Analyzing the Impact of Performance and Personal Attributes on Football Attackers' Market Value

Shayan D'Souza, Daniel Du, Derek Li

Abstract

This analysis investigates how performance statistics and personal attributes influence the market value of football attackers, identifying key factors like career performance and league strength as significant predictors while acknowledging limitations like broader contextual influences.

Introduction

In the football world, player valuations shape team strategies and financial planning. They inform decision-making for clubs and agents, enhancing the efficiency of player transfers and ensuring fair contract negotiations. This report explores the question: How do a football attacker's performance statistics and personal attributes influence their market value?

We hypothesize that the following factors affect player market value, reported by website Transfermarkt: attributes including age, attacking position, and contract length; recent performance statistics such as goals, assists, and minutes played in the current season, as well as the league played in; and historical performance measured by career goals and assists per game (Transfermarkt, n.d.).

Existing research supports this focus. Metelski (2021) studied Euro 2020 participants finding that younger players have greater market value due to their potential. He, Cachucho, and Knobbe (2015) applied machine learning to market valuations, observing that forwards' values were easier to predict through goal and assist totals. Muller, Simons, and Weinmann (2017) emphasized the higher market value of those appearing in Europe's strongest leagues. Understanding the magnitude of these relationships offers clubs and analysts a data-driven approach to evaluating players and making transfer decisions.

After gathering existing web-scraped data, we conduct our analysis by fitting a linear model to estimate and interpret relationships between the above factors

and market value in millions (Cariboo, 2024). The model coefficients quantify changes in market value for increases in our variables, aligning with the goal of understanding these factors' influence. Linear regression handles multiple predictors, allowing a comprehensive assessment and interpretation.

Methods

We begin by fitting a linear regression model with 9 predictor variables representing the hypothesized factors. Players are categorized by positions “Winger”, “Attacking Midfielder”, or “Centre Forward”. All other variables are continuous, including league strength, measured by an official coefficient; values, as well as the players per league, are listed in Appendix Figure A (Transfermarkt, n.d.).

To ensure model validity, we extract the model’s fitted and residual values to verify the assumptions and conditions of multiple linear regression. First we check the conditional mean response condition with a response vs fitted values graph, and the conditional mean predictor condition with response vs predictor graphs for each predictor. This ensures predictors correctly explain variation in the response variable and that residual patterns are random.

Afterwards, we check for the uncorrelated errors, linearity, and constant variance assumptions by looking at a residual vs fitted values graph, residual vs predictor graphs for each predictor, and pairwise scatterplots between all predictors. We also verify there is no multicollinearity between predictors by checking that the Variance Inflation Factor (VIF) for each predictor is not significantly high. We check the normality assumption with a normal Q-Q plot.

Performing the residual analysis, we found our model violated the constant variance and linearity assumption, as well as the mean response condition. To mitigate this, we apply a variance stabilizing transformation on our response variable to address its skew. Next, we apply variance stabilizing and box-cox power transformations to any predictor variables, ensuring constant error variance and linear relationships. We reverify all prior assumptions and conditions for our fully transformed model to ensure no further violations are present.

We then use an ANOVA test to verify a significant linear relationship between

the response and at least one predictor exists, as well as t-tests on all predictors to quantify the statistical significance of each. We note insignificant predictors and fit a reduced model excluding them. After checking this reduced model meets all prior assumptions and conditions without issue, we perform a partial F-test to determine if more variance in the response variable is explained by the full model. Since the partial F-test was insignificant, we conclude that the predictors indicated to be insignificant by the t-test do not explain any additional variance in the response variable, so we can exclude those predictors.

Performing another ANOVA test on the reduced model confirmed there was a significant linear relationship between the response and at least one predictor, while t-tests indicated that all remaining predictors were significant. Finally, we construct confidence intervals for coefficients and the mean response to determine the precision of our coefficients and model estimates. To determine the predictive accuracy of the model, we also construct prediction intervals for the response variable and verify their accuracy.

