Analysis on the Value of FIFA Players

San Jose State University

Abstract

Since 1993, over 100 million copies of the game FIFA, a video game about soccer, have been sold. One feature of the game is the career mode in which we take over the management role. To be a successful manager, one is concerned about the: How do I obtain the maximum value of players I am seeking for to yield the most profits from selling him? In this analysis, we want to help managers answer that question. Specifically, we will study factors that affect the value of the player. Taking that into account, we want to build and compare various models, then select the one predicting a player's value with high accuracy.

1. BACKGROUND AND SIGNIFICANCE

FIFA is one of the most popular soccer games in the world. Since the game's inception in 1993, over 100 million copies of the game have been sold across all platforms (Businesswire, 2010). Soccer is a competitive sports game where each team has 11 players. There are 4 main classes of positions: forwards (commonly known as strikers as well), midfielders (can be offensive or defensive), defender (also known as full-backs and centre-backs), and lastly the goalkeeper (protecting the team's goal post). Each team requires 1 goalkeeper and the other 10 are composed of forwards, midfielders, and defenders. The selection of these positions illustrated the type of formation the team is playing. For example, the 4-4-2 formation is the most widely used formation in soccer. The first number indicates the number of defenders, second number indicates the number of midfielders, and the third number indicates the number attackers. So the 4-4-2 formation would have, 4 defenders, 4 midfielders, and 2 attackers. The attributes between defenders, midfielders, and attackers varies. Defenders are more skilled in physical attributes like strength and balance, and in skill attributes like tackling. Midfielders, are generally well-rounded across most attributes due to the fact they are required to play offense and defense during a match. Forwards, are very skilled in physical attributes like speed, agility, and quickness, and skill attributes such as passing, dribbling, and shooting. All of these attributes compile into a player's overall rating.

The dataset was scraped from the website https://sofifa.com, and updated at https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset/data by Aman Shrivastava, a Kaggle user and data analyst at Citigroup, Pune, Maharashtra, India. Initially, there are 17,981 observations with 70 variables. Our variables of interest are:

1. **ID** – a player's identification in substitution of Name.
2. **Value** – measured in Euro; how much money one needs to pay to adopt this player. This will be our response variable.
3. **Overall** – current rating of player on 0-99 scale.
4. **Potential** – the rating on 0-99 that a player may achieve after training and winning.
5. **Nationality** – nationality of a player, there are 165 different countries.
6. **Club** – current club to which a player belong; there are 648 clubs.
7. **Preferred.Positions** – a string that list positions that are suitable with a player; there are 15 different positions with 3 categories: Goalkeeper – GK, Attacker – ST, RW, LW, RM, CM, LM, CAM, CF and Defender – CDM, CB, LB, RB, RWB, LWB.
8. **Age** – player's age.
9. **Performance Attributes** (each shows the corresponding attribute's rating on 0-99 scale of a player.) - **Acceleration, Aggression, Agility, Balance, Ball.control, Composure, Crossing, Curve, Dribbling, Finishing, Free.kick.accuracy, Heading.accuracy, Interceptions, Jumping, Long.passing, Long.shots, Marking, Penalties, Positioning, Reactions, Short.passing, Shot.power, Sliding.tackle, Sprint.speed, Stamina, Standing.tackle, Strength, Vision, Volleys**

Value have records in currency format, e.g. €95.5M. We converted them into an exact number, e.g. 95.5 * 1,000,000 = 95,500,000. However, after that, we found out that there are 256 records whose **Value** is 0. In some special events, FIFA introduces untradeable players to award gamers. In the scope of this analysis, our goal of predicting value is to earn profits, so we disregard these 256 players.

Concurrently, records of performance attributes are stored as strings; and even worse, some of them are unevaluated expressions, e.g. "79+1". We evaluated those expressions and converted all records of performance attributes into integers. Then, we realized that two observations whose **IDs** are 204846 and 239740 respectively, have negative performance attributes, **Interceptions** and **Marking** specifically. At the same time, four observations whose **IDs** are 233795, 219255, 238656 and 240611 respectively, have performance attributes over

100, **Aggression**, **Stamina**, **Crossing** and **Sprint.speed** specifically. We believed they are data collecting errors as performance attributes are from 0 to 99. Hence, they are removed.

We decided to focus on predicting values of attackers and defenders only, so we removed 1984 players that are goalkeepers. In this dataset, a goalkeeper is recognized by the fact that **Preferred.Positions** indicates "GK." In the end, we were left with 15,735 observations.

Furthermore, records of **Preferred.Positions** are quite inconvenient to do analysis. We wanted to categorize a player's preferred positions in a more general way. That is, "Attacker" if its preferred positions are the subset of ST, RW, LW, RM, CM, LM, CAM, CF only; "Defender" if its preferred positions are the subset of CDM, CB, LB, RB, RWB, LWB only; "Both" if its preferred positions are the subset of both groups. We stored the results in a new variable called **Attacker.Or.Defender**. Later, when we mention "preferred position," we are referring to this variable.

Lastly, regarding convenience to do analysis, we categorized by continent of origin instead of their nationality. We stored the results in a new variable called **Continent**.

2. OBJECTIVES

Objective 1

This objective comprises of three parts. In the first part **,** we were concerned whether a player's continent of origin influences his value. It means, on average, there is at least one continent whose players have value different from those coming from other continents. If that happens to be the case, we are also interested in understanding some characteristics regarding the difference like its size and direction.

In the second part, we determined whether a player's preferred position, attacker, defender or both, could affect his value. In other words, we want to know if we should expect a higher value from an attacker, a defender or someone who can perform well in both positions. A further analysis will also be conducted to understand how a player's value varies depending on his preferred position.

Finally, the third part was dedicated to studying the interaction effect between the factors mentioned, continent of origin and preferred position. In the last two parts, we assumed that each factor has an exclusive contribution to a player's value. However, in reality, a particular combination of factors can also lead to a secondary result on the response, namely interaction effect, that cannot be observed by merely looking at each factor's effect separately. This part will help us find out if such effect exists.

Objective 2

This objective is about constructing a predictive model, specifically linear regression, for a player's value in reliance on his age, value, continent of origin, overall and potential ratings, all performance attributes and preferred positions. In most cases, this is not a simple task as it sounds, since various challenges may arise. They are:

- Not all predictors have some useful contributions to the response, so feature selection is necessary in order to filter redundant variables.
- Sometimes, by nature, there could exist high correlations between certain predictors, also known as multicollinearity. While this may not affect the model fit, the estimates of coefficients will be less accurate. Removing one of the two highly correlated predictors or applying a shrinkage method can overcome this issue***.***
- Linear regression relies heavily on some assumptions of residuals, namely independence, normality and homogeneity of variances, which we need to verify (Objective 2.3). Often, they are not satisfied by default, and thus, we need at least one transformation to resolve the problems. However, as a result, the model can become complex and cumbersome.

- With such a large dataset like ours, the presence of outliers is guaranteed. When being influential, outliers hurt the model in the same manner with multicollinearity. Outliers need to be examined for its potential to be influential.
- Different criteria of feature selection and residual sum of squares minimization can result in different models. If we do not have a specific goal about inferential power for our model, i.e. a subset of variables we want to keep, we should choose the one with the most predictive power. That means to have all models obtained predict new data, and compare their mean squared errors.

Successful accomplishment of this objective will enable us to not only make compelling statements about the response-predictor relationship, but also to predict a future player's value with high accuracy.

3. METHODS

Two-way ANOVA

Two-way ANOVA tests for the true average responses associated with different levels of two factors or treatments(Devore, 2012, pp. 433-434). To be specific, we are concerned with the model

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where α_i is the effect of factor A at level i, β_j is the effect of factor B at level j, and γ_{ij} is the interaction effect between factor A and B. Three hypotheses to be considered are:

$$H_{0AB}: \gamma_{ij} = 0 \; for \; all \; i,j \;\; versus \;\; H_{aAB}: at \; least \; one \; \gamma_{ij} \neq 0$$
$$H_{0A}: \alpha_1 = \ldots = \alpha_I = 0 \;\; versus \;\; H_{aA}: at \; least \; one \; \alpha_i \neq 0$$
$$H_{0B}: \beta_1 = \ldots = \beta_I = 0 \;\; versus \;\; H_{aB}: at \; least \; one \; \beta_i \neq 0$$

Graphical Residual Analysis

This refers to the use of the residual plot and QQ-plot to check the assumptions of linear regression models. Ideally, we want to see "confetti in the box" and no discernible pattern in the residual plot. At the same time, dots in the QQ-plot should not depart too much from the straight line.

