Predict the Future Potential of Football Players

November 2021

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Statistics Project

MSBA 6120

Team members: Vivian Ho, Priyanka Rajendra Bhosale, Qujiaheng Zhang, Manish Jain

Index

List of Tables

Executive summary

A. Short Summary

Fédération Internationale de Football Association(FIFA) creates potential ratings of players every year based on various player attributes and performance.

Players' potential ratings will directly affect their market value, which makes improving potential ratings an important goal for all players and team managers.

However, the impact of different player attributes on the player's potential rating is unknown. Therefore, by building a regression model, we identified the key factors affecting the potential rating and proposed a plan to increase the potential.

B. Introduction

Our goal is to predict the player's potential and find out the important factors that affect the potential. We used FIFA 2015 data and selected data from the top leagues. We build a regression model based on the data on 3 major aspects, physical attributes, market value, and footballing skills. Physical attributes include age, pace, preferred foot, and physic. Market value includes international reputation and wage. Footballing skills include passing, defending, shooting, and dribbling.

C. Analysis

Player's age, wage, international reputation, and playing skills are important factors in deciding a player's potential. First of all, age has a negative impact on potential. Secondly, wages are the most critical factor, which, together with international reputation, has a positive impact on potential. Third, defense, dribbling ,and shooting are fundamental soccer skills, and increasing the scores of these skills can also increase the potential rating. Shooting and defending are complementary attributes on the court, which are also proven by the model.

D. Result and Conclusion

Based on the above analysis, we made two main recommendations. For older players, increasing players' salaries through short-term contracts can help boost players' market value. For young players, helping players improve their shooting and dribbling skills can help improve players' potential scores, thereby boosting their market value.

Main Body

A. Introduction

Fédération Internationale de Football Association(FIFA) in association with EA sports creates skill ratings for players every year based on physical tests and performance over the last season.

These metrics are created every year and affect the player's perception in the world of football. One such metric is the potential rating of a player, which signifies the highest possible overall rating, from 0 to 100, can a player achieve throughout his career. This metric is calculated for each player based on all the attributes of the player. As a player's potential signifies the future importance of a player, it plays an important part in determining the value of a player in the global market.

However, the impact of different player attributes on the player rating is unknown. If organizations such as football clubs and scouting organizations understood the importance of every player attribute that contributes to the potential rating, they could in-turn help the player increase it by improving on those important attributes.

The focus of the analysis is to analyze and create a statistical model to understand the relationship between potential and all other variables.

B. Understanding the Data

The data used for the analysis was created in 2015 September. The variables and attributes were measured and calculated based on the player's performance in the seasons before that time period. Many attributes of our analysis are calculated metrics in themselves, and hence are in a fixed range value. The variables are as defined below:

Variable	Type	Range	Description
Potential	Interval	0-99	Maximum possible overall rating that can be achieved in the future
Overall	Interval	0-99	Current overall rating of the player
Age	Interval	-	Player age
League Level	Ordinal	{1,2,3,4,5}	Level of the league player plays in ranked country wise
Wage	Interval	-	Current salary of the player in euros
Preferred Foot	Nominal	{'Left', 'Right'}	Preferred foot of the player

Pace	Interval	0-99	Rating of speed of running with the ball
Shooting	Interval	0-99	Rating of ability to shoot the ball
Passing	Interval	0-99	Rating of ability to pass the ball
Dribbling	Interval	0-99	Rating of ability to dribble with the ball
Defending	Interval	0-99	Rating of ability to defend
Physic	Interval	0-99	Rating of player’s physical strength
International Reputation	Ordinal	{1,2,3,4,5}	Popularity of the player

Table 1: Variable description

For our analysis, we have only selected players from the top teagues (League level = 1) in all countries as the data collection for them is more accurate due to higher commercial coverage of the leagues. This gives us a sample of 9,307 players.