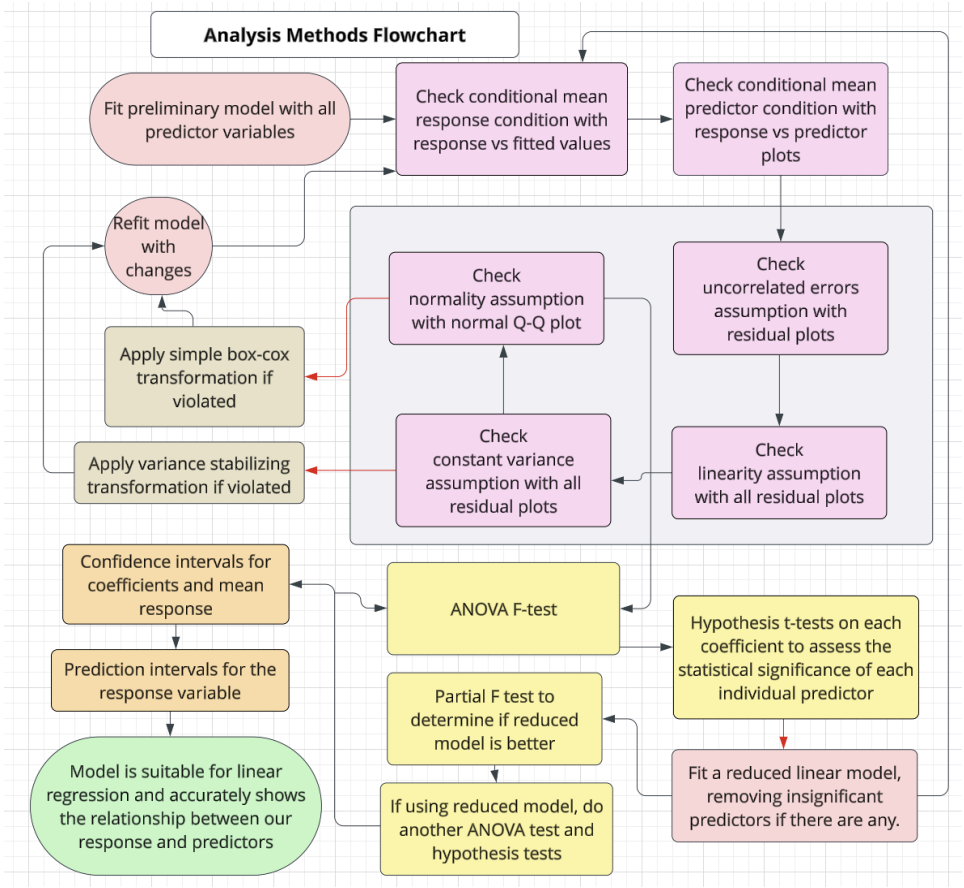


Figure 1: Rough flowchart of methods used to determine final model

As it passes all violation checks, our model is appropriate for linear regression. Using prediction intervals, we can estimate the market value range of a new player given their statistics and attributes. We use the confidence intervals of our model coefficients to find the magnitude of the effect of each predictor on market value, answering our research question.

Results

As outlined in the previous section and summarized in Figure 1, we fit a linear model with all hypothesized predictors and perform diagnostic tests. Plotting fitted values against true response variable values as seen in Figure 2, we determine that a log transformation on our response variable is appropriate to address its high right-skewness and stabilize the variance.

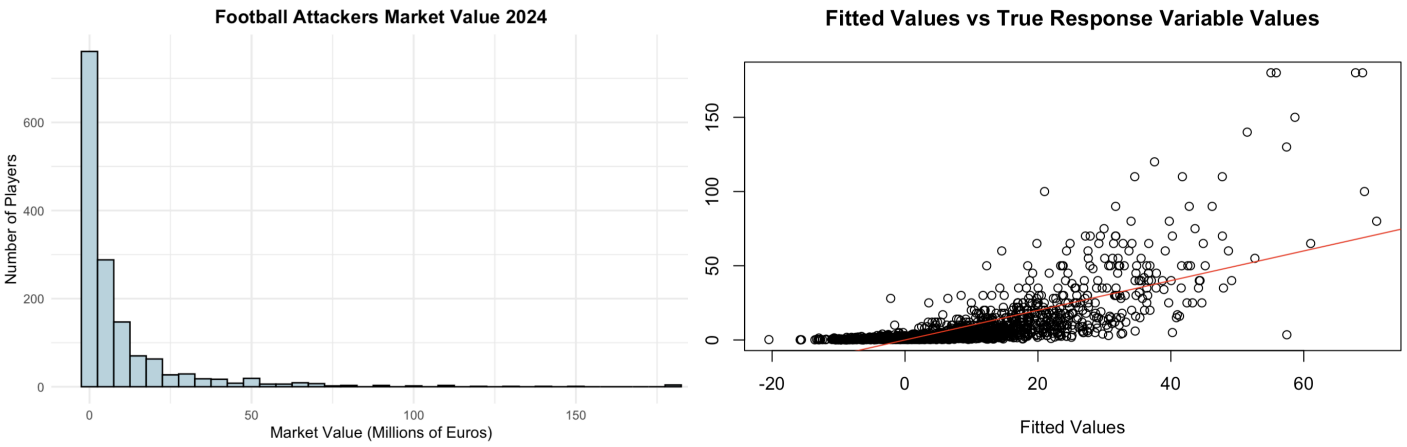


Figure 2: Plots indicate need for log transformation on response variable.