Box-Cox Transformation

Box-Cox transformation is one function of power transformation that helps stabilize variance and make the data more normal (Seal, 1967). The one-parameter Box-Cox transformations are defined as:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & if \; \lambda \neq 0 \\ ln(y_i) & if \; \lambda = 0 \end{cases}$$

The parameter λ is estimated using the profile likelihood function. In most case, if λ is close to 0, a log transformation is appropriate.

Tukey's Procedure

After conducting an ANOVA, we want to investigate which means are different from one another, which can be done using Tukey's Procedure. Let $Q_{\alpha,m,\nu}$ denote the upper- tail α critical value of the Studentized range distribution with m numerator df and ν denominator df (Devore, 2012, pp. 402-403). Let $w = Q_{\alpha,I,I(J-1)} * \sqrt{\frac{MSE}{J}}$, where I is the number of groups and J is the number of observations in each group. Tukey's procedure involves listing the sample means in increasing order. Any pair of sample means that do not differ by less than w are said to be significantly different (Devore, 2012, pp. 402-403).

Multiple Linear Regression

Multiple linear regression is a quantitative method to model the statistical relationship between a continuous response and one or more explanatory variables. To be specific, it focuses on the conditional mean of the response given the values of explanatory variables using the least squares approach where sum of squared errors is minimized (A., 1946). Unlike non-linear regression, linear regression is easier to fit and observe the response-predictor relationship. It can be mathematically stated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_i x_{ip} + \varepsilon_i$$

where p is the number of explanatory variables involved, and the error term ε_i is assumed to be independent and identically normally distributed with mean 0 and equal variances σ^2.

Model Utility Test

Model utility test asserts whether at least one of explanatory variables involved is useful in predicting the response in linear regression model (Devore, 2012, pp. 561-562). The parameter of interest in this case is all explanatory variables' coefficients. Specifically,

$H_o: \beta_1 = \beta_2 = \ldots = \beta_p$ or there is no useful relationship between the response and any of explanatory variables.

$H_a: At\ least\ \beta_j \neq 0\ (j = 1, 2, \ldots, p)$ or there is at least one explanatory variable having useful relationship with the response.

Test statistic value: $\mathrm{f} = \frac{SSR\ /\ p}{SSE\ /\ [(n-(p+1)]} = \frac{MSR}{MSE}$, which under H_o, $\mathrm{f} \sim \mathrm{F}[p,\ n-(p+1)]$

f can be seen as the ratio of the amount of variation explained divided by the amount of variation unexplained under the current linear model of use. Rejecting H_o will allow us to conclude that there is at least one explanatory variable having useful relationship with the response.

t-test for Regression Slope

t-test for regression slope asserts whether a specific explanatory variable is useful in predicting the response. Its parameter of interest is the observed variable's coefficient in the model. It is conducted after model utility to find out which specific set of explanatory variables are useful in predicting the response in the linear regression model. We want to test the following hypotheses:

$$H_o: \beta_j = 0\ (j = 1, 2, \ldots, p)$$
$$H_a: \beta_j \neq 0$$

Test statistic value: $\mathrm{t} = \frac{\beta_j - \widehat{\beta}_j}{\sqrt{s^2 + s_{\hat{Y}}^2}} \sim t[\mathrm{df} = \mathrm{n} - (\mathrm{p} + 1)]$

Rejecting H_o will allow us to conclude that the observed variable is useful in predicting the response.

Cook's Distance

Cook's distance D_i is a measure for identifying influential data points. Calculating Cook's distance involves deleting each observation one at a time, and refitting the regression model on the rest observations (Simon & Young, n.d.). We then observe the difference between fitted values of the model using all observations and the model with the i^{th} observation deleted. Such difference can summarize how much influence the i^{th} observation has on the analysis. We can compute Cook distance's D_i using the formula

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

where h_{ii} is the i^{th} observation's leverage. There are some guidelines on how to interpret a Cook's distance D_i. If $D_i > 0.5$, the i^{th} observation may be influential. If $D_i > 1$, the i^{th} observation is quite likely to be influential.

Variance Inflation Factor

The variance inflation factor (VIF) is used to measure a variable's problematic collinearity. We usually concern when VIF value is greater than 10. The VIF of a variable is the ratio of its coefficient estimate's variance in the full model divided by the same quantity in the model where it is the only predictor (James, Witten, Hastie, & Tibshirani, 2014, pp. 101-102). In a mathematically simpler form, it is

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R-squared from a regression of X_j onto all of the other predictors.

Forward Stepwise Selection

"Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model" (James, Witten, Hastie, & Tibshirani, 2014, p.207). Below is the algorithm for forward stepwise selection.

Forward Stepwise selection:

1. Let Mo denote the null model, which contains no predictors
2. For k=1,...,p-1:
    a. Consider all p-k models that augment the predictors in Mk with one additional predictor
    b. Choose the best among these p-k models, and call it Mk+1. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a best single model form M0,....,Mp using cross-validated prediction error, Cp (AIC), BIC, or adjusted R^2

Cross-Validation

Cross-validation is a resampling method to estimate the test error of a model when it predicts the response on new data. It is used when we want to evaluate the model's predictive power without a designated dataset for testing. Cross-validation randomly divides observations into k groups, or folds, of approximately equal size (James *et al.,* 2014, pp. 181-183). Each fold will be sequentially held out, and treated as a testing dataset for the model established using observations from the remaining $\mathrm{k} - 1$ folds. In each time doing so, the mean squared error is also recorded. The, the test error's estimate, called $CV_{(k)}$ will be

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

The choice of k is regulated by the bias-variance trade-off (James *et al.,* 2014, pp. 183-184). When k grows large to the original data's size, the training model is fitted with a larger amount of data, so we will have an approximately unbiased estimates of the test data. However,

there will be high a high amount of variance associated with it. When k shrinks to 1, we will have much lower variance in prediction, but our estimates are biased.

The Lasso regression

The lasso utilizes a different approach from ordinary least squares. The lasso coefficients, $\hat{\beta}_R^L$, minimize the quantity (James *et al.,* 2014, pp. 219) :

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{1}^{p}|\beta_j|$$

The lasso forces the coefficients associated with unnecessary variables to be 0, so it performs variable selection.

4. Results

Objective 1

For this objective, we established a two-way ANOVA model

$$\mu_{Value} = \mu_{Continent} + \mu_{Attacker.Or.Defender} + \gamma + \varepsilon \qquad (1)$$

where γ represents the interaction effect between a player's continent of origin and his preferred position. From Table 1a, individual F-tests show significant p-values for **Continent** and **Attacker.Or.Defender**, but not for their interaction effect. That corresponds to the interaction plot in Figure 1 where three lines are very close to parallel. For that reason, we can disregard γ, the interaction effect in the model, and conclude that on average, the value of attackers, defenders or those who play in both positions, does not differ significantly base on their continents of origin.

After dropping the interaction term, we wanted to check the assumptions for the model

$$\mu_{Value} = \mu_{Continent} + \mu_{Attacker.Or.Defender} + \varepsilon \qquad (2)$$

Two assumptions of residuals, normality and homogeneity of variances, are not satisfied. First, there is a curvy departure of dots at the right end in Figure 2a. Second, stripes of dots in Figure 3a increase for larger fitted values, suggesting that variances are not constant. To resolve those violations, we investigated a Box-Cox transformation and found out the optimal lambda $\lambda = -0.1010101$. This indicates that a natural log transformation is appropriate. Looking at Figure 2b and 3b, log transformation definitely relaxes the assumptions of residuals better.

From Table 1b, the two main factors are both significant as the p-values associated with their F-tests are a lot less 1%. However, among them, continent of origin is a more significant factor than preferred position. If we take a look at Figure 1 again, a player's value does not differ much base on their preferred position. In contrast, we can observe a clear difference in value base on their continents of origin.

The Tukey's procedure in Table 2 reveals us two things. First, the values of players coming from different continents will differ significantly from each other, except for Africa vs. South America and Asia vs. Oceania. Second, different preferred positions will result in different values. Furthermore, by Figure 4, it is worth noting that if a player comes from Africa or South America, or if he is at least an attacker, he is likely to have a value higher than the average.

Objective 2

We started with a full model in which all possible predictors are taken into account. For convenience, we presented the model as an R formula object

Value ~ Age + Continent + Overall + Potential + Acceleration + Aggression + Agility +

Balance + Ball.control + Composure + Crossing + Curve + Dribbling + Finishing + Free.kick.accuracy + Heading.accuracy + Interceptions + Jumping + Long.passing + Long.shots + Marking + Penalties + Positioning + Reactions + Short.passing + Shot.power + Sliding.tackle + Sprint.speed + Stamina + Standing.tackle + Strength + Vision + Volleys + Attacker.Or.Defender (3)

where the first level in increasing alphabetical order will be the baseline when creating dummy variables. That is, “Africa” for **Continent** and “Attacker” for **Attacker.Or.Defender**. A summary of this full model is displayed in Table 4. Most individual t-tests result in significant p-values, meaning many of the observed predictors have useful relationships with the response. However, this model seems to fit the data poorly; Multiple R-squared = 0.4896 and Adjusted R-squared = 0.4884. Additionally, the residuals of this model are not normally distributed (Figure 5a), and independent of the fitted values.

