Investigation of Player Performance in the English Premier League

Jaehyeon Park, Nivetha Sathish, Domenico Oppedisano
Huy Dang, Felipe Viencio-Heap

September 7, 2024

Abstract

This study investigates the relationship between individual player performance metrics and the number of appearances in the English Premier League. Using various regression techniques like multiple linear regression and lasso regression, we analyze a dataset to identify key factors influencing player appearances. The multiple linear regression model performed best ($MSE = 93.8, R^2 = 0.976$). Still, the lasso regression model was selected as the best model ($MSE = 101, R^2 = 0.974$) because it considers multi-collinearity, providing a more realistic model than the linear regression model. The findings highlight the importance of defensive contributions in player performance. Limitations include the exclusion of position-specific metrics and the need for further research on position-based models to broaden the applicability of this study. This research offers insights for clubs and analysts, providing a better understanding of how performance metrics impact player appearances in professional football.

Keywords: Player Performance, Multiple Linear Regression, Lasso Regression, Multicollinearity.

1 Introduction

The English Premier League (EPL) is regarded as one of the most competitive football leagues globally, attracting top talent and significant global viewership. Understanding the relationship between player performance metrics and the number of appearances is crucial for clubs, coaches, analysts, and players. However, the challenge lies in accurately analyzing and interpreting those metrics to derive actionable insights essential for scouting, performance optimization, and strategic

decision-making. This study explores data related to specific performance indicators, such as goals scored, assists, and defensive actions, which influence the number of appearances players make throughout their careers in the EPL.

1.1 Literature Review

The prestige and cultural prominence of the EPL have made it the focus of many research studies. Past studies have investigated its role in global football culture as well as analyzed match and player statistics in order to gain insight into the factors contributing to the success of specific teams and players in such a highly skilled and competitive league. The general public interest in what makes certain teams and players successful aligns with researchers who aim to understand the game better and gain insight into what gives a particular team or player an edge over others. Some examples of previous studies in this area include an investigation into the evolution of physical and technical performances in the EPL (Bradley, 2015), a study of the relationship between age and match-related physical and technical-tactical performance in elite football (Rey, 2022), and the relationship between goal-timing and match outcome (Zhao, 2019). These studies used datasets of

In efforts to better understand what allows teams and players to be successful in the EPL, various studies have been done to investigate factors behind performance and how it varies from match to match. For example, a study by Bush et al. (Bush, 2015) found that the match-to-match variability of technical performance metrics of EPL players, such as tackles and interceptions is higher than the match-to-match variability of physical performance metrics, such as total distance covered during a match. The study also found that different player positions showed different levels of variability - defensive players showed higher variance with regards to offensive actions and offensive players showed higher variance with regards to defensive actions (Bush, 2015). This kind of information can provide teams with an idea of how many matches are required to accurately assess a player's physical and technical performance and how to do so based on the position that they play, which is often a challenge with analyzing the value of different football players since playing in different positions prevents players from being compared directly in terms of the same performance metrics.

In particular, the number of appearances that a player makes can be a strong indicator of

their importance within the team and performance. A previous study by Gomez et al. (Gomez, 2019), analyzing players in the French, German, Italian, and Spanish leagues, found that players categorized as "important" (those having played more minutes and made more appearances) performed differently than those considered "less important." The study also found that players often increase their efforts and appear more in seasons leading to a contract renewal (Gomez, 2019). Players were found to increase their appearances and improve shooting and passing accuracy and defensive actions as they approached the end of their contracts - likely to strengthen their bargaining position during contract negotiations (Gomez, 2019). Another study by Varmus et al. (Varmus, 2020) investigated the impact of the proportion of foreign football players' appearances on the success of teams in their home leagues and in European competitions. They concluded a significant positive correlation exists between the proportion of appearances made by foreign players and the success of football clubs. The success of foreign players (measured by appearances) is linked to the overall success of teams (Varmus, 2020). This emphasizes the importance of the number of appearances that a player makes for analyzing player and team success. Boyden at al. have also looked into the reasons why certain players have longer careers in Major League Soccer (MLS) in the US and found that the main contributors to a longer career are games played (i.e. the more appearances a player makes, the longer their career in major league soccer is), minutes played, and number of assists (Boyden, 2010). The dataset comprised of information in the MLS all-time player registry on approximately 1100 players who started careers in MLS between 1996 and 2007 - in addition to statistics such as appearances and performance metrics, demographic variables such as age, nationality, and position were compiled using external sources (Boyden, 2010). Career expectancy was estimated using life table analysis, while the likelihood of player exit was examined through event-history analysis using a logit regression model. The analysis incorporated covariates such as age, ethnicity, games played, and league size (Boyden, 2010). Overall, extensive research has been conducted on factors that contribute to team and player success in elite football, with one of the leading indicators of success for a player being metrics related to the length of time a player has spent in a league and the duration of their career within a particular league. As a result, predicting the number of appearances a player will make in their career in a specific league could be very valuable as an indicator of success.