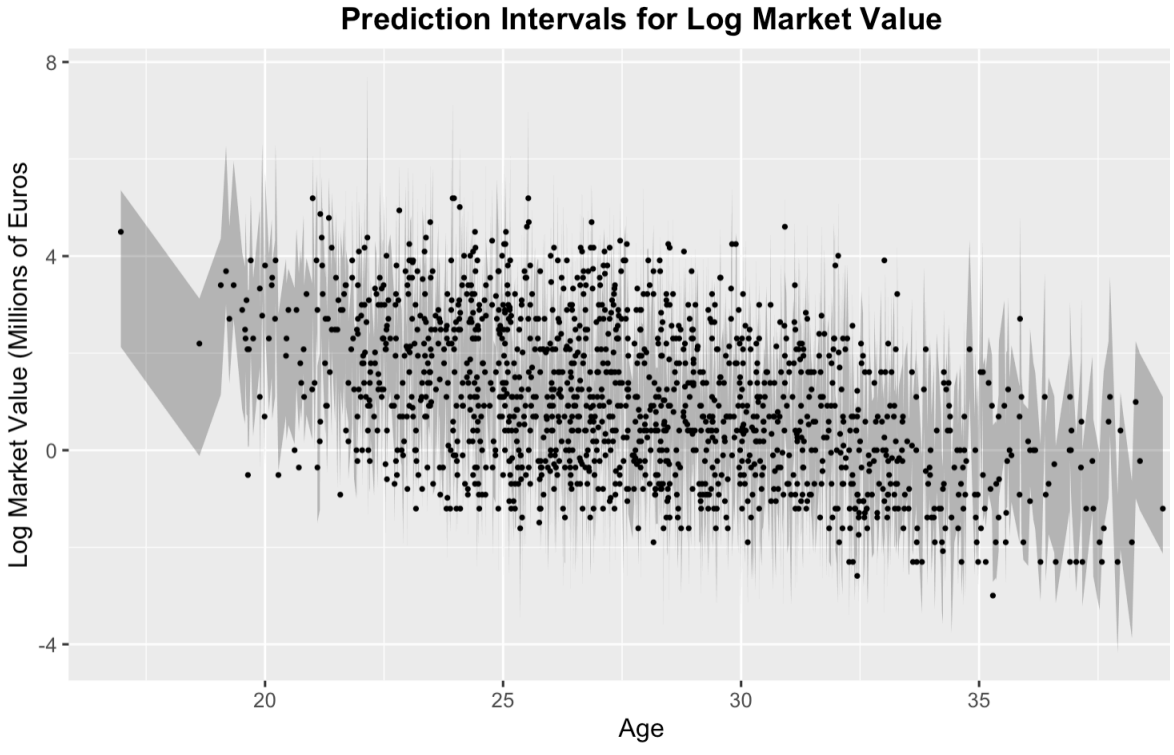
We also applied transformations to problematic predictors. We applied a box-cox power transformation to league coefficient and season minutes, deciding to use the square root of the league coefficient, and season minutes to the power of ⅔. As variables representing season goals and season assists were distributed with a right skew, we applied a log(x + 1) transformation to stabilize variance, adding 1 to the predictor to account for zero values while still preserving the relative order of the data. Evidence of residual plots justifying our transformations on the predictors and response variable are in Appendix Figure B and Figure C, respectively.

