ISSN: 3006-7251(Online)
MBSTU Journal of Science and Technology
DOI: https://doi.org/10.69728/jst.v11.60
Journal Homepage: https://journal.mbstu.ac.bd/index.php/jst

Enhancing Predictive Accuracy in Forecasting Football World Cup 2022 Outcomes

Umme Habiba[1,2*]
[1] School of Mathematical and Statistical Science with Interdisciplinary Applications, The University of Texas Rio Grande Valley, Texas, USA
[2] Department of Mathematics, Mawlana Bhashani Science and Technology University, Bangladesh

ARTICLE INFO

Article History
Submission: 20 January, 2025
Revision: 25 May, 2025
Accepted: 01, June 2025
Published: 30 June, 2025

Keywords
Simple Linear Regression Model, Multiple Linear Regression Model, Pearson Regression Model, Log-transformation, Football forecast 2022, Soccer Power Index

ABSTRACT

This study evaluates key predictive metrics for the 2022 FIFA World Cup outcomes, comparing the predictive accuracy of the simple linear regression (SLR) and multiple linear regression (MLR) models. The analysis examines model fit and accuracy using statistical metrics and addresses key assumptions such as homoscedasticity, autocorrelation, and multicollinearity. Significant variables include the Soccer Power Index (SPI), offensive and defensive strengths, and simulated goal differences. To improve model assumptions, log-transformation is employed for the dependent variable. The findings demonstrate that MLR models outperform SLR in predictive accuracy, contributing to advances in sports forecasting methodologies. Recommendations are also provided for future sports analytics applications.

1. Introduction

Accurate sports forecasting requires analyzing key metrics that influence game outcomes, such as offensive and defensive ratings, team performance indices, and match scenarios. In this project, predictive models are applied to forecast goals scored during the 2022 FIFA World Cup. Using regression analysis, the study evaluates the predictive capabilities of the simple linear regression (SLR) and multiple linear regression (MLR) models. The analysis focuses on: Predictor Variables: spi (Soccer Power Index): Higher SPI values indicate stronger teams are likely to score more goals. Global Offensive Strength (global o): Represents a team's scoring potential. Global Defensive Strength (global d): Reflects a team's ability to prevent goals, inversely impacting scores. Simulated Goal Difference (sim goal diff): Captures expected outcomes based on forecast scenarios. Outcome Variable: Goals scored by each team in a match. Incorporating team-specific performance metrics, such as offensive and defensive strengths, significantly enhances the predictive accuracy of regression models for sports outcomes (Jingzheng, 2023).

To improve the assumptions of the model and address the deviations in residual normality, a log-transformation of the dependent variable (goals scored) is applied. This transformation helps mitigate the skewness in the data and ensures better adherence to regression assumptions. This project aims to compare the predictive accuracy of SLR and MLR models, validate statistical assumptions, and explore the implications of these models for sports analytics. Király and Qian (2017) introduced a structured log-odds modeling approach that unifies the Bradley-Terry and Elo models, facilitating supervised probabilistic prediction of sports outcomes with applications in both batch and online learning settings. Maher (1982) proposed a bivariate Poisson model to predict football match results, accounting for the number of goals scored by each team and incorporating factors such as team attack and defense strengths.

By leveraging the log-transformation and identifying effective predictors, the study contributes to the advancement of sports forecasting methodologies, offering practical insights for future applications in tournaments and other sports domains. Machine learning models and historical player data have also been applied to predict football match outcomes, analyzing the impact of individual player statistics on final scores and demonstrating the profitability of such models in betting scenarios (Peters & Pacheco, 2022).

*Corresponding author: habiba.math@mbstu.ac.bd

2. Data Collection and Preparation

The dataset 'wc forecasts.csv' is used, containing multiple predictive metrics and team statistics. The dataset contains a total of 256 observations, each representing team-level statistics for individual matches in the 2022 FIFA World Cup forecasts. Predictors: spi: Soccer Power Index, representing the overall strength of a team. global o: Offensive strength metric. global d: Defensive strength metric. sim wins, sim ties, and sim losses: Simulated probabilities for match outcomes based on historical data. Outcome Variable: goals scored: Number of goals scored by each team in a game. Data preprocessing included loading the data and confirming the variables' suitability for regression analysis.