1.2 Objective

While previous studies have examined factors influencing player performance and selection, most have focused on single metrics or specific seasons rather than providing a holistic view across multiple seasons and metrics. Our study differs by taking a comprehensive approach: analyzing a broader set of performance metrics over an extended period and considering their cumulative impact on player appearances. This approach allows us to address the research question: *How do individual player performance metrics influence the number of player appearances in the English Premier League?*

The dataset used in this study consists of player statistics from multiple EPL seasons, focusing on key performance metrics that potentially influence appearances. Analyzing cumulative effects over time presents a significant challenge, as the analysis must account for controlling for confounding factors like age and position. The study employs a quantitative approach to analyze how individual player performance metrics influence the number of appearances in the EPL.

This research project is directly relevant to the learning objectives of STA302. It applies advanced regression techniques and diagnostic tools to real-world data, demonstrating how statistical methods can address complex problems in sports analytics.

2 Methodology

2.1 Research Design

The study takes a novel approach because it moves beyond the traditional, intuition-based selection of variables commonly used in sports analytics. Instead of relying on preconceived beliefs about which metrics might influence player appearances, we employ a quantitative analysis to identify the most influential variables objectively. This method allows us to uncover relationships and insights that may not be immediately apparent through intuitive analysis alone, offering a more accurate and comprehensive understanding of the factors that impact player appearances in the EPL.

2.2 Dataset

We used the data publicly available on Kaggle for the EPL dataset (Kanabar, 2020). This dataset includes statistics for players across multiple seasons, such as `Wins`, `Assists`, and `Yellow Cards`.

The dataset initially consisted of 58 variables and 571 entries. However, we preprocessed the dataset to focus on the most relevant variables and entries. Specifically, we excluded all entries whose position was goalkeeper, as the performance metrics of goalkeepers differ significantly from those of outfield players. Additionally, we removed position-specific metrics such as `Saves`, `Penalties scored`, and `Accurate long balls` since these values are consistently omitted for certain positions. For example, `Penalties scored` for defenders will be omitted because forwards or midfielders generally deal with penalty kicks. After preprocessing, our refined dataset consists of 16 variables and 501 entries.

2.3 Model Formulation

2.3.1 Variable Selection

We employed stepwise regression using the Akaike Information Criterion (AIC) as our selection criterion to identify the most influential variables (Akaike, 1974). Stepwise regression, which includes both forward and backward elimination methods, is suitable for this analysis as it allows for the inclusion or exclusion of variables based on their contribution to model fit. Using AIC as the selection criterion is critical because AIC balances model fit with complexity, penalizing the addition of unnecessary variables. This ensures the final model is accurate and avoids overfitting while retaining the most predictive variables.

Our methodology ensures that only significant variables are selected by evaluating the contributions of each variable to the model. During the stepwise regression, variables that do not significantly improve the AIC of the model are excluded. This approach directly supports our objective of creating an accurate and reliable model by focusing on the most meaningful predictors and eliminating those that do not add substantial value.

We began by fitting a full model (`modAll`) using all selected variables and compared it to an intercept-only model (`mod0`). The initial AIC values provided a baseline for comparison as we performed forward and backward elimination to refine the model.

Through backward elimination, we arrived at an optimal model (`optimodel`) with a subset of the initial variables. This model was further validated through forward selection, confirming the validity of our variable selection process. The adjusted R^2 values for both methods were nearly identical, indicating a strong model fit.