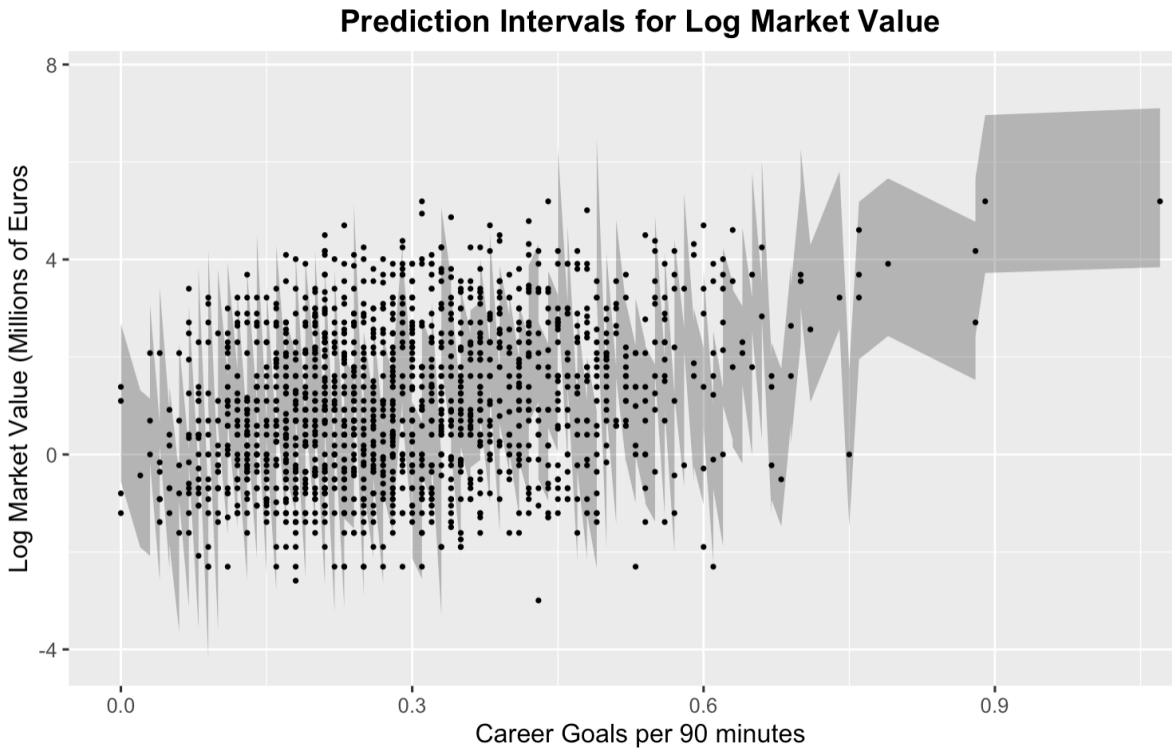
After performing an ANOVA f-test and hypothesis tests, we found that the log-transformed season assists predictor was potentially insignificant in our model, and verified this by fitting a reduced model and performing a partial f-test; this suggested to remove log-transformed season assists, as they were not a significant explainer of any variance in the log market value. Intuitively, single-season assists have little predictive power, as one season is a small sample size and assists are affected by outside factors such as teammate quality and team tactical instruction.

Thus, our final model variables included age, position, contract length, career goals per 90 minutes, career assists per 90 minutes, log-transformed season goals, the square root of the league coefficient, and season minutes to the power of ⅔.

Our final model has a multiple and adjusted R-squared value of 0.73, meaning that 73% of the variability in the log market value is explained by our predictors. We can assess the reliability and practical utility of the model by creating prediction intervals to see how the model predicts individual outcomes for unseen data while accounting for uncertainty. In Figure 3, we analyze 95% prediction intervals for the fitted values, depending on age and career goals per 90 minutes; 95% of our real response values fall within their respective intervals, confirming that our model is correctly specified and not overfit.

Figure 3: Real log market values plotted with shaded regions representing 95% prediction intervals for log market value, x-axes age and career goals per 90.





As our final model, outlined in Figure 4, is reliable and explains a significant portion of variance in log market value, we use its coefficients to draw conclusions about the impact of predictors. Table 1 contains each model coefficient value:

Coefficient	**Value**
β_0: Intercept (represents when all predictors are 0 or baseline)	-1.030
β_1: Age (in years)	-0.115
β_2: Position (Winger) (compared to baseline position, Att. Midfielder)	-0.114
β_3: Position (Centre Forward) (compared to baseline position)	-0.402
β_4: Contract length (years until expiry)	0.159
β_5: Square root of league strength coefficient	0.361
β_6: Log of goals scored in the current season	0.211
β_7: Minutes played in the current season to the power of 2/3	0.006
β_8: Career goals scored per 90 minutes	2.373
β_9: Career assists per 90 minutes	3.658

Table 1: Each model coefficient and their coefficient value, representing change in log market value.

$$\log(\text{MarketValue}) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Position (Winger)} + \beta_3 \cdot \text{Position (Centre Forward)}$$
$$+\beta_4 \cdot \text{Contract Length} + \beta_5 \cdot \sqrt{\text{League Strength}} + \beta_6 \cdot \log(\text{Goals Scored})$$
$$+\beta_7 \cdot \text{Minutes Played}^{2/3} + \beta_8 \cdot \text{Career Goals/90} + \beta_9 \cdot \text{Career Assists/90} + \epsilon$$

Figure 4: Final model. Interpretation of this model is used to determine the effects of various player attributes and statistics on market value.

The coefficient value represents the change in the log market value (in millions) for a one-unit increase in each coefficient. Since there are transformed variables, including the response, we will have to account for this when interpreting coefficients.

An example of interpretation is as follows:, a one-unit increase in contract length results in an 0.159 increase in the log market value. When converting to real market value in millions values, we calculate a percent change for the coefficient, finding that one additional year in contract length corresponds to a 17.2% increase in a player's market value, with all other variables held constant.

$$\text{Percent Change} = \left(e^{\beta} - 1\right) \times 100$$
$$e^{0.159} \approx 1.172$$
$$\text{Percent Change} = (1.172 - 1) \times 100 = 17.2\%$$

The two position coefficients are interpreted in comparison to the baseline position, attacking midfielder. A centre-forward has 0.402 lower log market value, or equivalently is worth 33% less than an attacking midfielder with equal statistics.