3. Methodology

In this section, two linear regression models are explained to fit the model to obtain the accuracy of the models. Also, log-transformation is used to get better regression assumptions.

A Pearson correlation heatmap is shown to visualize the multicollinearity.

3.1. Model Objectives and Variable Descriptions

The objective of this analysis is to evaluate and compare the predictive performance of various linear regression models in forecasting the number of goals scored in FIFA World Cup 2022 matches. The study employs both Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) frameworks, along with a log-transformed variant, to address deviations from regression assumptions.

The outcome (dependent) variable is

• goals scored: Number of goals scored by a team in a match.

The predictor (independent) variables are defined as follows:

• spi: Soccer Power Index — a composite score representing the overall strength of a team. Higher values indicate stronger teams.

• global o: Global offensive strength — a numerical indicator of a team's ability to

score goals.

• global d: Global defensive strength — a measure of a team's ability to prevent goals; lower values are indicative of better defensive performance.

• sim goal diff: Simulated goal difference — the expected net goal margin derived from simulation models for each match.

These variables are selected based on their statistical relevance and contextual importance in football analytics. All regression models were estimated using the ordinary least squares (OLS) method and evaluated for model fit and assumption validity in the subsequent sections. The models capture differences based on forecasted values

3.2. Simple Linear Regression (Jingzheng, 2023)

The simple linear regression equation can be represented as

goals scored = $\beta_0 + \beta_1 \cdot$ spi + ε

where β_0 = intercept, β_1 = coefficient of spi, ε = error term

This model serves as a baseline to compare with more complex models.

3.3. Multiple Linear Regression (Wang *et al.*, 2023)

The multiple linear regression equation can be represented as:

goals scored = $\beta_0 + \beta_1 \cdot$ sim goal diff + $\beta_2 \cdot$ global o + $\beta_3 \cdot$ global d + ε

where β_0 = intercept, β_1, β_2, β_3 = coefficients of sim goal diff, global o, and global d, ε = error term

All models are fit using ordinary least squares (OLS) estimation. Coefficients are interpreted, and statistical significance is assessed via p-values.

3.4. Log-Transformation (Simon *et al.*, 2021)

The dependent variable goals scored is positively skewed, which justifies the potential use of a log transformation to improve normality and stabilize the data for regression analysis.

log(goals scored) = $\beta_0 + \beta_1 \cdot$ sim goal diff + $\beta_2 \cdot$ global o + $\beta_3 \cdot$ global d + ε

where β_0 = intercept; β_1, β_2, β_3 = coefficients of sim goal diff, global o, and global d; ε = error term

This transformation stabilizes variance and improves model fit with respect to regression assumptions.

3.5. Pearson Regression (Dufera *et al.*, 2023)

To obtain correlation among the predictor variables and response variables. We fit this model in R code and we obtain the results.

Each model is fit to the data, and coefficients, significance levels, and residuals are analyzed to assess predictive power.

4. Real data applications

In this section, wc forecasts.csv data is analyzed for World Cup 2022.

4.1. Data Set Description

In this section, we obtain the five-number summary for real dataset variables.

Table 1: Five-Number Summary

Variable	Min	25%	Median	75%	Max
spi	48.16	67.78	74.07	82.86	93.66
global o	1.28	1.72	1.96	2.50	3.43
global d	0.24	0.55	0.65	0.80	1.32
sim wins	0.00	0.82	1.00	2.00	2.81
sim ties	0.00	0.00	0.84	1.00	2.32
sim losses	0.00	0.87	1.00	1.61	3.00
sim goal diff	-10.09	-2.00	0.00	2.00	8.54
goals scored	0.60	2.03	3.87	5.00	10.61
goals against	0.31	2.00	3.00	4.76	11.39

The five-number summary is used to describe the distribution of a dataset in a clear, concise manner. It provides insight into the data's spread, center, and variability using five key statistics:

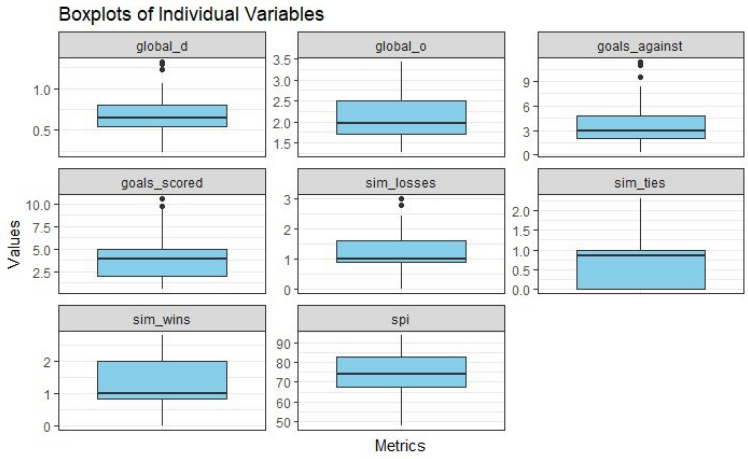

Figure 1: Five number summary

• Minimum: The smallest value in the dataset, showing the lower extreme. First Quartile

• (Q1): The 25th percentile, which indicates the value below which 25% of the data falls.

• Median: The middle value (50th percentile), showing the central tendency. Third Quartile

• (Q3): The 75th percentile, which indicates the value below which 75% of the data falls. Maximum: The largest value in the dataset, showing the upper extreme.

• Predictor Variability spi: Moderate spread (range: 45.5). global o: High concentration in lower values (Q3 = 2.50, Max = 3.43). global d: Smaller range compared to spi and global o. sim goal diff (2.00): The top-performing 25% of teams in simulations have a positive goal difference.

• Regression Implications: Greater variability in predictors (like spi) typically results in more reliable coefficient estimates.

4.2. Simple Linear Regression Model: Results Interpretation

The simple linear regression model is fitted using goals scored as the dependent variable and spi as the predictor variable. The key results are interpreted as follows:

• Residuals: The residuals range from -3.4686 to 5.0234, with a median close to zero (0.0704). This suggests a balanced distribution of residuals around the predicted values, indicating no major skewness or bias.

4.2.1. Coefficients and their Interpretation

• Intercept (-4.2804): When the predictor spi is zero, the model predicts an average goals scored of -4.2804. While this value lacks practical meaning (since goals cannot be negative), it is a necessary component of the regression equation. Coefficient of spi (0.1081): For every one-unit increase in spi, the predicted goals scored increase by 0.1081. The extremely low p-value ($p < 2.2\times10-16$) indicates that spi is a statistically significant predictor.

• Model Summary: Residual Standard Error (1.61): The residual standard error suggests that, on average, the actual values of goals scored deviate from the model's predictions by approximately 1.61 goals. R-squared (0.3496): About 34.96% of the variation in goals scored is explained by the predictor spi. Adjusted R-squared (0.347): After accounting for the number of predictors, 34.7% of the variation in goals scored is explained by the model, confirming the predictor's meaningful contribution. F-statistic (136.5, p-value $< 2.2\times10-16$):

The overall model is statistically significant, meaning the relationship between spi and goals scored is unlikely to have occurred by chance.
The predictor spi is statistically significant and positively associated with goals scored, but the R^2 value indicates that other factors not included in the model explain a majority of the variability in goals scored.

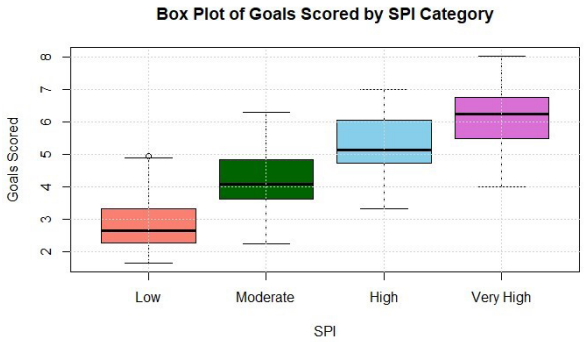

Figure 2: Relationship between spi and goals scored

The plot shows a positive correlation, indicating that higher spi values are associated with an increased number of goals scored.

