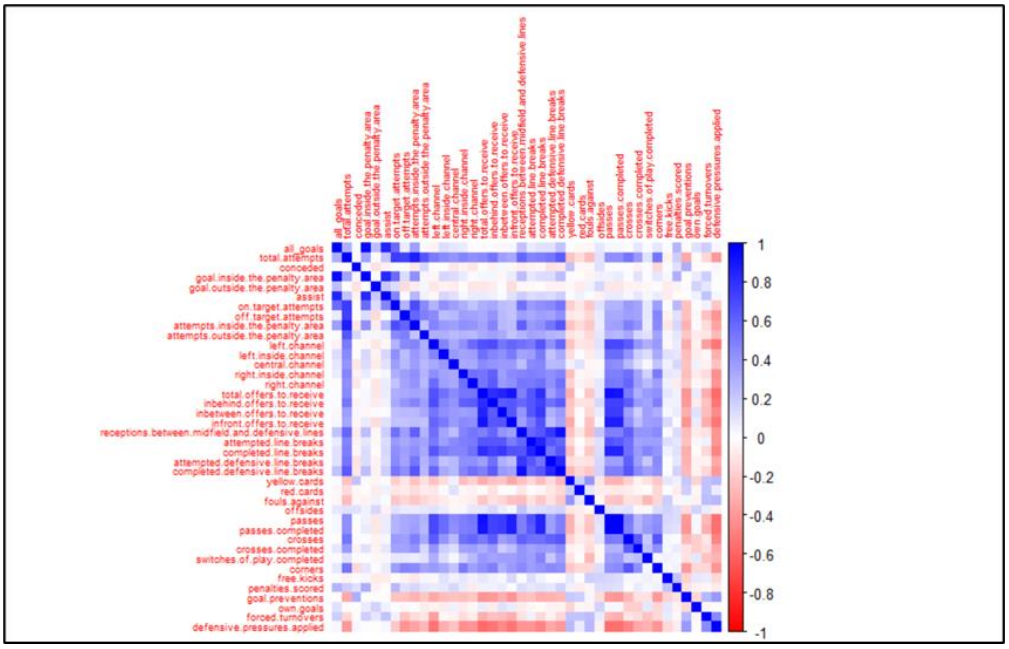

Fig. 2 Spearman correlation matrix of all variables

Fig. 2 shows the spearman correction matrix of all variables where the darker colours represent stronger relationship among them. A correlation coefficient of 0.9545 for "goal inside the penalty area" suggests an even stronger positive correlation. This implies that there is a high positive association between the number of goals scored and the goals that were scored from inside the penalty area. Scoring goals from inside the penalty area is

frequently associated with valuable goal-scoring opportunities. When teams take shots from close range, they increase the likelihood of scoring due to the shorter distance to the goal. The fact that a team scores goals from within the penalty area often indicates that they are creating high-quality scoring chances. These opportunities, originating from close proximity to the goal, are known for their superior likelihood of resulting in successful goals. For "assist", a correlation coefficient of 0.8409 indicates a strong positive correlation. As the number of assists increases, the number of goals also tends to increase. A positive correlation implies that as assist increases, the number of goals scored also tends to increase. Players who consistently contribute assists typically play a pivotal playmaking role within the team. Their skill lies in creating opportunities for their teammates to score goals, potentially leading to an increased overall goal count. The presence of a substantial number of assists may suggest efficient team coordination and communication. Players who excel in connecting with their teammates to establish scoring opportunities significantly enhance the team's overall effectiveness in scoring goals.

3.2 Prediction number of goals

3.2.1 Poisson regression

Prior to making predictions, the initial step involves constructing the Poisson regression model. The methodology chosen for this purpose is backward elimination, aiming to identify the most suitable model. The backward elimination analysis is utilised to eliminate variables that do not significantly contribute to the model. This iterative process continues until a predetermined stopping criterion is satisfied.