The distribution of the target variable player potential is shown in *figure 1.* Majority of the players are in the potential range 67 - 68 as representing the quality of general players in the top leagues. The players on the bottom half are usually substitutes or reserve players with minimal first team opportunities. Players with higher potential are top players of the league with great promise or consistent performers. More detailed information of the distributions and exploratory analysis of predictors can be seen at the appendix A.

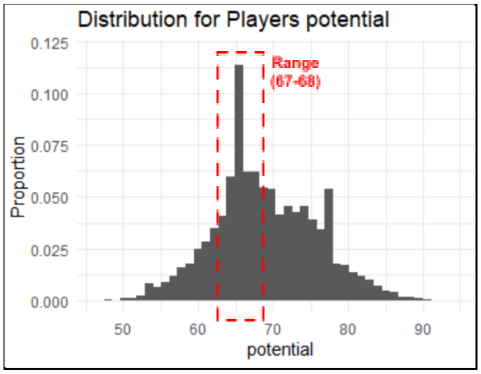


Figure 1

C. Finding and interpretation

Using a statistical model, (the model selection detail can be seen in appendix B) we were able to identify relationships between the player attributes and the potential rating that explained 80% of the change in the value of potential based on the factors available.

a. **Player Age** - Age has a negative relationship with their potential, which makes sense, because the younger players are, the more flexible they are and the more space for improving they have. On average, when a player’s age turns one year older, their potential decreases 0.7 units, keeping their wage, reputation, shooting, dribbling and defending ability fixed

b. **Player skills** - skills like shooting, defending, and dribbling, have all positive relationships with their potential. This is also easy to understand, because the potential measures a player’s ability to help a team win in future games, when comparing two players with the same profile, the player who has higher football skills will be considered as the one who has higher potential to the team. However, since these variables are highly related to each other (as shown in the appendix C), we cannot say how much exact impact a single variable has on the potential rating based on the model

c. **External valuations** - wage and international reputation also have positive relationships with potential rating, and among them wage is the most important predictor since it has the highest weight in this calculation
d. Lastly, we find that shooting and defending are complementary attributes and they jointly affect a player's potential. We assume that based on a player's position on the pitch, attackers or defenders, the potential calculation has different weights for the shooting and defending attributes. For example, when computing the potential of a forward, we focus more on their attacking ability, so the weight on how their defending ability influences their potential would be less than that when we measure the potential of a defender and vice versa. This effect is exactly captured and proofed by the negative measure of the relationship between shooting and defending in the model

D. Limitations of the Analysis

This is a preliminary study and we need to do further analysis as there are the following limitations.

To begin with, as the assumption checking part in the appendix D shows, the model doesn’t hold the residual normality assumption, which suggests that there may be other important factors that are not considered in the model. We actually included some predictors outside of the scope of a player's football skills, such as wage and reputation. On the one hand, this is because we believe that measuring a player’s potential is a complex thing, since a player’s reputation or wage do influence a player’s state of mind, thus further influencing a player’s potential. On the other hand, this also suggests that some other variables that are directly related to player future performance need to be discovered and introduced.

Besides, some predictors in this dataset, such as player’s shooting and dribbling skills, and player’s wage and reputation, have high correlation with each other, which hinders the interpretation of the impact of how different attributes contribute to a player's potential.

E. Conclusions & Recommendations

The factors Age, Wage, International Reputation, and skills such as shooting, passing, dribbling, and defending are able to explain how the potential of a player is rated. Using these factors, any change in the value of these factors can warrant a change in the player’s potential rating next year.

As shown in figure 2, a player’s market value increases exponentially with the player’s overall rating. This can be leveraged as an opportunity for teams to improve their player's market value by helping them improve upon the factors that affect the same.