For that reason, we investigated a Box-Cox transformation and found out the optimal lambda $\lambda = -0.06060606$. This indicates that a natural log transformation is appropriate. Looking at Figure 5b and 6b, log transformation definitely relaxes the assumptions of residuals better. At this point, from Table 5, we can tell that the model with log transformation on **Value** fits the data very ; Multiple R-squared = 0.9745 and Adjusted R-squared = 0.9744. However, it is worth noting that by individual t-tests, not all predictors are indicated to have significant relationships with the response. Hence, we suspected the presence of multicollinearity in the model. Later, we wished to perform variable selection for the model, so removing multicollinearity is necessary to make sure that coefficients estimates are reliable.

To resolve multicollinearity, we chose to remove predictors that are highly correlated with at least one of the rest predictors. We followed this algorithm. Under the full model after log transformation on **Value**, we calculate each involved predictor’s variance inflation factor. We remove the predictor with highest variance inflation factor that is greater than 10. Then, we refit the model without that variable, and see if all variance inflation factors are less than 10. If yes, we will remove the predictor with highest variance inflation factor and refit the model. The process is repeated until there are no variance inflation factors that are greater than 10. By doing so, in the end, the predictors we had removed, one by one, are **Standing.tackle**, **Overall** and **Marking**. Table 6 shows all variance inflation factors after removing those predictors. Now, we can see that predictors are more significant to predicting the response (Table 7).

One lucky thing happened is that although we have many outliers (see Figure 6b, many points have standardized residual greater than 3 or less than -3), they are not influential to our model. That can be verified in Figure 7 where even the highest Cook’s distance is less than 0.01.

Lastly, we applied forward selection on the most updated model. Three stopping criteria we chose were maximum adjusted R-squared, minimum Cp and minimum BIC. At the same time, we ran a Lasso regression with predictors in model (3) after a log transformation on the response. We ended up with four different models. Table 8 summarizes the predictors selected by those four models. The model recommended by maximum adjusted R-squared criterion comes with the most variables, while the model recommended by Lasso regression comes with the least variables. To evaluate predictive power, we performed 30-fold cross-validation 10 times on the models recommended by maximum adjusted R-squared, minimum Cp and minimum BIC. The package **glmnet** in R that we used for the Lasso regression has its own function to do cross-validation. In the end, the Lasso regression model has the lowest mean squared error 0.04927. The other models have mean squared errors around 0.2, in which the model recommended by maximum adjusted R-squared criterion is the lowest, and the model recommended by minimum BIC criterion is the highest (Figure 8).

5. Conclusion

The analysis told us that a player with high value is usually at least an attacker; or comes from Africa and South America. That is not a big surprise in real life. Attackers are those who bring victory to the team, so it is easy to understand that they will get more recognitions and rewards. Africa and South are well-known for their investments in soccer, so their native players get trained intensively, and possess many desirable skill sets. Those factors can tell us some insights on how their values are high.

In terms of predicting a player's value, it is clear that the lasso regression has the highest predictive power. However, as the lasso regression allows its coefficient estimates to be biased, inference about the response-predictor relationship cannot be made. The model with balance between prediction and inference would be the one recommended by the minimum BIC criterion, since it does not include so many predictors, while its mean squared error is not bad compared to the other two stopping criterion in forward selection.

In conclusion, if all we care about is a numeric answer for value, we should use the one in Table 9. If we want to see the relationship between the response and some important predictors, we should use the one in Table 10.

6. References

A., F. J. (1946). Sequential Analysis of Statistical Data: Applications. *Journal of the Royal Statistical Society,109*(4), 505. doi:10.2307/2981341

Devore, J. L. (2012). *Probability and Statistics for Engineering and the Sciences* (8th ed.). Australia: Brooks/Cole, Cengage Learning.

EA SPORTS FIFA Soccer Franchise Sales Top 100 Million Units Lifetime. (2010, November 04). Retrieved May 8, 2018, from https://www.businesswire.com/news/home/20101104006782/en

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: With applications in R*. New York: Springer.

Seal, H. L. (1967). Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. *Biometrika,54*(1-2), 1-24. D oi:10.1093/biomet/54.1-2.1

Simon, L., & Young, D. (n.d.). Identifying Influential Data Points (I. Pardoe, Ed.). Retrieved May 6, 2018, from https://onlinecourses.science.psu.edu/stat501/node/340

7. Appendix

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Continent	5	4966020420910069	993204084182014	33.395	<0.0001 ***
Attacker.Or.Defender	2	2378756376779254	1189378188389627	39.991	<0.0001 ***
Continent:Attacker.Or.Defender	10	345242814771866	34524281477187	1.161	0.312
Residuals	15717	467436017138217408	29740791317568		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1a. Two-way ANOVA table for model (1).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Continent	5	1192	238.37	130.62	<0.0001 ***
Attacker.Or.Defender	2	253	126.57	69.36	<0.0001 ***
Residuals	15727	28699	1.82		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1b. Two-way ANOVA table for model (2) after log transformation.

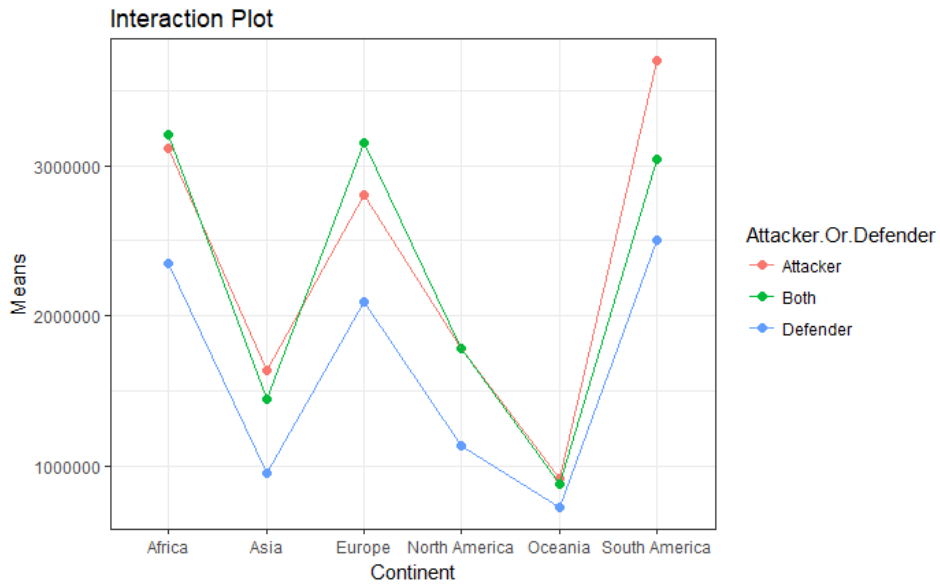

Figure 1. Interaction plot of **Continent** and **Attacker.Or.Defender**.

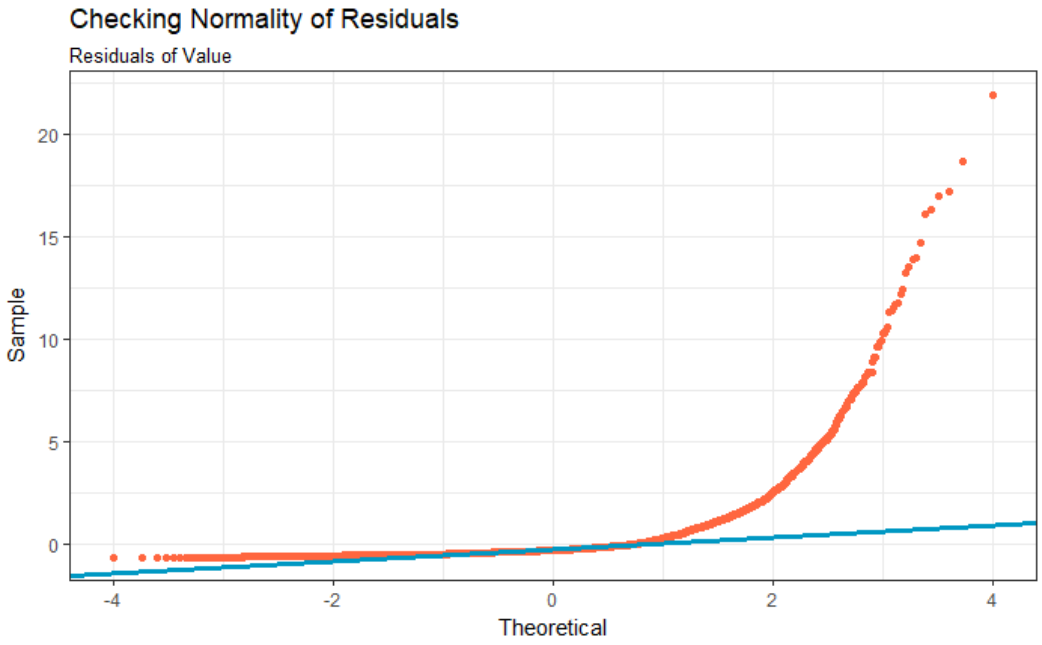

Figure 2a. QQ-plots of model (2) before log transformation.