Note that the coefficients are not precise values, and have a standard error. To increase interpretation precision, in Figure 5 we calculate 95% confidence intervals for each coefficient, representing the range in which we are 95% confident the true effect of the predictor on market value lies. From our model results, we can make conclusions about the effect of different player attributes on statistics on player market value, and the real-world implications of our findings.

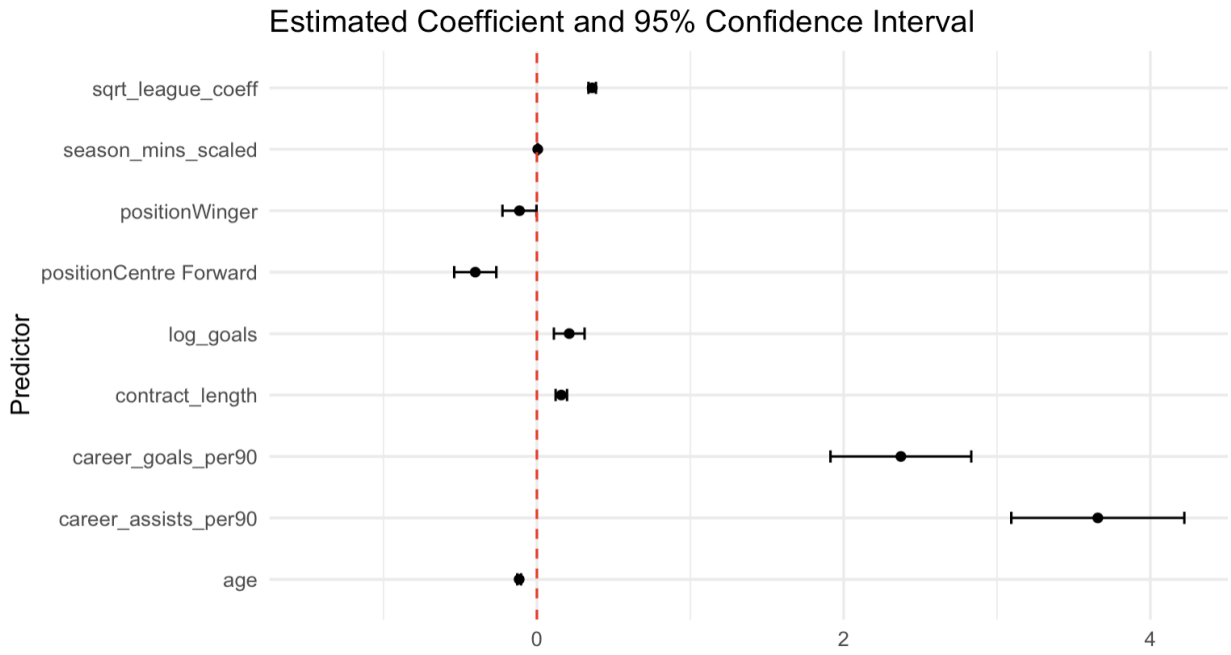


Figure 5: Estimate coefficient values and 95% confidence intervals for each predictor variable

Conclusions and Limitations

The results demonstrate that the most significant impact on player value is from career statistics; players with higher career goals and assists per 90 have significantly higher values, highlighting the importance of consistent long-term performance. For example, from the coefficient 2.37, a 0.1-unit increase in career goals per 90 corresponds to a 0.237 increase in log market value (97% increase in market value), showing that a player with 1 more goal for every ten 90-minute matches in his career is worth about twice as much.

Current season performance influences market value to a lesser degree; log-transformed goals has a sizable effect, but the effect of minutes played in a season is much smaller, and current-season assists were found to be insignificant. This is likely due to a season's small sample size and player dependence on factors such as team ability and tactics increasing the effect of variance; players need to perform across longer periods of time to prove their worth.

With all other stats and attributes held equal, attacking midfielders are worth the most, followed by wingers and centre forwards. This is intuitive considering the opportunities to contribute offensively for each position. Value is increased by playing in a stronger league, and decreased by an increase in age; this affirms the

conclusions of Metelski (2021), who observed that younger players are more valuable, and Muller et al. (2017), who emphasized the role of league strength. A longer contract length increases market value, which is intuitive as teams have less incentive to sell players on long deals.

There are some limitations to the analysis; the model R-squared value of 0.73 suggests that 27% of variance in market value remains unexplained. It would be worth examining additional factors such as wages, injuries, and external economic conditions. Additionally, the market value itself does not fully reflect player worth; different clubs value players at different prices for various reasons including personal familiarity, tactical fit, personality, and salary cost. Furthermore, grouping an attacker's position into three broad categories does not fully encompass the tactical differences between roles in different teams; future studies could employ more granular classifications.