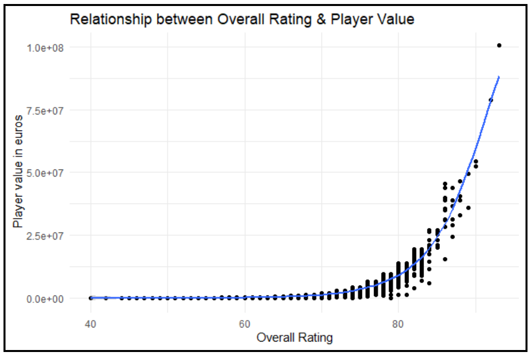


Figure 2

Teams can help improve their player’s values using the following strategies to ensure a higher price for a player they wish to sell in the near future:

- For older players, increasing the wage of the player using a short-term contract can be help drive up player’s market value
- For younger players, helping players i[1]improve their shooting and dribbling skills can help improve a player’s potential ratings and in-turn drive up their market value

F. References

We get the data from kaggle “FIFA 22 complete player dataset”. This data has been scraped by kaggle from the publicly available website ‘https://sofifa.com’.
Link: https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset

[1]

Appendix

A. Exploratory Data Analysis

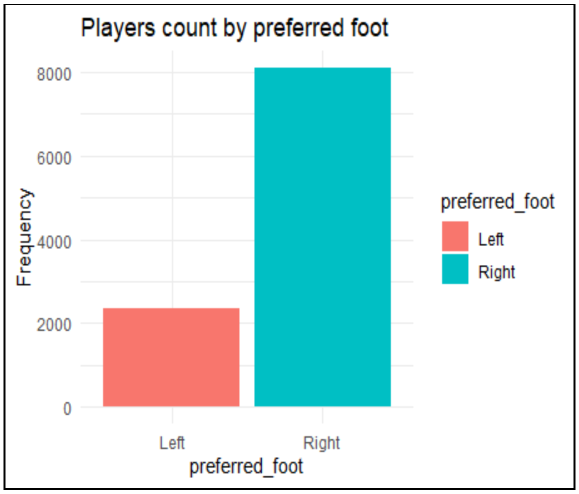
The dependent variable in this study is the potential of the football player. The independent variables fall into three main categories

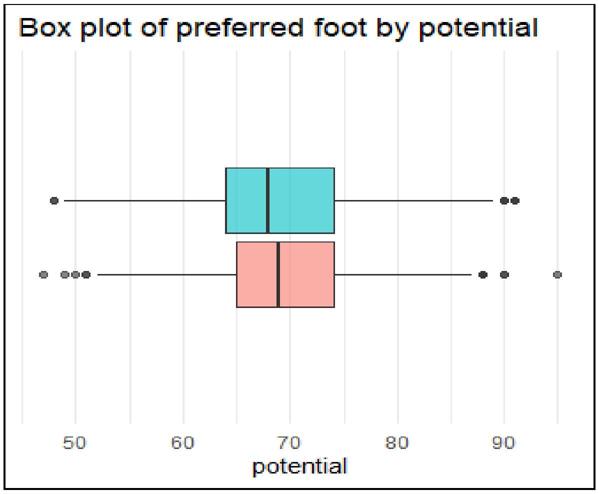
- The physical attributes of a player
- The players’ football skills
- Global recognition of players

There is no missing data in any of the variables. The sample size of data is 9307. We analyze various predictor variables with the response variable the dependent variable to get an idea of the nature of the relationship between them.

a. Preferred foot

Preferred foot is a categorical variable that indicates if the preferred shooting foot of the player is right or left. Preferred foot is a physical attribute of the player that cannot be changed late in the careers. We see that there are more players with the preferred foot as right than left, which is expected. It is observed from the box plot shown below that the players with the left preferred foot have slightly higher median potential than the ones with the right preferred foot.