4.3. Multiple Linear Regression Model: Results Interpretation

The multiple linear regression model is fitted using goals scored as the dependent variable and sim goal diff, global o, and global d as predictor variables. The key results are interpreted as follows:

• Residuals: The residuals range from -2.63632 to 2.93339, with a median close to zero (-0.07032). This suggests a balanced distribution of residuals around the predicted values, indicating no major skewness or bias.

• Coefficients and their Interpretation: Intercept (-2.69414): When all predictors (sim goal diff, global o, and global d) are zero, the model predicts a baseline goals scored of -2.69414. Although this value lacks direct practical meaning, it is essential for the model. sim goal diff (0.51590): For every one-unit increase in sim goal diff, the predicted goals scored increase by 0.51590, holding other predictors constant. The extremely low p-value (< 2e-16) indicates that sim goal diff is highly statistically significant in the model. global o (1.56403): For every one-unit increase in global o, the predicted goals scored increase by 1.56403, holding other predictors constant. The p-value (1.63e-10) indicates that global o is also highly statistically significant. global d (4.65622): For every one-unit increase in global d, the predicted goals scored increase by 4.65622, holding other predictors constant. The very low p-value (1.70e-10) shows that global d is a significant predictor.

• Model Summary: Residual Standard Error (1.24): The residual standard error indicates that, on average, the predicted goals scored deviate from the actual values by approximately 1.24 goals. R-squared (0.6175): About 61.75% of the variation in goals scored is explained by the predictors included in the model. Adjusted R-squared (0.613): After adjusting for the number of predictors, the model explains 61.3% of the variation in goals scored , indicating a strong fit without over fitting. F-statistic (135.6, p-value $< 2.2 \times 10-16$): The overall model is statistically significant, with the low p-value indicating that at least one of the predictors contributes significantly to explaining goals scored. The predictors sim goal diff, global o, and global d are all statistically significant contributors to the model, with sim goal diff having the highest relative impact. The model explains a substantial proportion of the variation in goals scored, suggesting its usefulness for prediction in this context.

5. Comparison

In this section, comparisons between two linear regression models are shown.
The comparison between the Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) models highlights key differences in their performance. The MLR model explains a significantly higher proportion of the variability in goals scored ($R^2 = 0.6175$) compared to the SLR model ($R^2 = 0.3496$). The residual standard error is lower in the MLR Table 2: Comparison between two models.

Table 2: Comparison between two models.

	SLR Model	**MLR Model**
Multiple R-squared:	0.3496	0.6175
Adjusted R-squared:	0.347	0.613
Residual standard error:	1.61	1.24
F-statistic:	136.5	135.6
p-value:	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$

Model (1.24) than in the SLR model (1.61), indicating better prediction accuracy. Both models are statistically significant with p-value $< 2.2 \times 10^{-16}$. Overall, the MLR model performs better by incorporating additional predictors, leading to improved explanatory power and precision. The box plot below displays a comparison of the residuals for two models.

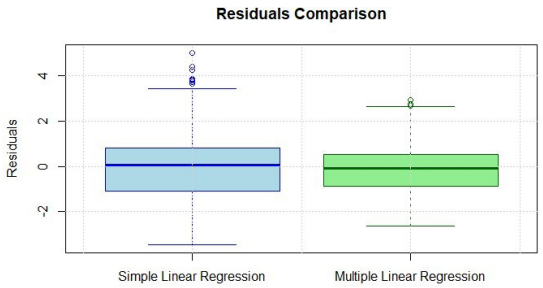

Figure 3: Residual Comparison

The boxplot compares the residuals from two regression models:

• Simple Linear Regression: The residuals exhibit

greater variability with more extreme outliers (above 4).

• Multiple Linear Regression: The residuals have a smaller range, indicating better model fit and reduced variability.