A model encompassing all potential predictors is initiated. The regression analysis is performed, and the coefficients, p-values, and relevant statistics are noted. This criterion will be based on a significance threshold which is *p*-value less than 0.1. After undergoing 34 iterations, the final model is displayed in equation 9.

$$\log(\lambda) = -0.2913 + 0.5302x_1 + 0.8559x_2 - 0.0445x_3 + 0.0717x_4 + 0.1021x_5 - 0.0012x_6 \quad (9)$$

A one-unit increase in "goal.inside.the.penalty.area" is associated with a 0.5302 increase in the log number of goals. Similarly, a one-unit increase in "goal.outside.the.penalty.area" is associated with a 0.8559 increase in the log number of goals. A one-unit increase in " attempted.defensive.line.breaks " is associated with a decrease of 0.0445 in the log number of goals. A one-unit increase in "completed.defensive.line.breaks" is associated with a 0.0717 increase in the log number of goals. The coefficient is 0.1021, suggesting that a one-unit increase in "yellow.cards" is associated with a 0.1021 increase in the log number of goals. A one-unit increase in "passes" is associated with a decrease of 0.0012 in the log count of goals.

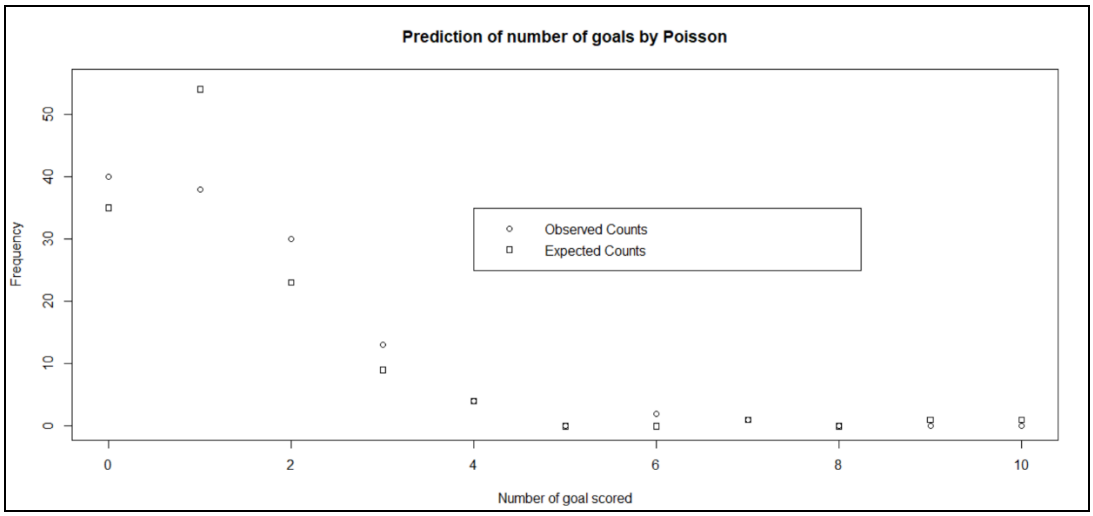

Fig. 3 Prediction of number of goals by Poisson regression

Fig. 3 exhibits the observed and expected counts utilizing the Poisson model. Circles denote observed counts, while squares represent expected counts. Across the categories of obtaining 0, 2, 3 and 6 goals, the observed count exceeds the expected counts, except in the category of achieving 1 goal where the expected counts surpass the observed counts. Categories representing 4, 5, 7, and 8 goals depict overlapping circles and squares in the graph. Remarkably, the graph illustrates a close correspondence between observed and expected counts for the number of goals in the categories of achieving 9 and 10 goals.

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 2 (2024) p. 349-358

3.2.2 Overdispersion test

Overdispersion can lead to underestimation of standard errors and incorrect inferences. Resolving overdispersion ensures more reliable and accurate statistical analysis. Statistical software packages like R, along with dedicated libraries such as qcc, provide extensive tools for testing and addressing overdispersion in statistical models. The outcome of the test will encompass a p-value. The *p*-value associated with the test is 0.0062092. A small *p*-value indicates evidence against the null hypothesis of no overdispersion. Therefore, the assumption of equidispersion (equal variance and mean) in a Poisson distribution may be violated.