Ethics Discussion

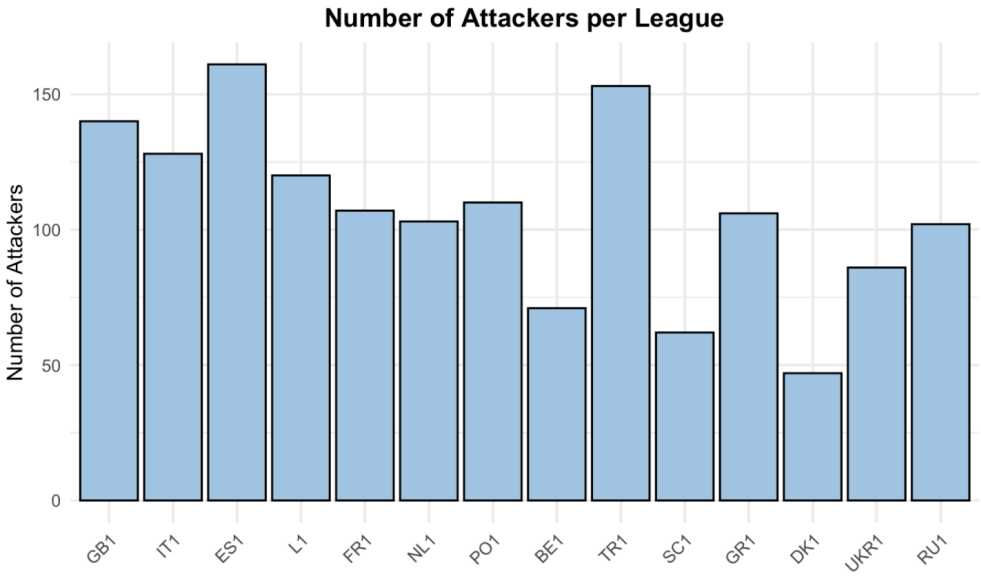
In our analysis, we decided to rely on manual selection methods instead of automated ones. The most important reason for this strategy is in the need for interpretability, fittingness to the research question, and ethical considerations. Manual selection allows us to select such predictors that might bring domain knowledge in football and statistical relationships to the model to keep it focused on the variables that are of foremost relevance to market value. This approach is in tune with our objective to explore the individual effects of performance statistics and personal attributes on the valuation of players.

While automated methods can optimize model performance, they may exclude meaningful ones based on statistical criteria. This may result in either overfitting or misinterpretation, especially where domain knowledge is significant. That is, automated methods may prefer less relevant variables to established determinants, such as league strength, against established football valuation frameworks.

From an ethical point of view, either can be used responsibly and therefore are valid. However, manual selection is more ethical in terms of transparency and accountability since one gives reasons for the variables to be included on relevance and the contribution that needs to be measured. The automation methods are likely to foster negligence on the grounds of variable inclusions and exclusions when these are inconsistent with goals or even domain knowledge.

Manual selection is chosen here, not because automated methods are unethical, but in this particular case, it fits our focus on interpretability and responsible model building to ensure fairness and rigour in the evaluation of player attributes.