2.3.2 Model Diagnostics

With the optimal model identified, we conducted several diagnostic tests to ensure the validity of our assumptions. We examined the residuals of the model for normality using a histogram, a Q-Q plot, and the Shapiro-Wilk test (Shapiro and Wilk, 1965) and checked for homoscedasticity using the Breusch-Pagan test (Breusch and Pagan, 1979) (see Appendix A).

Given that the initial diagnostics suggested potential non-normality and heteroscedasticity, we explored potential transformations for the dependent variable (`Appearances`) and one of the covariates (`Clearances`). After applying square root and logarithmic transformations, we rechecked the model diagnostics, resulting in improved residual behavior and a better model fit (see Appendix B).

2.3.3 Final Linear Model Selection

The final model, which incorporates the selected variables and necessary transformations, was validated using Partial F-tests (see Appendix C). We employed Partial F-tests because of their ability to compare nested models, where one model is a subset of the other. This comparison helps determine whether the inclusion or exclusion of specific variables significantly affects the model's predictive power. By testing whether a reduced model (with fewer variables) performs as well as the full model, Partial F-tests help ensure that the final model is efficient.

For instance, comparing the optimal and simpler models that excluded `Age` (`fSecondBest`) yielded a p-value of 0.1109. This indicated that removing `Age` did not significantly influence the model, suggesting that while `Age` might be intuitively relevant, it was not statistically crucial in the presence of other variables. Similarly, further tests revealed that excluding `Tackles` (moving from `fSecondBest` to `fThirdBest`) also did not significantly impact the model, as indicated by $p = 0.0824$.

Furthermore, when comparing the `fThirdBest` model (which excludes `Age` and `Tackles`) with a

model that also excludes `Hit Woodwork` (`fFourthBest`), $p = 0.00523$ indicated that `Hit Woodwork` is a critical variable. Its exclusion would lead to a substantial loss. This finding emphasizes the importance of offensive actions in determining player appearances.

Lastly, the comparison between the optimal model and a more complex model that included `Yellow Cards` (`bSecondBest`) showed $p = 0.2957$, confirming that adding `Yellow Cards` did not improve the model. This reinforces the decision to retain a simpler model without unnecessary variables.

These partial F-tests confirmed that the reduced model did not significantly lose accuracy compared to the full model and prevented overfitting the model. By excluding variables that did not add meaningful value to the performance of the model, we developed a final model that is both simpler and more interpretable. The reduced model, focusing on key predictors such as `Wins, Fouls, Interceptions, Blocked Shots, Offsides, Crosses, Clearances, Goals,` and `Hit Woodwork`, clearly represents the factors influencing player appearances in the EPL. This approach enhances the usefulness of the model in practical applications, offering solid insights with less complexity.

2.3.4 Influential Points

We also identified and addressed influential points that can disproportionately impact the model's coefficients, potentially distorting the results. We focused on using Cook's Distance as a diagnostic measure to detect these influential points, mainly because we wanted to assess the overall influence of individual players on the model (Cook, 1977). Cook's Distance evaluates how much the fitted values change when each observation is removed, thus providing a comprehensive view of each player's impact on the regression model. Points with a Cook's Distance exceeding the threshold of 4/n were flagged as influential.

After identifying these influential points, we visualized them using plots to assess their impact further. The flagged points were removed from the dataset, and the regression model was refitted. This allowed us to observe model performance changes and ensure that specific players with extreme values did not unduly influence our results.

Finally, we reexamined the residuals of the cleaned model for normality and homoscedasticity, confirming that the assumptions underlying our regression analysis were met. By addressing these influential points, we developed a more reliable model that better reflected the overall patterns in

the data.

2.3.5 Multi-Collinearity and Regularization

The Variance Inflation Factor (VIF) analysis conducted on the `optimal_cleaned` model reveals that most covariates have VIF values between 4 and 10, indicating moderate multi-collinearity among the predictors (See Figure 1). However, one covariate has a VIF value exceeding 10, suggesting a higher degree of multi-collinearity that could potentially impact the stability of the coefficient estimates and indicating that regularization is necessary.