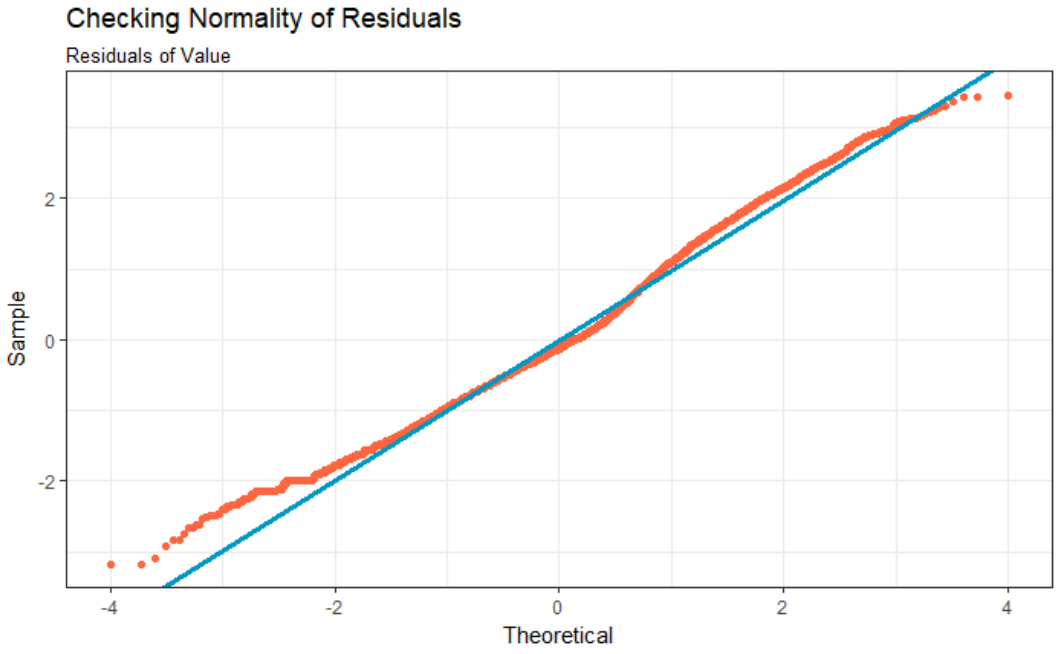

Figure 2b. QQ-plots of model (2) after log transformation.

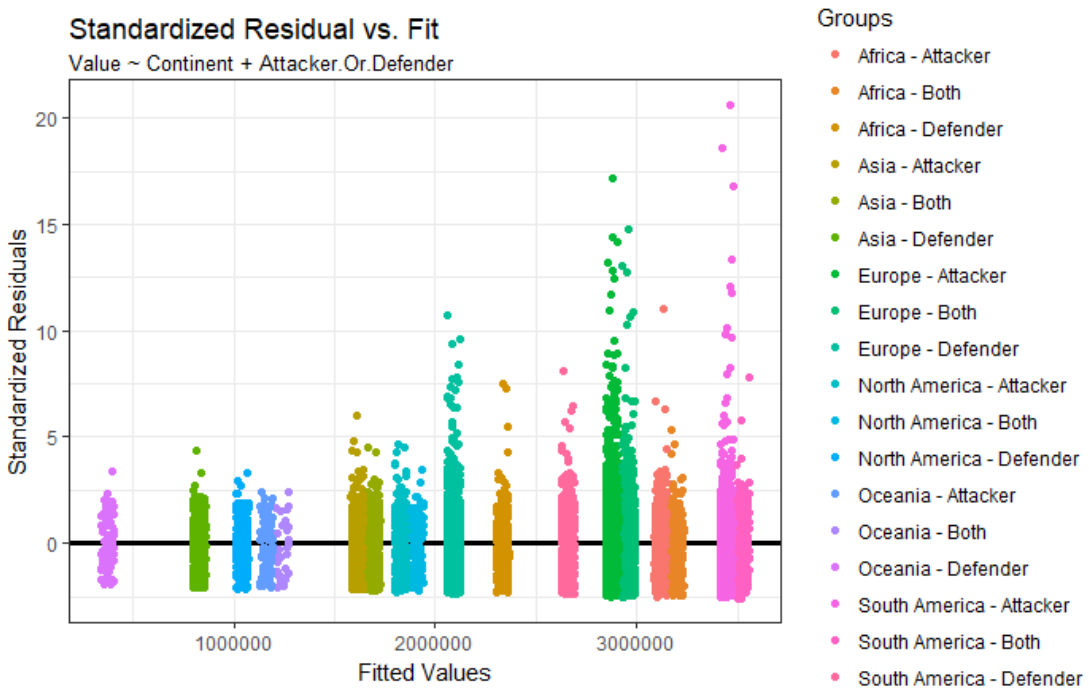

Figure 3a. Residual plots of model (2) before log transformation.

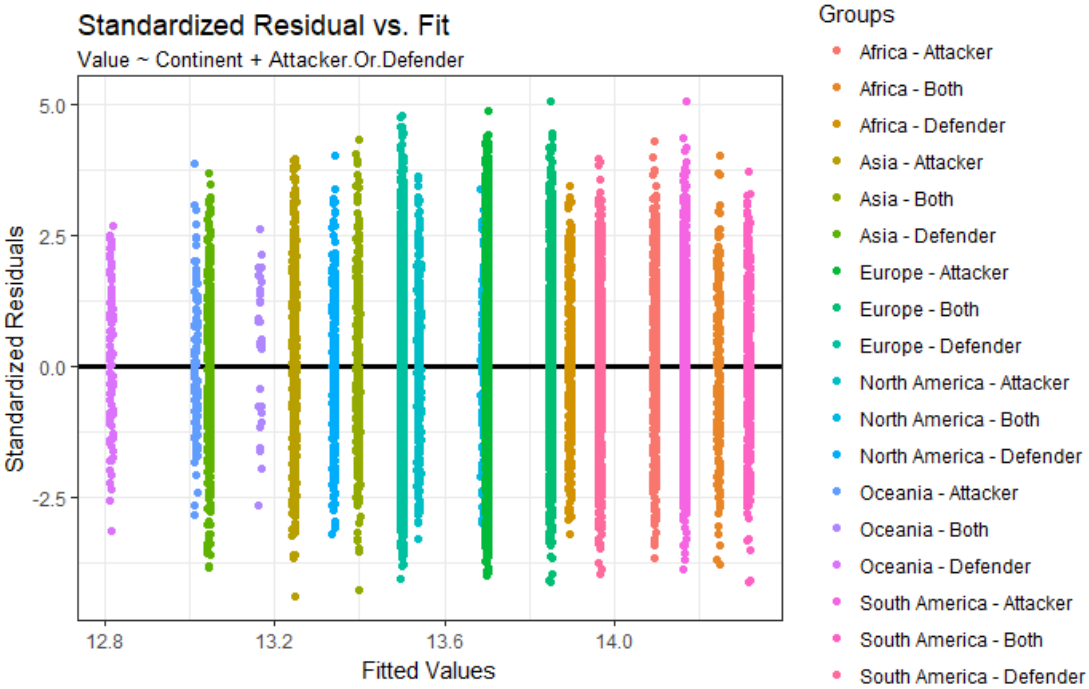

Figure 3b. Residual plots of model (2) after log transformation.