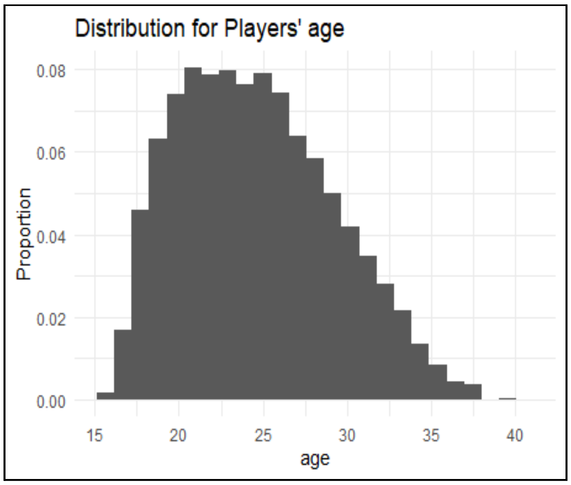


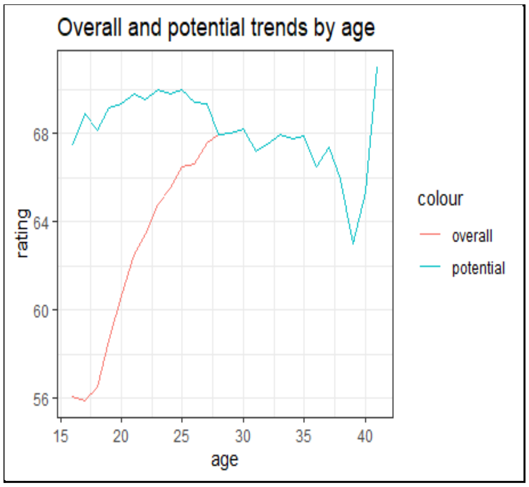


b. Age

Age is an interval level variable indicating the player’s age at the time the data was collected. We can see that a large proportion of players are between age 20 and 27, with a rare case of a player continuing to play even beyond 40 years of age.

Additionally, as we can see, as the age increases a player starts to close into his expected potential and the difference between the potential and the overall vanishes. This suggests that younger players have a higher area for improvement in their potential rating compared to players above the age of 30, as seen from the graph below.

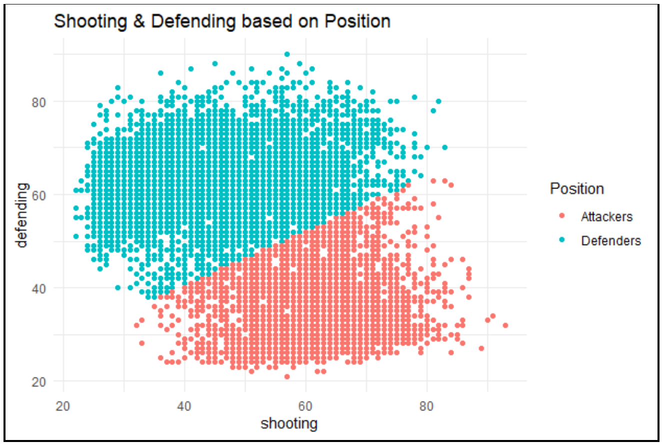




c. Shooting and Defending Skill

We have observed that in the data, there is no categorical distribution between attacking players and defensive players, however these two skills should hypothetically be complementary.

Investigating this by grouping players into two different groups based on their shooting and defending skills, we see that the two attributes separate attacking and defensive players well. Thus, we can conclude that some relationship exists between the shooting and defending ratings and its impact on the potential of a player.