3.2.3 Quasi-Poisson regression

The initial step in this process involves constructing the Quasi-Poisson regression model, with a chosen methodology of backward elimination aimed at identifying the most suitable model. This iterative process continues until the predefined stopping criterion is satisfied, determined by a significance threshold of a *p*-value less than 0.1. After 30 iterations of this backward elimination process, the final model is presents in equation 10, with the selected predictors, their coefficients, and associated statistics.

$$\begin{aligned}\log(\mu) = &-0.5412 + 0.5785x_1 + 0.8993x_2 + 0.0506x_3 + 0.0187x_4 - 0.0385x_5 \\ &+ 0.0567x_6 + 0.0987x_7 - 0.0017x_8 + 0.6752x_9\end{aligned} \tag{10}$$

The provided output is from a Quasi-Poisson regression model, which is a type of generalized linear model suitable for count data with overdispersion. A one-unit increase in "goal.inside.the.penalty.area" and "goal.outside.the.penalty.area" is associated with a 0.5786 and 0.8993 increase in the log number of goals. Similarly, a one-unit increase in "left.inside.channel" is associated with an increase of 0.0506 in the log number of goals. A one-unit increase in the "right.channel" predictor is associated with a 0.0187 increase in the log number of goals. An increase in "attempted.defensive.line.breaks" by one unit is associated with a decrease of 0.0385 in the log number of goals. Also, an increase in "completed.defensive.line.breaks" by one unit is associated with a 0.0567 increase in the log number of goals. A rise in the "yellow.cards" variable by one unit correlates with a 0.0987 increment in the log number of goals. Conversely, a one-unit increase in the "passes" variable is linked to a reduction of 0.0017 in the log number of goals. Similarly, an increase of one unit in the "own.goals" variable is connected to a 0.6752 rise in the log number of goals.

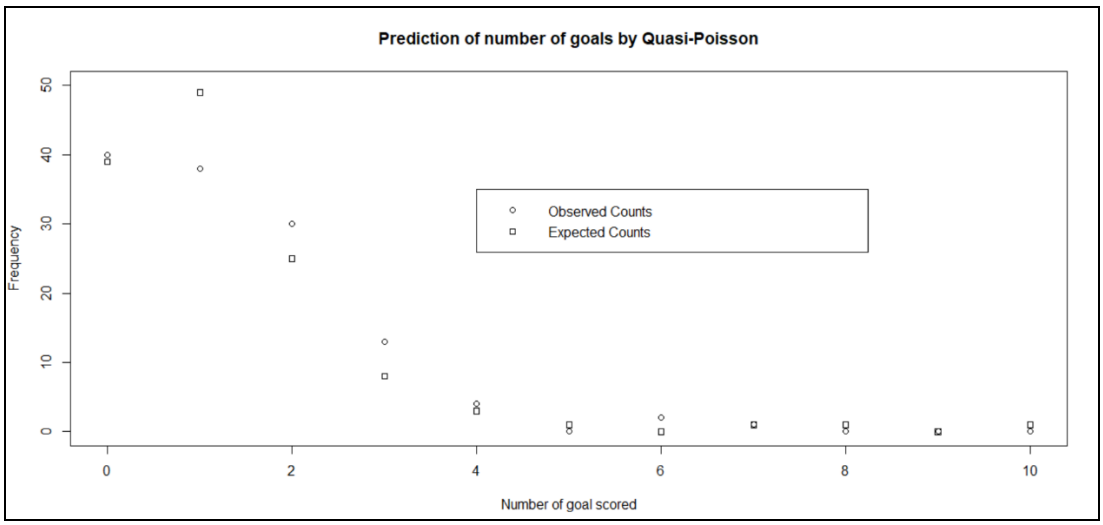

Fig. 4 Prediction of number of goals by Quasi-Poisson regression

Fig. 4 exhibits the observed and expected counts utilizing the Poisson model. Circles denote observed counts, while squares represent expected counts. In the categories corresponding to scoring 0 to 4 goals, the observed count surpasses the expected counts, except in the case of the number of goals in category 1, where the expected counts are higher than the observed counts. Notably, the graph demonstrates a close alignment between observed and expected counts for the number of goals in categories 4, 5, 8, and 10. For the categories representing 7 and 9 goals, there is an overlap between the observed count and the expected count, indicating an equality in counts.