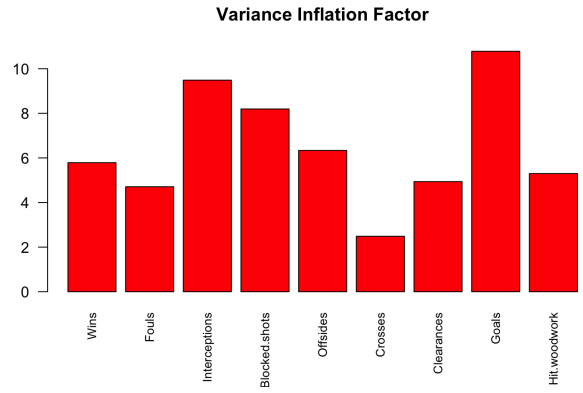


Figure 1: Variance Inflation Factors of the Covariates.

Given the presence of multi-collinearity identified by the VIF analysis, lasso regression was employed to improve the model (Tibshirani, 1996). The lasso regression model was fitted using `glmnet` package in R. Cross-validation was performed to determine the optimal value of the regularization parameter λ, which was then used to refit the model. The final lasso regression model `lasso_model_best` reflects this regularization, providing a more reliable model that addresses multi-collinearity concerns identified by the VIF analysis.

2.3.6 Model Validation

Cross-validation was employed because it assesses the performance of the model on different subsets of data, which reduces the risk of overfitting. In the process, k-fold cross-validation (where $k = 10$) was used instead of leave-one-one-cross-validation (LOOCV) due to the lower computational cost, making it more practical for our large data set. The data was split into ten subsets, with the

model trained on nine and tested on the remaining, repeating this process until each subset served as a test set. The Root Mean Squared Errors (RMSE) from all folds were then averaged to provide a more reliable estimate of the model performance.

The cross-validation results show a strong model fit (See Appendix D), with a cross-validated $R^2 = 0.9721$, which suggests that the model explains approximately 97.2% of the variance in player appearances. This high R^2 value demonstrates the model's reliability in predicting appearances based on the selected performance metrics. Additionally, the average Root Mean Squared Error ($RMSE = 10.57$), relative to the typical range of the player appearances, indicates that the model has a reasonable level of accuracy.

Overall, these results demonstrate that the model is accurate at predicting player appearances while still avoiding overfitting and ensuring generalizability. The model's strong performance suggests it could be reliably used in actual player scouting and performance analysis, providing valuable insights to clubs and coaches.

3 Results

Variable	**Linear Regression**	**Variable**	**Lasso Regression**
`Wins`	0.621101	`Wins`	0.52422381
`Blocked Shots`	0.553854	`Blocked Shots`	0.39138905
`Offsides`	0.447027	`Hit Woodwork`	0.29076598
`Fouls`	0.231333	`Offsides`	0.26299731
`Interceptions`	0.116846	`Fouls`	0.23561601
`Clearances`	0.052697	`Interceptions`	0.14094288
`Crosses`	0.047868	`Goals`	0.09089333
`Goals`	-0.308624	`Crosses`	0.04863233
`Hit Woodwork`	-0.961677	`Clearances`	0.04455158

Figure 2: The parameters ordered from most important to least important as determined by the linear regression model and the lasso regression model.

3.1 Linear Regression Model

The linear regression model (`optimal_cleaned`) demonstrates a strong fit, explaining approximately 97.59%. The model's residual standard error ($RSE = 9.815$) and significant F-statistic ($p < 2.2e-16$) further validate the model. Key predictors such as `Wins`, `Fouls`, `Interceptions`, and `Blocked Shots` show significant positive associations with appearances, with `Wins` having the greatest impact ($\beta = 0.621, p < 2e-16$). However, offensive actions such as `Goals` ($\beta = -0.309, p = 0.022$) and `Hit Woodwork` ($\beta = -0.962, p = 0.027$) are negatively associated with appearances, suggesting that players who score more goals or frequently hit the woodwork may have fewer overall appearances.

3.2 Lasso Regression Model

The lasso regression model (`lasso_model_best`) yielded $R^2 = 0.9721$, indicating that it explains about 97.21% of the variance in player appearances. The residual standard error ($RSE = 10.68$) suggests that the model has a slightly higher average prediction error than the linear regression model.