The central tendency (median) of the residuals is closer to zero in both models, but the multiple regression model shows fewer extreme values. The multiple linear regression model demonstrates improved performance by reducing residual variability and minimizing the presence of outliers compared to the simple linear regression model.

6. Regression Diagnostics

Regression diagnostics (Fox, 2015) are used in this section to ensure the validity and robustness of the regression models. The following regression diagnostics (Sections 6.1 to 6.4) are based on the Multiple Linear Regression (MLR) model described in Section 4.3. This model uses goals scored as the dependent variable and includes sim goal diff, global o, and global d as predictor variables.

6.1. Breusch-Pagan Test for Heteroscedasticity

The Breusch-Pagan test checks for heteroscedasticity (non-constant variance) in the residuals. The results are

$$BP = 1.0505, \quad df = 3, \quad p\text{-value} = 0.789$$

The null hypothesis (H_0) of this test states that the variance of the residuals is constant (no heteroscedasticity). Since the p-value (0.789) is greater than 0.05, we fail to reject H_0. No evidence of heteroscedasticity is detected, suggesting that the assumption of constant variance is satisfied.

6.2. Durbin-Watson Test for Autocorrelation

The Durbin-Watson test examines the presence of autocorrelation (correlation between residuals). The results are

$$DW = 2.3161, \quad p\text{-value} = 0.9938$$

The null hypothesis (H_0) assumes that there is no autocorrelation in the residuals. The p-value (0.9938) is much greater than 0.05, indicating no evidence of autocorrelation. Additionally, a Durbin-Watson statistic (DW) close to 2 suggests no significant positive or negative autocorrelation. No autocorrelation is detected in the residuals.

6.3. Mean of Residuals

The mean of the residuals is calculated as

$$\text{Mean of Residuals} = 3.596332 \times 10^{-17}$$

For a properly specified linear regression model, the mean of the residuals should be approximately zero. The result confirms that the residuals have a mean close to zero, satisfying this assumption of the regression model.

6.4. Shapiro-Wilk Test for Normality of Residuals

The Shapiro-Wilk test assesses whether the residuals follow a normal distribution. The results are

$$W = 0.9615, \quad p\text{-value} = 2.374 \times 10^{-6}$$

The null hypothesis (H0) states that the residuals are normally distributed. Since the p-value (2.374×10^{-6}) is less than 0.05, we reject H0, indicating that the residuals deviate significantly from normality. The residuals are not normally distributed, which may affect the validity of hypothesis tests and confidence intervals. The error assumptions of the model can be seen in the table below.

Table 3: Error Assumptions

Assumption	Result	Conclusion
$E[\varepsilon i] = 0$	Mean = 3.6×10^{-17}	Holds
$Var[\varepsilon i] = \sigma^2$	p = 0.789	Holds (No heteroscedasticity)
$Cov[\varepsilon i, \varepsilon j] = 0$	p = 0.9938	Holds (No autocorrelation)
$\varepsilon i \sim N(0, \sigma^2)$	p = 2.374×10^{-6}	Violated (Residuals are not normal)

6.5. Bootstrap Analysis for MLR Coefficients

A bootstrap analysis (Efron & Tibshirani, 1994) with 1,000 resamples is conducted on the multiple linear regression (MLR) model. The original and bootstrap estimates of the coefficients are presented in the table below. The bootstrap results closely align with the original estimates, confirming the stability of the model.

Table 4: Bootstrap Analysis for MLR Coefficients

Coefficient	Original Estimate	Bootstrap Mean	Bootstrap SE
(Intercept)	-2.6941	-2.7083	0.7023
sim goal diff	0.5159	0.5152	0.0563
global o	1.5640	1.5692	0.1969
global d	4.6526	4.6564	0.6927

The model satisfies the assumptions of no heteroscedasticity (Breusch-Pagan test) and no autocorrelation (Durbin-Watson test). The residuals have a mean close to zero, as expected. However, the residuals deviate from normality, as shown by the Shapiro-Wilk test. This issue may need to be addressed, possibly through data transformation (e.g., log transformation) or robust statistical methods.