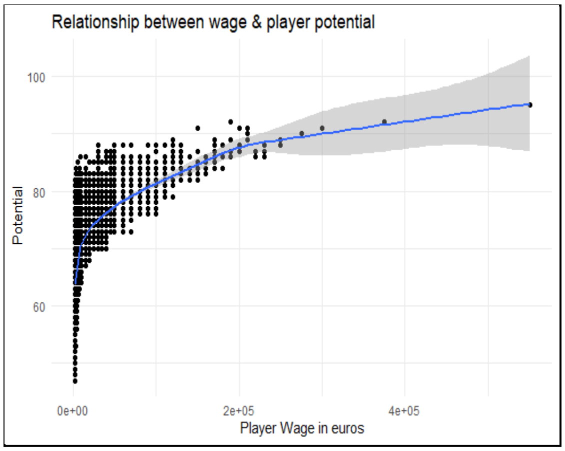


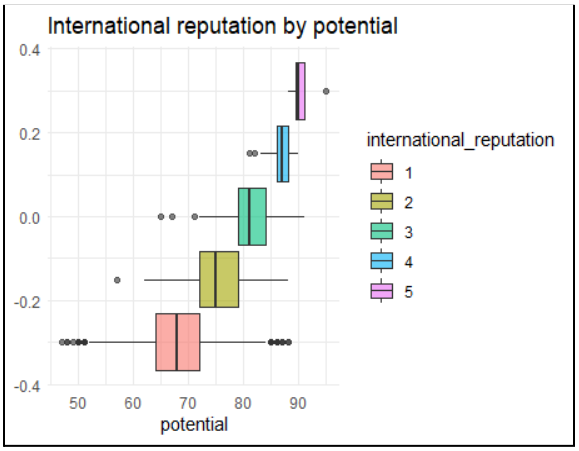
d. Players' Wage & International Reputation

Player's wage, market value, and international reputation should all be related to the potential rating of a player, as it acts as a metric to gauge the promise a player shows and thus their importance in the world of football.

Exploring the relationship between player's wage and potential, we can see that as the potential increases, the player wages increase exponentially.

Similarly, the international reputation of a player varies from 1 to 5 with 5 being the highest popularity. We can see from the box-plot that the potential is higher as the international reputation increases for a player.





B. Model selection

a) Hypothesized model

Potential = β0 + β1 * age+ β2 * wage_eur + β3 * pace + β4 * shooting + β5 * passing + β6 * dribbling + β7 * defending + β8 * physic + β9 * international_reputationF2 + β10 * international_reputationF3 + β11 * international_reputationF4 + β12 * international_reputationF5 + β13 * preferred_footFright + ϵ

The table below shows the results of various iterations of model building with different predictors

Predictors	**k (number of terms in equation)**	**Overall p-value**	**R Squared**	**error(s)**	**Highest p-value**
preferred_foot	2	0.0279			0.0279
international_reputation	2	<2e-16			<2e-16
age	2	4.46E-13	0.56%	6.807	4.46E-13
value_eur	2	< 2.2e-16	28.7%	5.765	< 2.2e-16

wage_eur	2	< 2.2e-16	42.4%	5.181	< 2.2e-16
pace	2	< 2.2e-16	12.2%	6.397	< 2.2e-16
shooting	2	< 2.2e-16	20.1%	6.103	< 2.2e-16
passing	2	< 2.2e-16	31.5%	5.651	< 2.2e-16
dribbling	2	< 2.2e-16	32.8%	5.597	< 2.2e-16
defending	2	< 2.2e-16	2.5%	6.741	< 2.2e-16
physic	2	< 2.2e-16	8.1%	6.545	< 2.2e-16
age , wage_eur, pace, shooting, passing, dribbling , defending , physic , international_reputation , preferred_foot	14	< 2.2e-16	67.5%	3.893	0.564 (pace)
age , wage_eur, shooting, passing, dribbling , defending , physic , international_reputation , preferred_foot	13	< 2.2e-16	67.5%	3.893	0.117 (preferred_foot_right)
age , wage_eur, shooting, passing, dribbling , defending , physic , international_reputation	12	< 2.2e-16	67.5%	3.893	3.71e-05 (passing)
age , wage_eur (with log(wage)), shooting, passing, dribbling , defending , physic , international_reputation	13 (added log(wage))	< 2.2e-16	75.9%	3.357	0.0821 (passing)
age , wage_eur (with log(wage)), shooting, dribbling, defending, physic, international_reputation	12	< 2.2e-16	75.9%	3.357	4.09e-13 (wage)

age , wage_eur (with log(wage)), shooting, dribbling, defending, international_reputation (with interactive relationship between shooting and defending)	12	< 2.2e-16	80.0%	3.043	0.008383 (wage)