3.3 Comparison of Poisson regression and Quasi-Poisson regression

Table 2 : Explanation of information criterion

Regression model	AIC	BIC	Deviance
Poisson regression	281.1933	301.1575	49.7028
Quasi-Poisson regression	-50.0853	347.7055	38.7935

The AIC values for the Poisson and Quasi-Poisson models are 281.1933 and -50.08531, respectively. In the context of AIC, lower values indicate a better balance between goodness of fit and model complexity. The Quasi-Poisson model, with its considerably lower AIC, suggests a more suitable yet effective representation of the data. The BIC values for the Poisson and Quasi-Poisson models are 301.1575 and 347.7055, respectively. BIC penalizes model complexity more strongly than AIC. While the Quasi-Poisson model has a higher BIC, suggesting increased complexity, it is essential to balance this against the other metrics. The deviance values for the Poisson and Quasi-Poisson models are 49.70277 and 38.79347, respectively. Deviance measures the lack of fit, and lower values indicate better model fit. In this case, the Quasi-Poisson model exhibits a lower deviance, suggesting a better overall fit to the data.

Given that the quasi-Poisson model exhibits lower AIC and deviance values than the Poisson model, despite having a higher BIC value, the assessment of standard errors and *p*-values is employed to determine the most suitable model.

Table 3 : Standard error and p-value of compared models

Variables	Standard error		*p*- value	
	Poisson	Quasi-Poisson	Poisson	Quasi-Poisson
Intercept	0.3096	0.1806	> 0.1	< 0.01
goal.inside.the.penalty.area	0.0501	0.0299	< 0.001	< 0.001
goal.outside.the.penalty.area	0.2213	0.1193	< 0.001	< 0.001
left.inside.channel	N/A	0.0164	N/A	< 0.001
right.channel	N/A	0.0093	N/A	< 0.05
attempted.defensive.line.breaks	0.0255	0.0134	< 0.1	< 0.01
completed.defensive.line.breaks	0.0339	0.0186	< 0.05	< 0.01
yellow.cards	0.0489	0.0256	< 0.05	< 0.001
passes	0.0006	0.0003	< 0.05	< 0.001
own.goals	N/A	0.2770	N/A	< 0.05

Table 3 displays the standard errors and *p*-values for the Poisson and Quasi-Poisson regression models, with "N/A" indicating variables that were not selected in the Poisson regression model. In comparison to the Poisson regression model, the Quasi-Poisson regression model consistently shows lower standard errors and lower p-values for all variables. In summary, the Quasi-Poisson model generally exhibits lower standard errors and comparable or lower *p*-values for the selected variables, suggesting improved precision and significance compared to the Poisson model.

3.4 Optimal model

The optimal model is derived from the quasi-Poisson regression as shown in equation 11.

$$\begin{aligned}\log(all\_goals) = {} & -0.5412 + 0.5785(\text{goal.inside.the.penalty.area}) \\ & +0.8993(\text{goal.outside.the.penalty.area}) \\ & +0.0506(\text{left.inside.channel}) + 0.0187(\text{right.channel}) \\ & -0.0385(\text{attempted.defensive.line.breaks}) \\ & +0.0567(\text{completed.defensive.line.breaks}) \\ & +0.0987(\text{yellow.cards}) - 0.0017(\text{passes}) + 0.6752(\text{own.goals})\end{aligned} \tag{11}$$

The predictors such as "goal.inside.the.penalty.area," "goal.outside.the.penalty.area," "left.inside.channel," "right.channel," "attempted.defensive.line.breaks," "completed.defensive.line.breaks," "yellow.cards," "passes," and "own.goals" were identified as variables frequently linked to a higher likelihood of number of goals.

After thorough analysis and consideration of various regression models, it has been determined that the most effective and fitting model is obtained through the utilization of the quasi-Poisson regression. Based on the research findings, a parallel outcome was observed with [18]. In situations involving over-dispersed data, it was noted that quasi-Poisson regression exhibited a better fit than Poisson regression.

4. Conclusion

In conclusion, this study successfully compared Poisson and quasi-Poisson regression analyses in the context of FIFA World Cup games, achieving four primary objectives. The research utilised a comprehensive dataset from the 2022 FIFA World Cup and identified influential factors affecting the number of goals scored.

In the quasi-Poisson model, the expected number of goals closely corresponds to the observed count, demonstrating improvements in matching observed and expected counts compared to the Poisson regression. The quasi-Poisson model accommodates overdispersion, and this is evident in the adjustments made to the expected counts. both models provide reasonable fits to the data, but the quasi-Poisson model may capture the variability in the data more accurately, especially in cases where overdispersion is present.