7. Log-transformations

By transforming the data, the residuals of the regression

model better adhere to the assumption of normality, which is critical for valid statistical inference. The regression model is fitted with the log-transformed dependent variable goals scored log and predictors sim goal diff, global o, and global d.

• Residual Standard Error: 0.2806 (on 252 degrees of freedom).

• Multiple R^2: 0.5659, Adjusted R^2: 0.5607.

• F-statistic: 109.5 (on 3 and 252 DF, $p < 2.2 \times 10^{-16}$).

Significant Predictors

The model is statistically significant and explains 56.6% of the variance in the log-transformed dependent variable. All predictors are highly significant and positively influence the outcome variable.

Table 5: Regression Coefficients and their Significance

Predictor	Estimate	p-value
sim goal diff	0.10632	2×10^{-16}
global o	0.33209	1.7×10^{-9}
global d	1.02798	4.25×10^{-10}

7.1. Shapiro-Wilk Normality Test (After Log Transformation)

The Shapiro-Wilk test was performed on the residuals of the log-transformed model. The

results are:

$$W = 0.97632, \quad \text{p-value} = 0.0002872$$

The null hypothesis (H_0) states that the residuals are normally distributed. Since the p-value (0.0002872) is less than 0.05, we reject H_0, indicating that the residuals still deviate significantly from normality, even after applying the log transformation. The log transformation improved the normality slightly (as W is closer to 1 compared to the original model).

8. Final goals scored vs. Actual Goals Scored

Here, the scattered plot is used to evaluate the accuracy of the regression model by comparing the predicted goals scored (y-axis) against the actual goals scored (x-axis).The red line represents the line of perfect prediction, where predicted values would equal actual values. Points close to the red line indicate that the model's predictions are accurate, as the predicted values are close to the actual values.

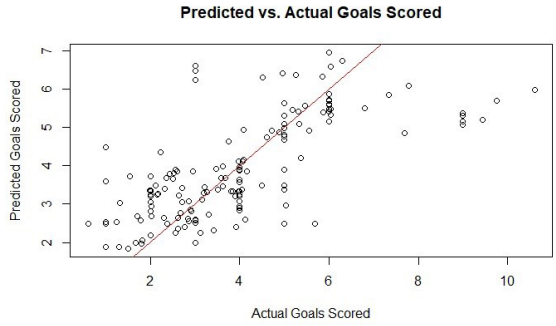

Figure 4: Scattered plot

Points far from the red line indicate prediction errors, with the vertical distance from the line showing the size of the error. In this plot, there is some clustering around the red line, especially for lower goal counts (around 2 to 5), suggesting that the model is relatively accurate in predicting lower to mid-range goal values. However, as actual goals increase, predictions become less accurate, with several points far from the line, indicating larger prediction errors for higher actual goal counts.

9. Multicollinearity

Variance Inflation Factor (VIF) (O'Brien, 2007) is used in the section to detect and address multicollinearity among predictors in the regression models.

Interpreting Variance Inflation Factor (VIF):

• VIF < 5: No serious multicollinearity.

• VIF 5–10: Moderate multicollinearity; consider addressing it.

• VIF > 10: High multicollinearity; corrective measures are needed.

This figure displays the VIF values for the predictors in the regression model below.

```
sim_goal_diff      global_o      global_d
     3.041625      2.320456      3.261092
No significant multicollinearity detected.
```

Figure 5: VIF

The image shows Variance Inflation Factor (VIF) values for three variables: sim goal diff, global o, and global d. Additionally, the message "No significant multicollinearity detected" is displayed.

From figure 5 we can see that all values are below 5, suggesting that multicollinearity is not a major concern for these predictors. The predictors sim goal diff, global o, and global d do not exhibit significant multicollinearity, so they can be used together in a regression model without causing issues related to redundancy or instability of coefficients.

9.1. Pearson Correlation

Here, a correlation heatmap is used to visualize the multicollinearity among variables.