3.3 Model Comparison

We selected the lasso regression model as the best model because the linear regression approach yielded unrealistic beta values, particularly the negative coefficient for `Goals`. This suggests that lower goal counts would lead to higher player appearances, which is counterintuitive and doesn't align with the expected relationship between performance metrics and appearances. Additionally, the significant multicollinearity observed, especially with `Goals`, further weakens the reliability of the linear regression model. Lasso regression addresses these issues by applying regularization, which not only considers the effect of multicollinearity but also produces more plausible and stable coefficient estimates, making it a more suitable choice despite the slightly higher residual standard error.

4 Conclusion

4.1 Discussion

Although the linear regression model achieved a slightly lower MSE and slightly higher R^2, the lasso regression model addressed the multicollinearity in the dataset better and showed more reasonable relationships between key performance indicators and appearances. The variables that explained the most variance in the final lasso regression model were `Wins`, `Fouls`, `Interceptions`, `Blocked Shots`, `Offsides`, `Crosses`, `Clearances`, `Goals`, and `Hit Woodwork`. As the number of appearances made by a football player throughout their career in the EPL is generally considered a sign of consistency in skill and importance, the determined values of the parameters align with what previous studies have found to be signs of importance in football players. For example, the study by Gomez et al. also linked perceived importance to more minutes played, more defensive actions (i.e., fouls, tackles, and interceptions), and more offensive contributions (i.e., assists and key passes). Overall, the final model accurately identifies the most important performance indicators in predicting number of appearances. The model ultimately shows an interesting relationship between performance metrics and the number of appearances, which could be used by teams or analysts in the future to understand the long-term value of a particular football player better. The best way to predict a particular football player's long-term value and potential is a topic that is heavily discussed amongst teams and the general public due to the difficulty of estimating a unique player's market value (Selma, 2023). More research could help teams and analysts understand this topic better and make contracting decisions more effectively.

4.2 Limitation and Further Research

As all position-specific metrics were ultimately disregarded in the final model due to them not applying to all players, it could be helpful to create separate models for the different types of positions to keep position-specific metrics as variables in the model and potentially reduce error. A limitation of this approach would be that for certain positions, there could be less data for the model to use. For example, the dataset for a model regarding goalkeepers would be significantly smaller than the data available for a model regarding midfielders. However, making separate models or even introducing interaction terms between position categories and position-specific

metrics to incorporate the effect of certain position-specific moves for players who play a particular position could allow for a better understanding of what is required of the different positions to make more appearances over their careers. For example, we could understand what performance metrics a midfielder would have that lead to more appearances than a goalkeeper would have.

Further, our model formulation using k-fold cross-validation causes sensitivity to data splitting. If outliers are present in the testing set, they can inflate the model's error, potentially leading to overestimating its predictive error. Although manually removing outliers could address this issue, it is impractical for our large dataset. While the effect of the outliers may not be entirely negligible, k-fold-cross-validation reduces the impact of outliers by averaging errors across all iterations, thus making our model sufficiently reliable.

This would be an area of further research because the four players with the most appearances in the EPL are midfielders, and five of the ten players with the most appearances in the EPL are midfielders. Gareth Barry, an English professional footballer who played midfielder in the Premier League from 1997 to 2018, currently holds the record for the most Premier League appearances, with 653 appearances for his career (Premier League, 2024). The relationship between the performance metrics and the number of appearances may also help teams decide which players can maintain consistent appearances when signing future players. Another area of interest is the relationship between the success of teams and the number of appearances each of their players has made. The findings of an investigation into the difference in success between teams who have players who continue to play in the EPL throughout a long period and teams who have players who make/have made fewer appearances in the EPL could result in teams being able to employ better strategies with regards to signing new players. This could also be related to the studies of the effects of a player's age on their match performance. Many areas associated with the number of appearances made by EPL players could be further researched to provide valuable information for teams and players. Football strategy is ever-evolving, especially at the elite level, and based on the connection we have found between player appearances and performance, there is great potential for future studies regarding this topic that could result in a deeper overall understanding of team and player success.

Access to Full Code

For the full code used in this analysis, please refer to the following URL:
https://github.com/huyxdang/STA302-

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716-723.