Tukey multiple comparisons of means

95% family-wise confidence level

Continent of Origin

	diff	lwr	upr	p adj
Asia-Africa	-0.85880461	-1.01000502	-0.70760421	0.0000000
Europe-Africa	-0.41433421	-0.53677411	-0.29189430	0.0000000
North America-Africa	-0.58571741	-0.76450523	-0.40692958	0.0000000
Oceania-Africa	-1.11971867	-1.40178868	-0.83764866	0.0000000
South America-Africa	0.04745507	-0.08933851	0.18424865	0.9217059
Europe-Asia	0.44447041	0.33924259	0.54969822	0.0000000
North America-Asia	0.27308721	0.10561646	0.44055795	0.0000498
Oceania-Asia	-0.26091406	-0.53595016	0.01412204	0.0745115
South America-Asia	0.90625968	0.78462972	1.02788964	0.0000000
North America-Europe	-0.17138320	-0.31342692	-0.02933948	0.0077219
Oceania-Europe	-0.70538446	-0.96572044	-0.44504848	0.0000000
South America-Europe	0.46178928	0.37857982	0.54499874	0.0000000
Oceania-North America	-0.53400126	-0.82511701	-0.24288551	0.0000026
South America-North America	0.63317248	0.47858473	0.78776022	0.0000000
South America-Oceania	1.16717374	0.89978680	1.43456068	0.0000000

Preferred Position (Attacker or Defender)

	diff	lwr	upr	p adj
Both-Attacker	0.1494052	0.0782050	0.2206054	0.0000027
Defender-Attacker	-0.1998996	-0.2558721	-0.1439271	0.0000000
Defender-Both	-0.3493048	-0.4233982	-0.2752114	0.0000000

Table 2. Tukey's procedure for model (2).

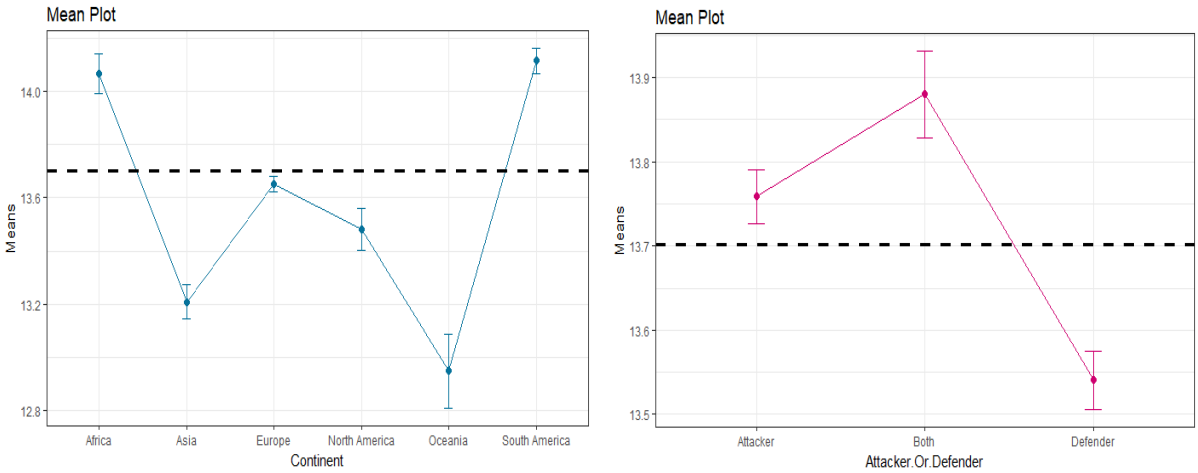

Figure 4. Mean plots for **Continent** and **Attacker.Or.Defender**. Each error bar represents its corresponding mean's 95% confidence interval.

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*	
(Intercept)	-33210521	705265	-47.089	< 0.0000000000000002	***
Age	-251388	15437	-16.285	< 0.0000000000000002	***
ContinentAsia	744996	159251	4.678	0.00000291941129	***
ContinentEurope	655000	127819	5.124	0.00000030195657	***
ContinentNorth America	85130	185187	0.460	0.645740	
ContinentOceania	837869	291440	2.875	0.004047	**
ContinentSouth America	-68010	143299	-0.475	0.635078	
Overall	561947	17022	33.013	< 0.0000000000000002	***
Potential	84462	13029	6.483	0.00000000009282	***
Acceleration	-1152	6136	-0.188	0.851023	
Aggression	-11015	3367	-3.272	0.001070	**
Agility	-15808	4814	-3.283	0.001028	**
Balance	10599	4396	2.411	0.015924	*
Ball.control	-57099	8222	-6.944	0.00000000000395	***
Composure	8330	5517	1.510	0.131063	
Crossing	-11193	4146	-2.700	0.006946	**
Curve	-4564	4273	-1.068	0.285496	
Dribbling	-1031	6295	-0.164	0.869919	
Finishing	10836	5033	2.153	0.031336	*
Free.kick.accuracy	8651	3799	2.277	0.022794	*
Heading.accuracy	-13286	4510	-2.946	0.003222	**

```
Interceptions                   1616       4828   0.335             0.737813
Jumping                         4027       3273   1.230             0.218561
Long.passing                   10902       5512   1.978             0.047943 *
Long.shots                    -16158       4653  -3.472             0.000517 ***
Marking                       -60601       6109  -9.920 < 0.0000000000000002 ***
Penalties                       4972       4253   1.169             0.242496
Positioning                     8240       4720   1.746             0.080824 .
Reactions                      62629       6926   9.043 < 0.0000000000000002 ***
Short.passing                 -23690       7739  -3.061             0.002209 **
Shot.power                    -19192       4529  -4.238     0.00002269763559 ***
Sliding.tackle                 29243       6925   4.223     0.00002425052169 ***
Sprint.speed                   -5723       5692  -1.005             0.314763
Stamina                        -6799       3758  -1.809             0.070418 .
Standing.tackle                17365       7178   2.419             0.015567 *
Strength                       -7700       4201  -1.833             0.066848 .
Vision                         28377       5054   5.615     0.00000001998046 ***
Volleys                        26404       4329   6.100     0.00000000108702 ***
Attacker.Or.DefenderBoth       98451     127375   0.773             0.439580
Attacker.Or.DefenderDefender  198330     151030   1.313             0.189142
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3931000 on 15695 degrees of freedom
Multiple R-squared:  0.4897, Adjusted R-squared:  0.4884
F-statistic: 386.1 on 39 and 15695 DF,  p-value: < 0.00000000000000022
```

Table 4. Summary table for model (3).

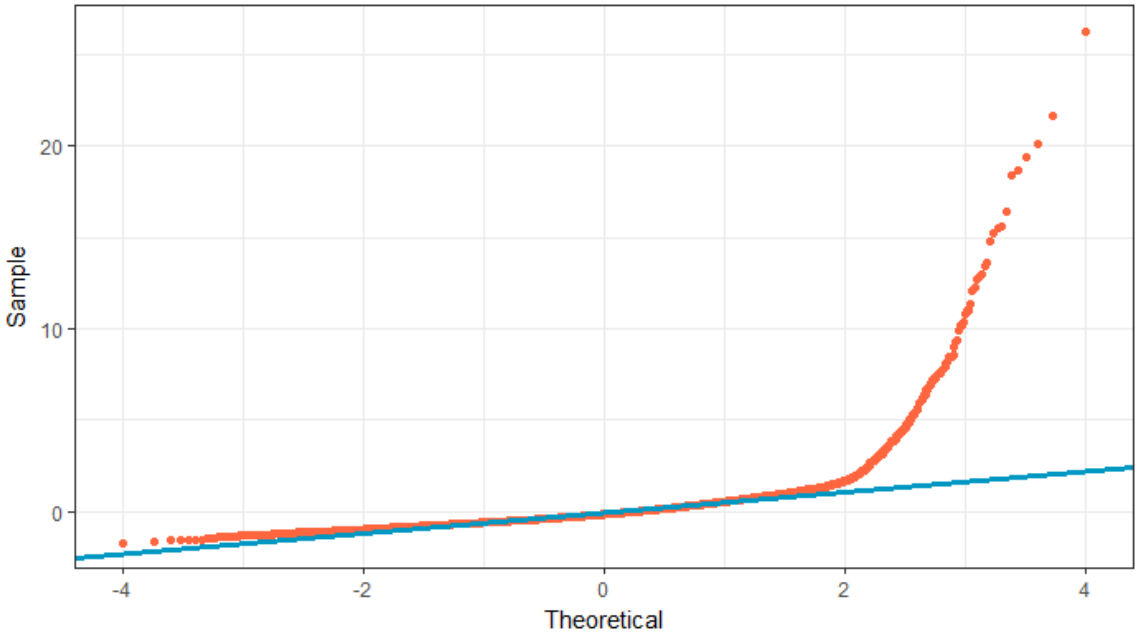

Figure 5a. QQ-plots of residuals model (3) before log transformation.