Table 2: Model results comparison

b) Model Selection Process

We first introduced all the variables of the hypothesized model into the model. The built model could predict 67.5% of the variance of potential. We found there was about a 57% chance of getting the observed sample results when the predictor pace had no relationship with potential while keeping other variables constant. So, we concluded that ‘pace’ did not have a significant relationship with potential while other variables were present and removed it from the model.

We repeated this process iteratively and removed the variables like pace, preferred foot, passing, and physics that we found had no significant relationship with the dependent variable potential, while other variables were present. The model could still predict 67.5% of the variance of potential, even after removing these variables and without an increase in error as well.

We then checked the model assumption and found that the residuals did not have a fixed standard deviation across all predicted y and across all the wage predictors. This means that some parts of the model are problematic. We then went back to check the scatter plots between wage and potential and discovered that it is better to be a logarithm relationship than a linear relationship. So we introduced the term log(wage) to our model and this increased the R-square explained by model to 75.9%.

At the same time, from a business perspective, the potential for a forward (attacker) and defender should have different levels of importance on skills of shooting and defending. So, we tested if there was an interactive relationship with shooting and defending and further removed the variables that did not explain the variance in the potential that was already being explained by other variables. The resulting model was able to explain about 80% of the variance in potential with an error of 3.043 (least observed error so far).
Comparing the final model with the former ones, we can see that we decrease the number of unnecessary variables to make the model more simple and reduce overfitting. Plus, we significantly increase the r-square and decrease the residual standard error. All the remaining variables have very low p-value, or significant relationships with potential and satisfy logical relationships from a business perspective. Based on this reason, we choose this model to be the final one. Here are the detailed results of the final model.

c) Final Model Results

Looking at the final model equations -

For international_reputation level1:

Potential = 9.077 - 0.7048*age+ - 0.000007812 * wage_eur + 3.041 * log(wage_eur) + 0.6293 * shooting + 0.6656 * defending + 0.1393 * dribbling - 0.009707 * shooting * defending

For international_reputation level2:

Potential = 10.0553 - 0.7048*age+ - 0.000007812*wage_eur + 3.041 * log(wage_eur) + 0.6293 * shooting + 0.6656 * defending + 0.1393 * dribbling - 0.009707 * shooting * defending

For international_reputation level3:

Potential = 11.946 - 0.7048 * age+ - 0.000007812 * wage_eur + 3.041 * log(wage_eur) + 0.6293 * shooting + 0.6656 * defending + 0.1393 * dribbling - 0.009707 * shooting * defending

For international_reputation level4:

Potential = 14.891 - 0.7048 * age+ - 0.000007812 * wage_eur + 3.041 * log(wage_eur) + 0.6293 * shooting + 0.6656 * defending + 0.1393 * dribbling - 0.009707 * shooting * defending

For international_reputation level5:

Potential = 13.9 - 0.7048 * age+ - 0.000007812 * wage_eur + 3.041 * log(wage_eur) + 0.6293 * shooting + 0.6656 * defending + 0.1393 * dribbling - 0.009707 * shooting * defending

The above coefficients are referred from the results below.