Boyden, N., Carey, J. (2010). From One-and-Done to Seasoned Veterans: A Demographic Analysis of Individual Career Length in Major League Soccer. *Journal of Quantitative Analysis in Sports*, **6**(4).

Bradley, P., Archer, D., Hogg, B., Schuth, G., Bush, M., Carling, C., Barnes, C. (2015). Tier-specific evolution of match performance characteristics in the English Premier League: it's getting tougher at the top. *Journal of Sports Sciences*, **34**(10).

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, **47**(5), 1287-1294.

Bush, M., Archer, D., Hogg, R., Bradley, P. (2015). Factors influencing physical and technical variability in the English Premier League. *International journal of sports physiology and performance*, **10**(7).

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15-18. https://doi.org/10.2307/1268249.

Gomez, M., Lago, C., Gomez, M., F, P. (2019). Analysis of elite soccer players' performance before and after signing a new contract. *PLoS ONE*, **14**(1).

Kanabar, R. (2020). All Time Premier League Player Statistics. Retrieved August 18, 2024, from [https://www.kaggle.com/datasets/rishikeshkanabar/premier-league-player-statistics-updated-daily].

Premier League. (2024). Gareth Barry Overview. Retrieved August 18, 2024 from [https://www.premierleague.com/players/1308/Gareth-Barry/overview].

Rey, E., Lorenzo-Martinez, M., Lopez-Del Campo, R., Resto, R., Lago-Penas, C. (2015). No sport for old players. A longitudinal study of aging effects on match performance in elite soccer. *Journal of Science and Medicine in Sport*, **25**(6).

Selma-Malagon, P., Debon, A., Domenech, J. (2023). Measuring the popularity of football players with Google Trends. *PLoS ONE*, **18**(8).

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3-4), 591-611.

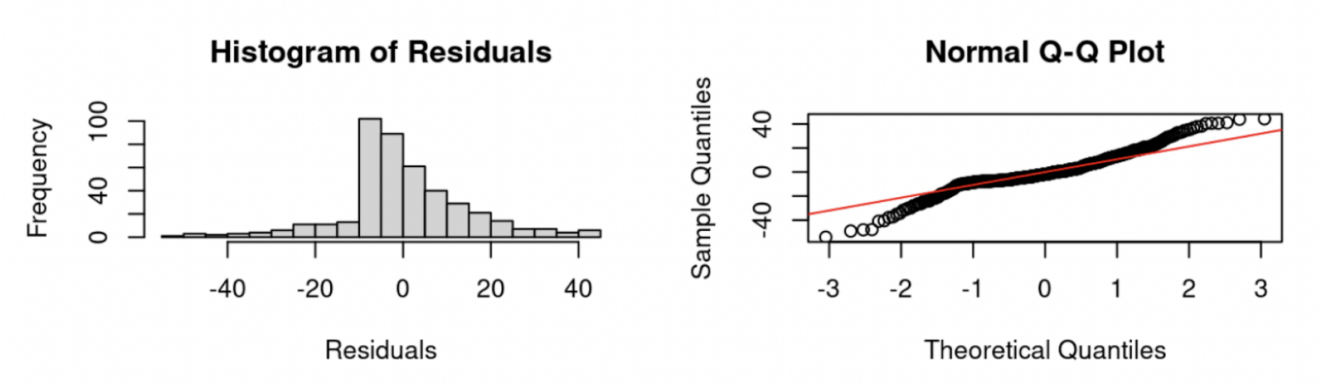
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**(1), 267-288.

Varmus, M., Kubina, M., Adamik, R. (2020). Impact of the Proportion of Foreign Players' Appearances on the Success of Football Clubs in Domestic Competitions and European Competitions in the Context of New Culture. *Sustainability*, **12**(1).

Zhao, Y., Zhang, H. (2019). Analysis of goals in the English Premier League. *International Journal of Performance Analysis in Sport*, **19** (5).