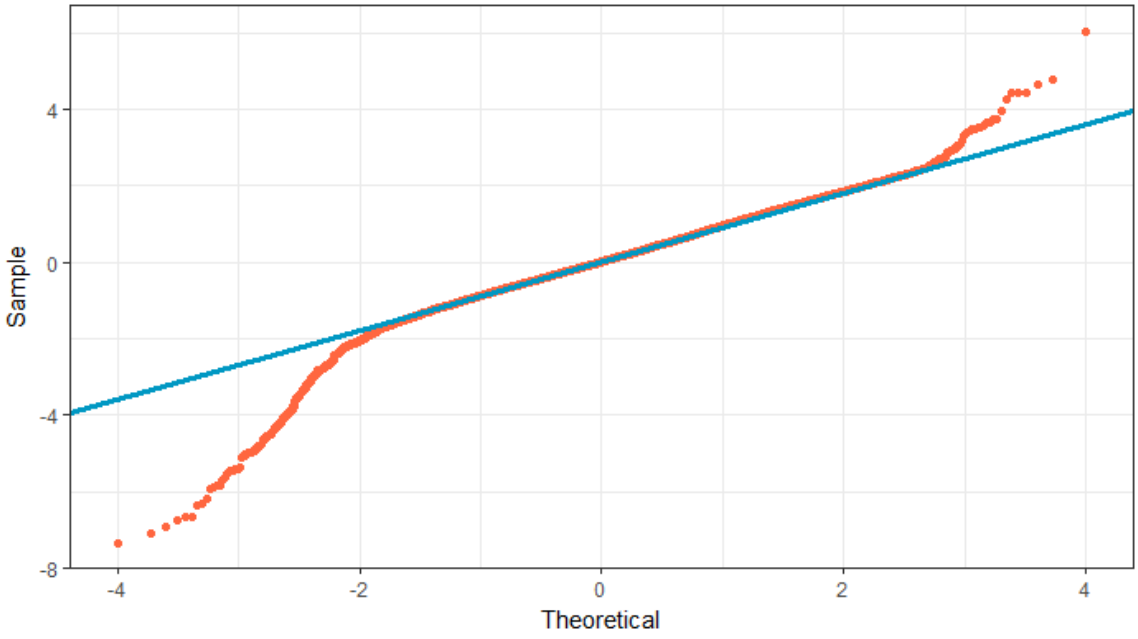

Figure 5b. QQ-plots of residuals in model (3) after log transformation.

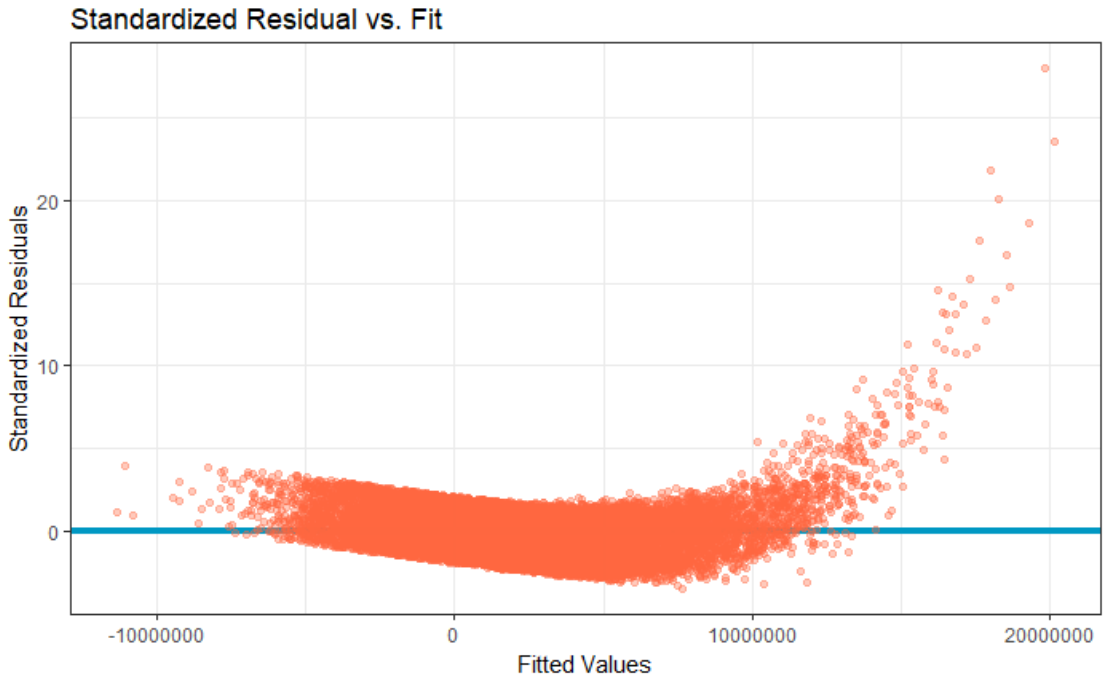

Figure 6a. Residual plots of model (3) before log transformation.