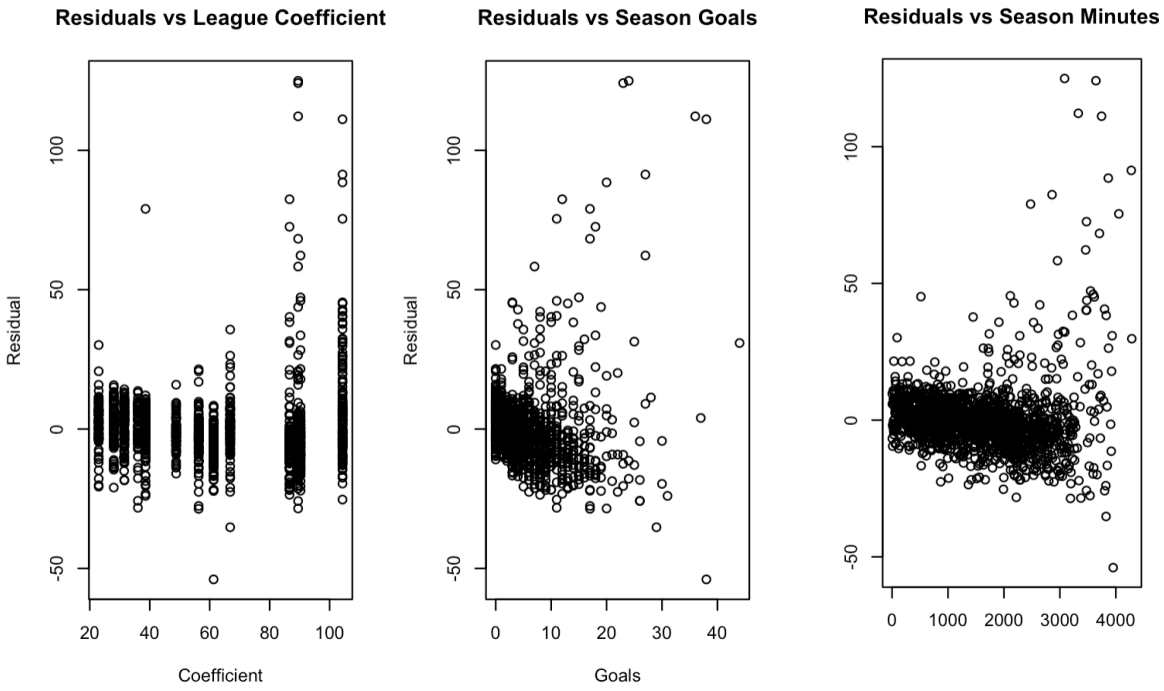
Appendix



League Strength Coefficient 2023-24

GB1	104.303	NL1	61.300	GR1	31.525
IT1	90.284	PO1	56.316	DK1	31.450
ES1	89.489	BE1	48.800	UKR1	28.000
L1	86.624	TR1	38.600	RU1	22.965
FR1	66.831	SC1	36.050	Source: Transfermarkt	

Figure A: Table of league coefficients and attackers per each league.



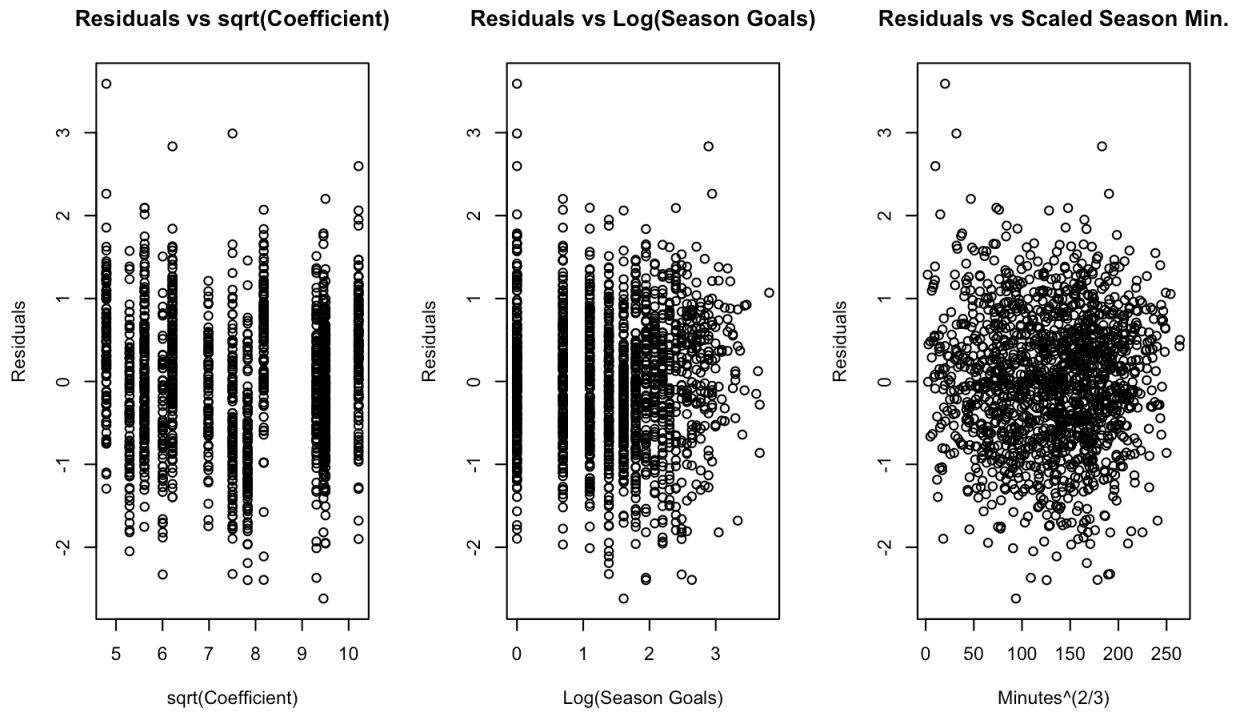


Figure B: Diagnostic plots of residuals for predictors league coefficient, season goals, and season minutes, before (above) and after (below) variance stabilizing and Box-Cox transformations.