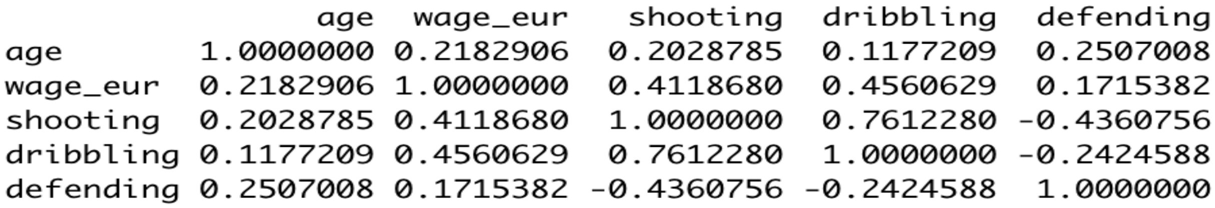
Coefficients	Estimate	P-value
Intercept	9.077	2.00E-16
age	-0.7048	2.00E-16
wage_eur	-7.05E-06	8.38E-03
log_wage	3.041	2.00E-16
shooting	0.6293	2.00E-16
defending	0.6656	2.00E-16
dribbling	0.1393	2.00E-16
international_reputationF2	0.9783	3.48E-15
international_reputationF3	2.869	2.00E-16
international_reputationF4	5.814	2.00E-16
international_reputationF5	4.823	2.27E-04
shooting*defending	-9.71E-03	2.00E-16

Table 3. Model coefficients

We have a high degree of belief that all the predictors have a significant relationship with the dependent variable potential, and these attributes also have real-world logical relationships with potential, so no further deletion is needed in this model selection. We then check the assumptions for the final model.

C. Correlation between selected predictors

The table below shows correlations between the selected variables.



```
                age  wage_eur   shooting  dribbling  defending
age       1.0000000 0.2182906  0.2028785  0.1177209  0.2507008
wage_eur  0.2182906 1.0000000  0.4118680  0.4560629  0.1715382
shooting  0.2028785 0.4118680  1.0000000  0.7612280 -0.4360756
dribbling 0.1177209 0.4560629  0.7612280  1.0000000 -0.2424588
defending 0.2507008 0.1715382 -0.4360756 -0.2424588  1.0000000
```

Table 4. Correlation coefficients

As we can see, age and wage do have a significant correlation with other predictors. However, other predictors, such as shooting and dribbling are highly correlated, which makes it impossible to find out how much impact the predictor shooting or dribbling individually has on the potential of a player while other variables are constant.