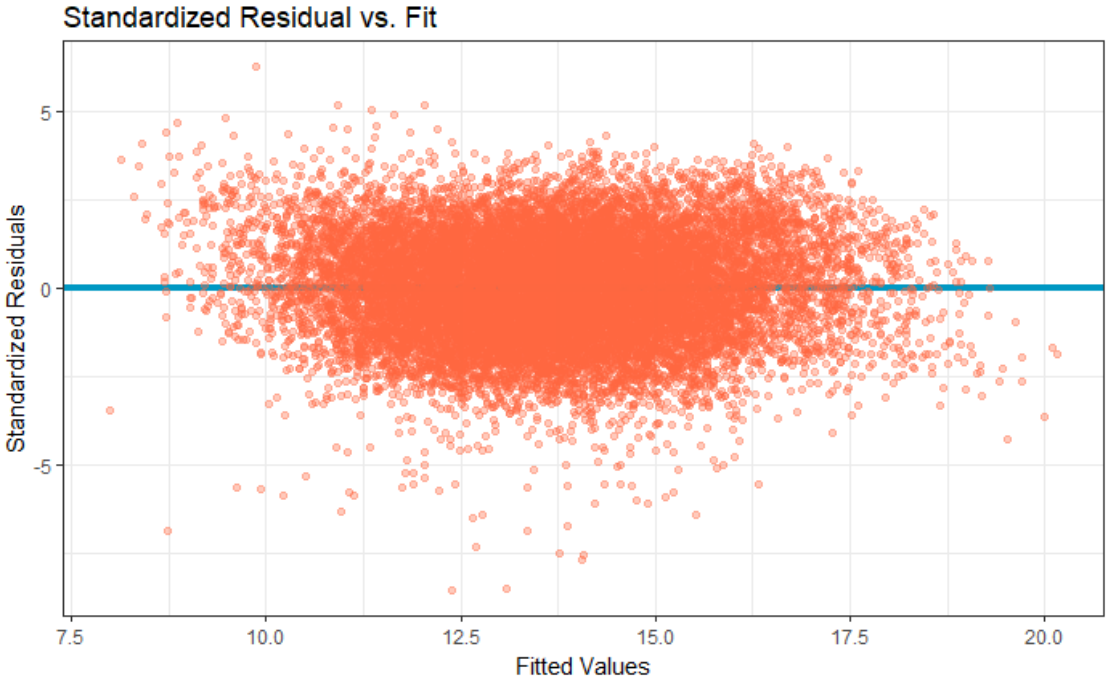

Figure 6b. Residual plots of model (3) after log transformation.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.38053721	0.03973103	34.747	< 0.0000000000000002	***
Age	-0.08536196	0.00086963	-98.159	< 0.0000000000000002	***
ContinentAsia	0.00200813	0.00897141	0.224	0.822887	
ContinentEurope	-0.00474625	0.00720065	-0.659	0.509815	
ContinentNorth America	-0.05509221	0.01043250	-5.281	0.000000130317939	***
ContinentOceania	-0.03610050	0.01641826	-2.199	0.027907	*
ContinentSouth America	-0.01053240	0.00807272	-1.305	0.192017	
Overall	0.21238793	0.00095893	221.484	< 0.0000000000000002	***
Potential	0.00349271	0.00073398	4.759	0.000001966860395	***
Acceleration	0.00013360	0.00034565	0.387	0.699125	
Aggression	-0.00007540	0.00018966	-0.398	0.690975	
Agility	-0.00082806	0.00027122	-3.053	0.002269	**
Balance	-0.00078127	0.00024766	-3.155	0.001610	**
Ball.control	-0.00195845	0.00046320	-4.228	0.000023699431200	***
Composure	-0.00014313	0.00031079	-0.461	0.645123	
Crossing	-0.00148496	0.00023355	-6.358	0.000000000209736	***
Curve	0.00009142	0.00024072	0.380	0.704118	
Dribbling	0.00014780	0.00035463	0.417	0.676849	
Finishing	0.00045871	0.00028352	1.618	0.105705	
Free.kick.accuracy	0.00084757	0.00021402	3.960	0.000075205939023	***
Heading.accuracy	0.00071521	0.00025405	2.815	0.004881	**
Interceptions	0.00045161	0.00027200	1.660	0.096865	.
Jumping	0.00012689	0.00018438	0.688	0.491333	
Long.passing	0.00114121	0.00031050	3.675	0.000238	***
Long.shots	-0.00092651	0.00026215	-3.534	0.000410	***
Marking	-0.00245723	0.00034414	-7.140	0.000000000000973	***
Penalties	0.00030851	0.00023962	1.288	0.197936	
Positioning	0.00067850	0.00026588	2.552	0.010722	*
Reactions	0.00225590	0.00039017	5.782	0.000000007531662	***
Short.passing	-0.00084055	0.00043596	-1.928	0.053871	.
Shot.power	-0.00023724	0.00025513	-0.930	0.352448	
Sliding.tackle	0.00035948	0.00039011	0.921	0.356808	
Sprint.speed	0.00123835	0.00032069	3.862	0.000113	***
Stamina	0.00229120	0.00021170	10.823	< 0.0000000000000002	***
Standing.tackle	0.00094965	0.00040437	2.348	0.018864	*

```
Strength                      -0.00066311  0.00023667  -2.802             0.005088 **
Vision                         0.00050883  0.00028470   1.787             0.073911 .
Volleys                        0.00070599  0.00024385   2.895             0.003795 **
Attacker.Or.DefenderBoth      -0.09754733  0.00717567 -13.594 < 0.0000000000000002 ***
Attacker.Or.DefenderDefender -0.17118316  0.00850827 -20.120 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2214 on 15695 degrees of freedom
Multiple R-squared:  0.9745, Adjusted R-squared:  0.9744
F-statistic: 1.536e+04 on 39 and 15695 DF,  p-value: < 0.00000000000000022
```

Table 5. Summary table for model (3) after log transformation.

	Variables	*Tolerance*	*VIF*
1	Age	0.3012548	3.319449
2	ContinentAsia	0.4325411	2.311919
3	ContinentEurope	0.2483052	4.027303
4	ContinentNorth America	0.5968754	1.675392
5	ContinentOceania	0.8238762	1.213775
6	ContinentSouth America	0.3292086	3.037588
7	Potential	0.2707366	3.693627
8	Value	0.5487373	1.822366
9	Acceleration	0.1863406	5.366518
10	Aggression	0.4041329	2.474434
11	Agility	0.2711331	3.688225
12	Balance	0.3417154	2.926412
13	Ball.control	0.1429537	6.995273
14	Composure	0.3277777	3.050848
15	Crossing	0.2830430	3.533032
16	Curve	0.2332162	4.287868
17	Dribbling	0.1526689	6.550121
18	Finishing	0.1443715	6.926576
19	Free.kick.accuracy	0.2978833	3.357019
20	Heading.accuracy	0.3688663	2.711009

21	Interceptions	0.1384070	7.225070
22	Jumping	0.6725937	1.486782
23	Long.passing	0.2026585	4.934410
24	Long.shots	0.1810923	5.522047
25	Penalties	0.3424898	2.919795
26	Positioning	0.2056045	4.863706
27	Reactions	0.2760440	3.622611
28	Short.passing	0.1629213	6.137934
29	Shot.power	0.2704380	3.697705
30	Sliding.tackle	0.1343775	7.441725
31	Sprint.speed	0.2277241	4.391279
32	Stamina	0.5554790	1.800248
33	Strength	0.3526330	2.835809
34	Vision	0.2266827	4.411453
35	Volleys	0.2369574	4.220168
36	Attacker.Or.DefenderBoth	0.4383741	2.281156
37	Attacker.Or.DefenderDefender	0.1939574	5.155771

Table 6. Variance inflation factors of all predictor after removing **Standing.tackle**, **Overall** and **Marking** one by one.

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*	
(Intercept)	-1.65668375	0.07582567	-21.849	< 0.0000000000000002	***
Age	0.02326542	0.00146007	15.934	< 0.0000000000000002	***
ContinentAsia	0.00682713	0.01823213	0.374	0.708070	
ContinentEurope	-0.05367649	0.01463448	-3.668	0.000245	***
ContinentNorth America	-0.10055396	0.02118209	-4.747	0.000002081537524	***
ContinentOceania	-0.06070564	0.03338854	-1.818	0.069059	.
ContinentSouth America	0.01414457	0.01639704	0.863	0.388354	
Potential	0.11441709	0.00109334	104.649	< 0.0000000000000002	***
Acceleration	0.00708189	0.00069992	10.118	< 0.0000000000000002	***
Aggression	0.00034130	0.00038519	0.886	0.375603	
Agility	-0.00063147	0.00055155	-1.145	0.252271	
Balance	-0.00324108	0.00050324	-6.440	0.000000000122628	***
Ball.control	0.01702766	0.00092455	18.417	< 0.0000000000000002	***
Composure	0.01287497	0.00062027	20.757	< 0.0000000000000002	***

```
Crossing                       0.00106495  0.00047422   2.246             0.024737 *
Curve                          0.00005368  0.00048967   0.110             0.912709
Dribbling                     -0.00082192  0.00072068  -1.140             0.254101
Finishing                      0.00554228  0.00057474   9.643 < 0.0000000000000002 ***
Free.kick.accuracy             0.00224154  0.00043492   5.154    0.000000258124057 ***
Heading.accuracy               0.01000103  0.00050810  19.683 < 0.0000000000000002 ***
Interceptions                  0.00316286  0.00050989   6.203    0.000000000567985 ***
Jumping                        0.00044782  0.00037462   1.195             0.231942
Long.passing                  -0.00056037  0.00063049  -0.889             0.374129
Long.shots                    -0.00163273  0.00053320  -3.062             0.002202 **
Penalties                     -0.00051474  0.00048715  -1.057             0.290701
Positioning                   -0.00394215  0.00053755  -7.334    0.000000000000235 ***
Reactions                      0.02972381  0.00075209  39.522 < 0.0000000000000002 ***
Short.passing                  0.01170214  0.00087887  13.315 < 0.0000000000000002 ***
Shot.power                     0.00133236  0.00051858   2.569             0.010200 *
Sliding.tackle                -0.00138060  0.00050396  -2.739             0.006160 **
Sprint.speed                   0.00611084  0.00065061   9.392 < 0.0000000000000002 ***
Stamina                        0.00789385  0.00042733  18.472 < 0.0000000000000002 ***
Strength                       0.00597146  0.00047708  12.517 < 0.0000000000000002 ***
Vision                        -0.00325782  0.00057749  -5.641    0.000000017161467 ***
Volleys                        0.00087386  0.00049559   1.763             0.077877 .
Attacker.Or.DefenderBoth      -0.17846605  0.01438244 -12.409 < 0.0000000000000002 ***
Attacker.Or.DefenderDefender -0.02843571  0.01699706  -1.673             0.094351 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4504 on 15698 degrees of freedom
Multiple R-squared:  0.8943, Adjusted R-squared:  0.8941
F-statistic:  3691 on 36 and 15698 DF,  p-value: < 0.00000000000000022
```

Table 7. Summary table of model (3) after log transformation and removing multicollinearity

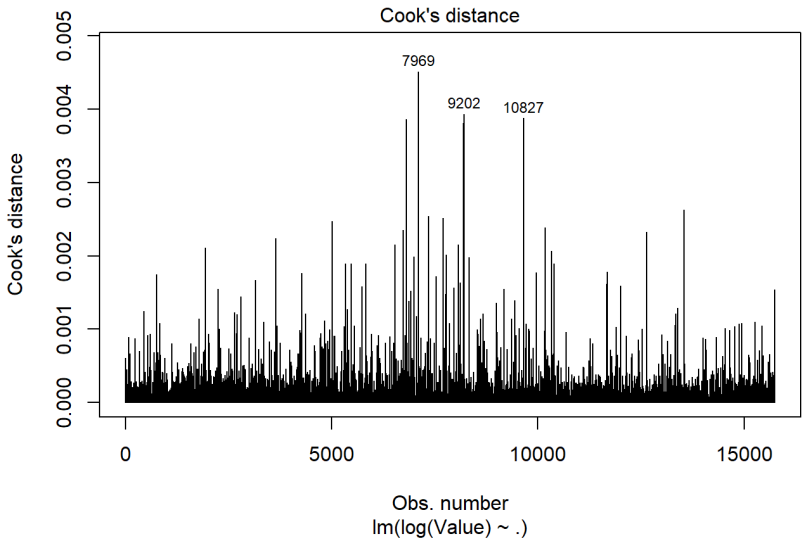

Figure 7. Cook's distance plot of model (3) after log transformation and removing multicollinearity.