The comparison of models using AIC, BIC, and deviance metrics favoured the quasi-Poisson model, demonstrating superior performance in terms of model fit and complexity. Despite a higher BIC, the quasi-Poisson model consistently exhibited lower standard errors and p-values, indicating enhanced accuracy and significance compared to the Poisson model. The quasi-Poisson model exhibits a superior fit as evidenced by its lower AIC value of -50.0853 and a reduced deviance value of 38.7935 compared to alternative models. Ultimately, this research contributes valuable insights to the football field, recommending the quasi-Poisson regression as the preferred model, especially in scenarios involving overdispersed data, for more accurate predictions of goal outcomes in FIFA World Cup games.

While acknowledging the model's limitations and the challenges associated with creating the best goal prediction model, there is optimism that this football prediction model represents a stride toward establishing a valuable decision-making tool for Malaysia football fields. It is hoped that a suitable model can be constructed to yield precise outcomes in future football field predictions.

Acknowledgement

The authors would like to thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussien Onn Malaysia for the support and resources provided during the final year project.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Kang Yue Teng, Norziha Che Him;* ***data collection:*** *Kang Yue Teng;* ***analysis and interpretation of results:*** *Kang Yue Teng, Norziha Che Him;* ***draft manuscript preparation:*** *Kang Yue Teng, Norziha Che Him. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Statistics & Data. (2020, October 14). Most popular sports in the world - (1930/2020) -. https://statisticsanddata.org/most-popular-sports-in-the-world/
[2] Kirkendall, D. T. (2020). Evolution of soccer as a research topic. Progress in Cardiovascular Diseases, 63(6), 723–729.
[3] Nguyen, Q. V. (2021). Poisson Modeling and Predicting English Premier League Goal Scoring. The New England Journal of Statistics in Data Science, 1–7.
[4] O'Leary, D. E. (2017). Crowd performance in prediction of the World Cup 2014. European Journal of Operational Research, 260(2), 715–724.
[5] Pielke, R. A. (2013). How can FIFA be held accountable? Sport Management Review, 16(3), 255–267.
[6] Fialho, G. G., Manhães, A., & Teixeira, J. A. (2019). Predicting Sports Results with Artificial Intelligence – A Proposal Framework for Soccer Games. Procedia Computer Science, 164, 131–136.
[7] Inan, T. (2020). Using poisson model for goal prediction in European football. Journal of Human Sport and Exercise.

[8] Hartono, P. G., Tinungki, G. M., Jakaria, J., Hartono, A. B., Hartono, P. G., & Wijaya, R. (2021). Overcoming overdispersion on direct mathematics learning model using the quasi poisson regression. Advances in Social Science, Education and Humanities Research.
[9] Edmondson, M., Luo, C., Islam, M. N., Sheils, N. E., Buresh, J., Chen, Z., Bian, J., & Chen, Y. (2022). Distributed Quasi-Poisson regression algorithm for modeling multi-site count outcomes in distributed data networks. Journal of Biomedical Informatics, 131, 104097.
[10] Zhang, X., Kano, M., Tani, M., Mori, J., Ise, J., & Harada, K. (2020). Prediction and causal analysis of defects in steel products: Handling nonnegative and highly overdispersed count data. Control Engineering Practice, 95, 104258.
[11] Razali, N., Mustapha, A., Yatim, F. A., & Aziz, R. A. (2017). Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). IOP Conference Series, 226, 012099.
[12] Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741–755.
[13] Loeb, S. (2017, March). Descriptive Analysis in Education: A Guide for Researchers. NCEE 2017-4023.
[14] Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). Descriptive statistics. International Journal of Academic Medicine, 4(1), 60.
[15] Lemenkova, P. (2018). R scripting libraries for comparative analysis of the correlation methods to identify factors affecting Mariana Trench formation. arXiv (Cornell University).
[16] Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. Journal of the American Statistical Association, 87(418), 451.
[17] Efron, B. (1986). Double exponential families and their use in generalized linear models. Journal of the American Statistical Association, 81, 709–721.
[18] Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. Briefings in Bioinformatics, 21(2), 553–565.
[19] Zhang, Y., & Meng, G. (2023). Simulation of an Adaptive Model Based on AIC and BIC ARIMA Predictions. Journal of Physics, 2449(1), 012027.
[20] Walker, R. (2022, December 7). World Cup 2022 - Portugal 6-1 Switzerland: Goncalo Ramos nets hat-trick as dropped Cristiano Ronaldo watches on. Sky Sports.