Table 6: Rule of Thumb for Interpreting the Size of a Correlation Coefficient

Size of Correlation	Interpretation
.90 to 1.00 (-.90 to -1.00)	very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to .30)	Negligible correlation

Figure 6: Scattered plot

Here, dark green represents strongly highly positive correlation and dark blue represents strongly highly negative correlation.

10. Results and Discussions

This section discusses the results of the fitted regression models and interprets the role of each predictor in explaining the outcome variable, goals scored.

The Simple Linear Regression (SLR) model using spi alone provides a basic estimate of team performance, but it explains only a limited portion of the variance (low R^2). In contrast, the Multiple Linear Regression (MLR) model that includes sim goal diff, global o, and global d significantly improves model performance. The R-squared value increases substantially, and residuals are more randomly distributed. Due to observed non-normality in residuals, a log-transformation of goals scored is applied. This transformation yielded improved residual plots and more stable coefficient estimates, indicating better adherence to model assumptions.

The analysis confirms that MLR provides a more robust and interpretable framework for forecasting match outcomes than SLR. These results are consistent with the descriptive insights observed earlier and align with practical expectations from football analytics.

11. Conclusion

This study evaluates the predictive performance of Simple Linear Regression (SLR) and Multiple

Linear Regression (MLR) models in forecasting goals scored during the 2022 FIFA World Cup. The results clearly demonstrate the superior performance of the MLR model over the SLR model. Specifically, the MLR model explains a significantly higher proportion of the variance in goals scored, as evidenced by higher R-squared values and lower residual standard errors.

Regression diagnostics confirm the validity of most model assumptions, including homoscedasticity and lack of autocorrelation. Although the residuals do not follow a normal distribution, this issue is addressed through a log-transformation of the dependent variable, which improves the model's adherence to normality.

Multicollinearity among predictors is evaluated using Pearson correlation and Variance Inflation Factor (VIF) analyses, both of which indicate no serious collinearity concerns. The MLR model's predictors simulated goal difference, offensive strength, and defensive strength are all statistically significant and contributed meaningfully to the model's accuracy.

A bootstrap analysis further confirmed the stability and reliability of the regression coefficients. The bootstrap means are closely aligned with the original estimates, indicating the robustness of the model's predictions.

While this study focuses on traditional regression methods for their interpretability and statistical rigor, it also acknowledges the potential of machine learning techniques to enhance predictive performance. Incorporating such models remains a promising direction for future work.

Overall, this study provides a strong statistical foundation for forecasting sports outcomes using interpretable models and paves the way for more advanced predictive analytics in sports domains (Baio & Blangiardo, 2010).

11.1. Future Direction

Further improvements may be achieved by incorporating machine learning models such as Random Forest, Support Vector Machine, and Gradient Boosting. These models can capture complex, non-linear relationships and interactions among variables, potentially leading to higher predictive accuracy and broader applicability in sports analytics.

References

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics, 37*(2), 253–264.

Dufera, A. G., Liu, T., & Xu, J. (2023). Regression models of Pearson correlation coefficient. *Statistical Theory and Related Fields, 7*(2), 97–106.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Fox, J. (2015). *Applied regression analysis and generalized linear models* (3rd ed.). Sage Publications.

Jingzheng, D. (2023). Linear regression to predict World Cup. *Highlights in Science, Engineering and Technology, 61*, 144–151.

Király, F. J., & Qian, Z. (2017). *Modelling competitive sports: Bradley-Terry-Elo models for supervised and on-line learning of paired competition outcomes*. arXiv preprint, arXiv:1701.08055.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica, 36*(3), 109–118.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi*

Medical Journal, 24(3), 69–71.

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation, 15*(1).

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*, 673–690.

Peters, G., & Pacheco, D. (2022). *Betting the system: Using lineups to predict football scores*. arXiv preprint, arXiv:2210.06327.

Simon, L., Heckhard, R., Weisner, A., Young, D., & Pardoe, I. (2021). *STAT 501: Regression Methods*. Online Notes, The Pennsylvania State University, Eberly College of Science.

Wang, S., Chen, J., & Chen, H. (2023). Player score prediction based on multiple linear regression model. *Advances in Engineering Technology Research, 4*(1), 246.