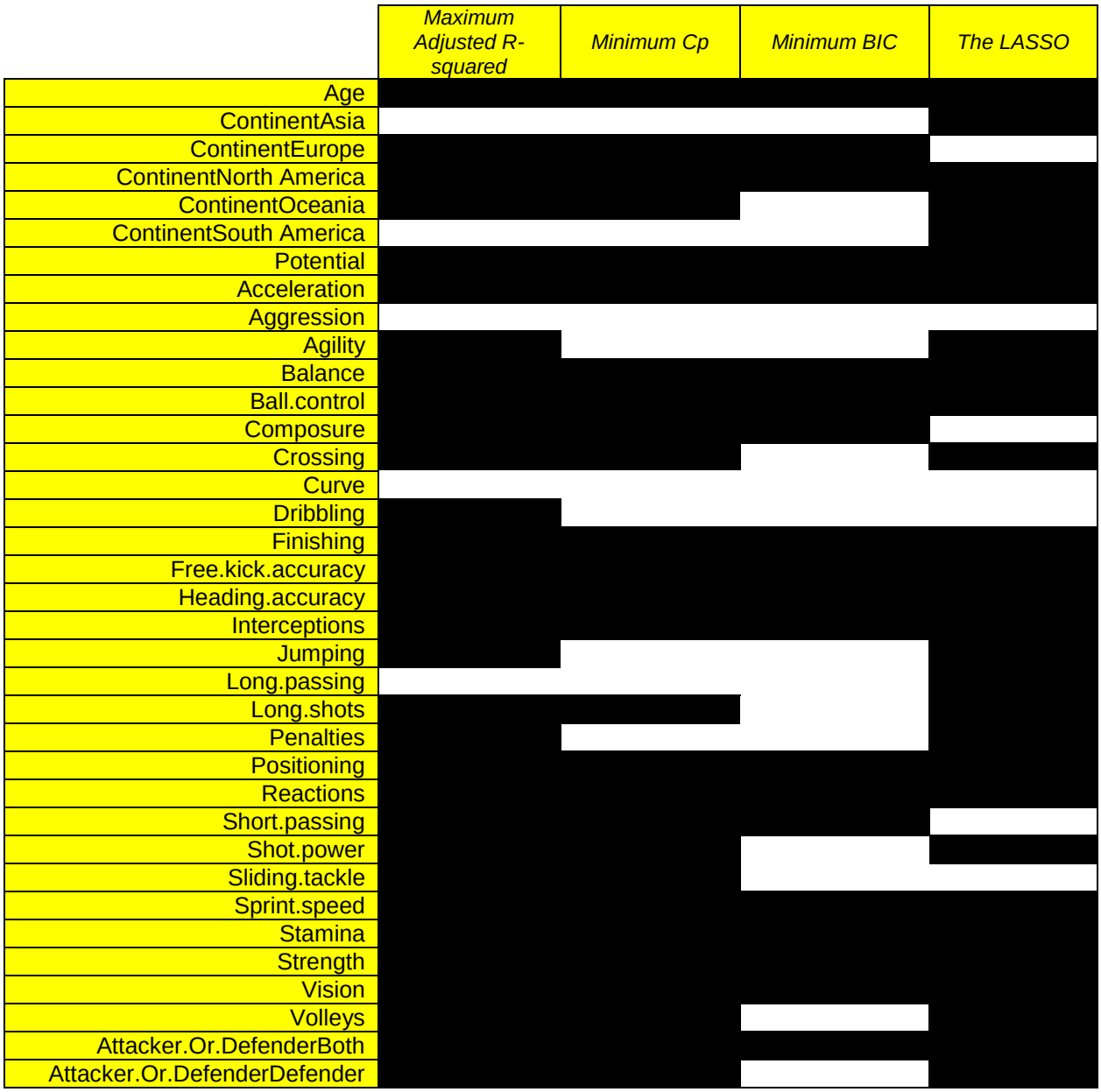

Table 8. Predictors selected by different criteria in forward selection and the LASSO.

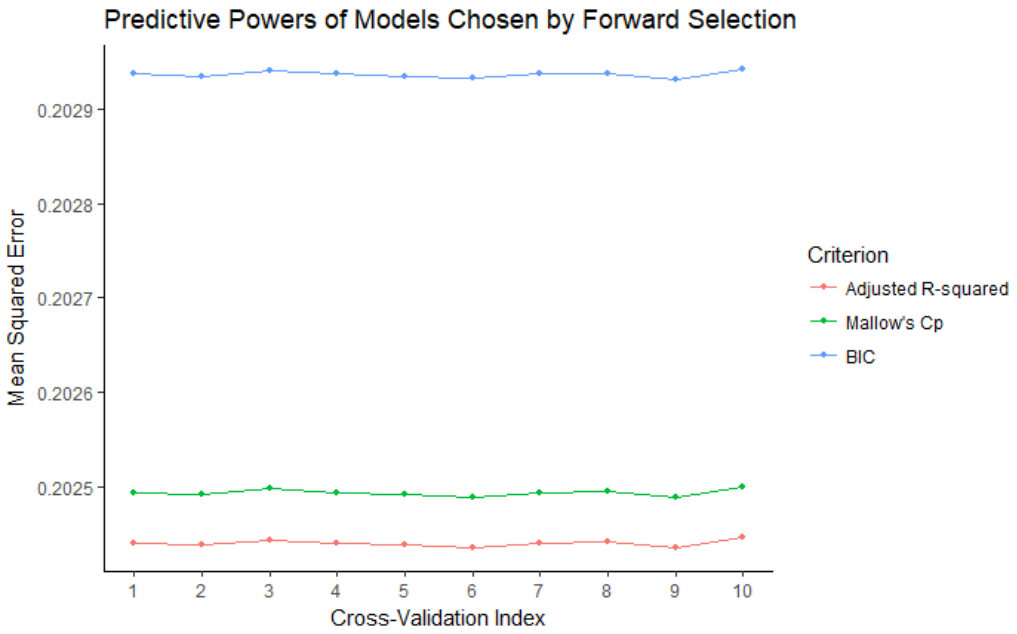

Figure 8. Predictive powers of models suggested by different criteria in forward selection.

Predictors	*Coefficient Estimates*
(Intercept)	1.339032222067
Age	-0.084241874843
ContinentAsia	0.003101426346
ContinentEurope	.
ContinentNorth America	-0.047549721531
ContinentOceania	-0.023515212014
ContinentSouth America	-0.002673320785
Overall	0.210404086087
Potential	0.004451565031
Acceleration	0.000005864430
Aggression	.
Agility	-0.000519484468
Balance	-0.000544435209
Ball.control	-0.001349553241
Composure	.
Crossing	-0.001206428013
Curve	.
Dribbling	.
Finishing	0.000286227587
Free.kick.accuracy	0.000723339526
Heading.accuracy	0.000597185409
Interceptions	0.000014585666
Jumping	0.000000199598
Long.passing	0.000531227803
Long.shots	-0.000503180947
Marking	-0.001228076787
Penalties	0.000171001358
Positioning	0.000500529351
Reactions	0.002304856470
Short.passing	.
Shot.power	-0.000005193652
Sliding.tackle	.
Sprint.speed	0.001119715136
Stamina	0.002212695035
Standing.tackle	0.000170830442
Strength	-0.000276166641
Vision	0.000355540565
Volleys	0.000530432935
Attacker.Or.DefenderBoth	-0.087551802852
Attacker.Or.DefenderDefender	-0.158411559284

Table 9. The lasso regression model where log(Value) is the response.

Predictors	***Coefficient Estimates***
(Intercept)	-1.697611107
Age	0.023764282
ContinentEurope	-0.056108489
ContinentNorth America	-0.104985949
Potential	0.114248703
Acceleration	0.006887190
Balance	-0.003260253
Ball.control	0.016766223
Composure	0.012908630
Finishing	0.005814528
Free.kick.accuracy	0.002284793
Heading.accuracy	0.010238346
Interceptions	0.001909744
Positioning	-0.003684959
Reactions	0.030056290
Short.passing	0.011333386
Sprint.speed	0.006258156
Stamina	0.007805930
Strength	0.006197198
Vision	-0.003373002
Attacker.Or.DefenderBoth	-0.169096355

Table 10. The model recommended by the minimum BIC criterion in forward selection