Appendix A

Results of original model (`optimal`)'s Historgram of Residuals and Q-Q Plot

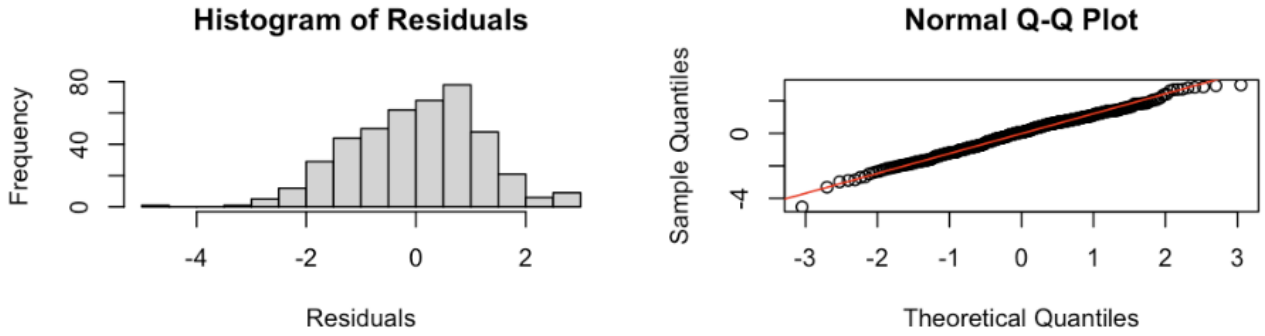


Results of original model (`optimal`)'s Shapiro-Wilk and Breusch-Pagan Tests

```
        Shapiro-Wilk normality test

data:  residuals(optimal)
W = 0.95119, p-value = 8.848e-11


        studentized Breusch-Pagan test

data:  optimal
BP = 131.88, df = 9, p-value < 2.2e-16
```

Appendix B

Results of transformed model(`optimal_sqrt`)'s Historgram of Residuals and Q-Q Plot



Results of transformed model(`optimal_sqrt`)'s Shapiro-Wilk and Breusch-Pagan Tests

```
        Shapiro-Wilk normality test

data:  residuals(optimal_sqrt)
W = 0.99411, p-value = 0.09188


        studentized Breusch-Pagan test

data:  optimal_sqrt
BP = 40.385, df = 9, p-value = 6.471e-06
```

Appendix C

Results of Partial F-Tests for Final Model Selection

```
Analysis of Variance Table

Model 1: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals + Hit.woodwork +
    Age + Tackles
Model 2: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals + Hit.woodwork +
    Tackles
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    422 91352
2    423 91905 -1   -552.31 2.5514 0.1109
Analysis of Variance Table

Model 1: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals + Hit.woodwork +
    Tackles
Model 2: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals + Hit.woodwork
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    423 91905
2    424 92563 -1    -658.6 3.0313 0.0824 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals + Hit.woodwork
Model 2: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    424 92563
2    425 94283 -1   -1720.2 7.8797 0.00523 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: Appearances ~ Wins + Fouls + Interceptions + Blocked.shots +
    Offsides + Crosses + Clearances + Goals + Hit.woodwork +
    Age + Tackles
Model 2: Appearances ~ Age + Wins + Goals + Hit.woodwork + Tackles + Blocked.shots +
    Interceptions + Clearances + Crosses + Yellow.cards + Fouls +
    Offsides
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1    422 91352
2    421 91115  1     237.2 1.096 0.2957
```

Appendix D

Results of Partial F-Tests for Final Model Selection

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-28.496  -6.291  -2.564   6.166  42.161

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.968419   0.730076   9.545  < 2e-16 ***
Wins           0.709847   0.042345  16.763  < 2e-16 ***
Fouls          0.282812   0.017901  15.799  < 2e-16 ***
Blocked.shots  0.537140   0.061592   8.721  < 2e-16 ***
Offsides       0.385092   0.064625   5.959 5.92e-09 ***
Crosses        0.057329   0.003980  14.405  < 2e-16 ***
Clearances     0.075026   0.003476  21.581  < 2e-16 ***
Goals         -0.368560   0.140183  -2.629  0.00892 **
Hit.woodwork  -1.107063   0.454508  -2.436  0.01533 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.31 on 370 degrees of freedom
Multiple R-squared:  0.9734,    Adjusted R-squared:  0.9728
F-statistic:  1691 on 8 and 370 DF,  p-value: < 2.2e-16
```