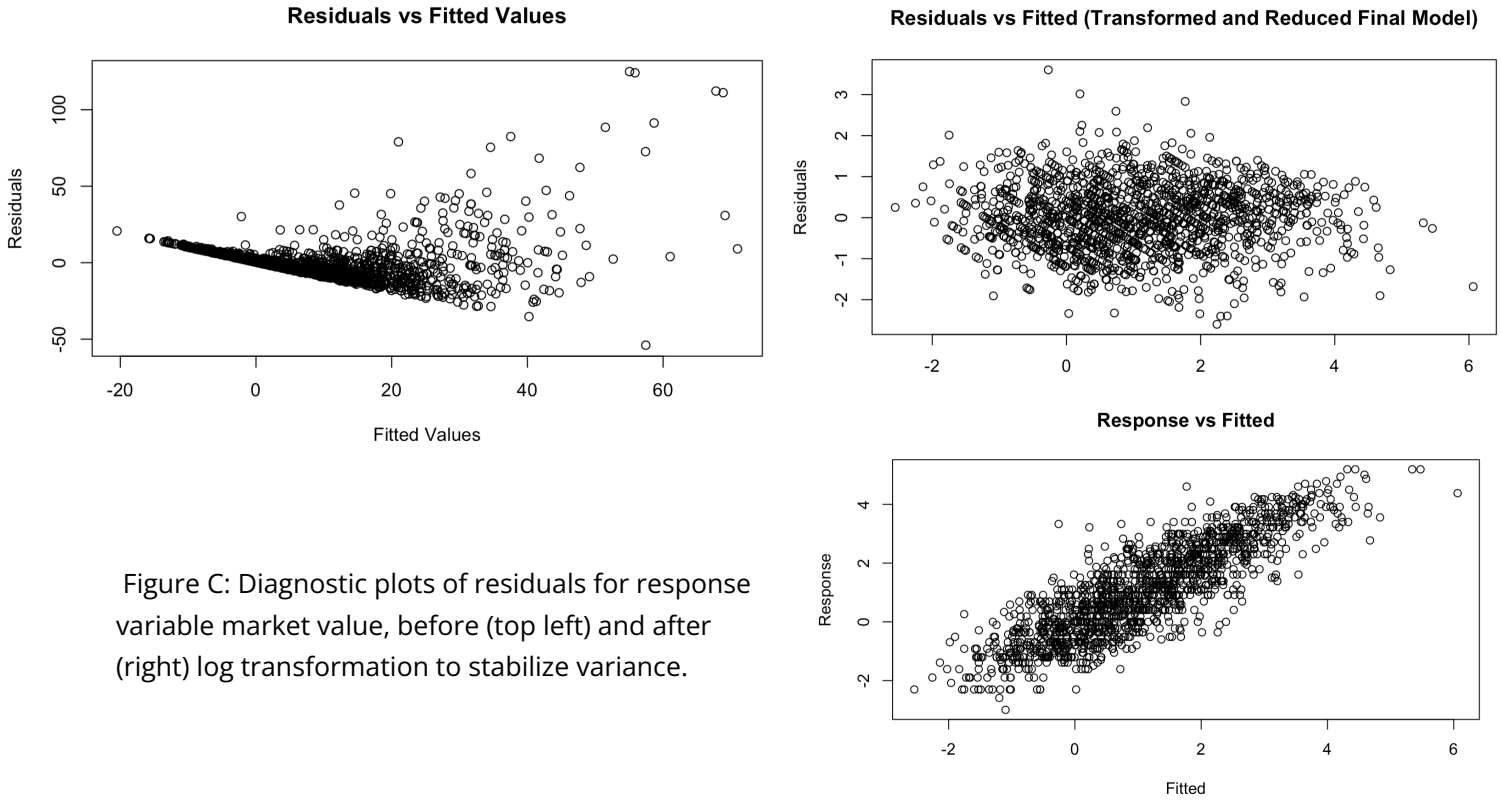


Figure C: Diagnostic plots of residuals for response variable market value, before (top left) and after (right) log transformation to stabilize variance.

References

Cariboo, D. (2024). *transfermarkt-scraper* [Code repository]. GitHub. https://github.com/dcaribou/transfermarkt-scraper

Cariboo, D. (2024). *Football Data from Transfermarkt*. Kaggle. https://www.kaggle.com/datasets/davidcariboo/player-scores/data

He, M., Cachucho, R., & Knobbe, A. J. (2015, June). Football Player's Performance and Market Value. In Mlsa pkdd/ecml (pp. 87-95). Link

Metelski, A. (2021). Factors affecting the value of football players in the transfer market. Journal of Physical Education and Sport, 21, 1150-1155. Link

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research, 263(2), 611-624. Link

Transfermarkt. (n.d.). *Transfermarkt*. https://www.transfermarkt.com