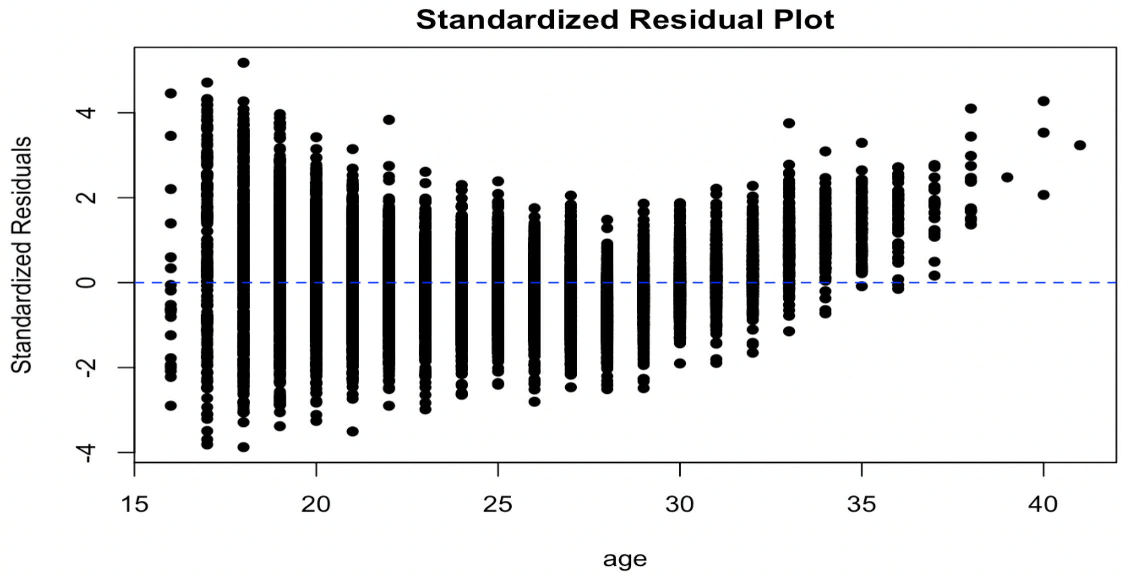
D. Model assumptions

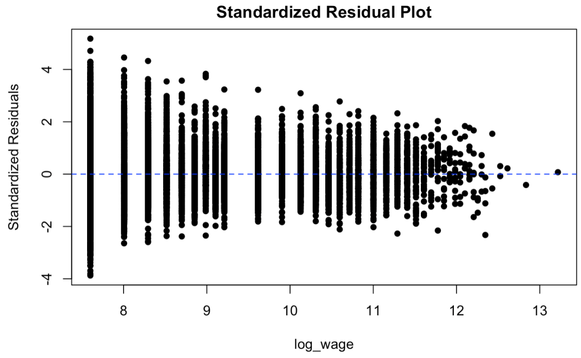
Following are the assumptions for our multi-linear regression model.

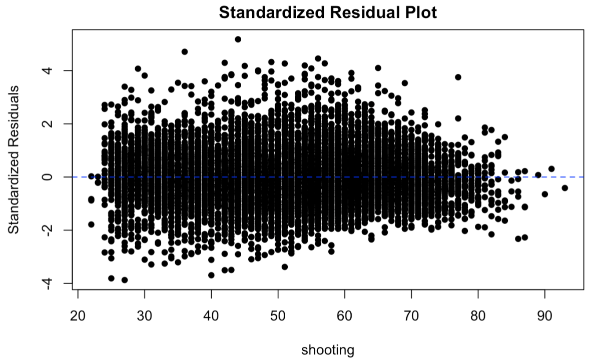
1. Random Sampling
2. Stability of process over time
3. The residuals of the model are normally distributed

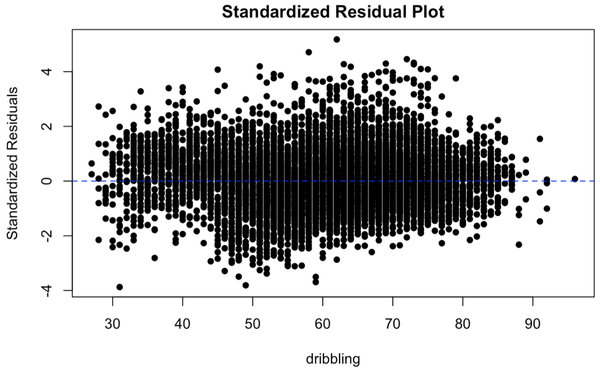
By making the random sampling and stability over time assumption, we assume that the sampling process is meaningful and does not have sampling bias. Specifically, the random sampling hypothesis assumes that all players are equally likely to be selected into the data sample. Since we use a population data that gathered all the players in the Fédération Internationale de Football Association in 2015, we can believe that the sampling process satisfies random sampling. For the stability over time assumption, since we use only one-year data, we can assume that the player's attribute features are fixed over the period when the data were collected.

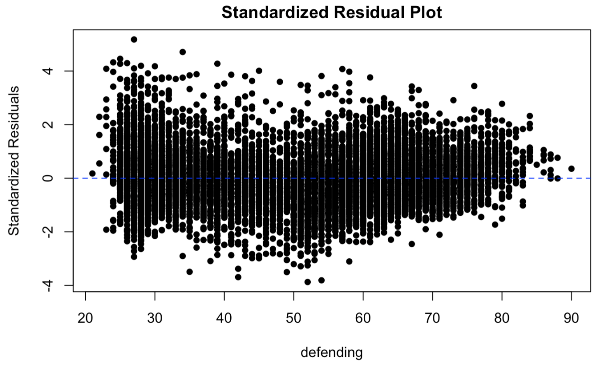
To test the standard error assumption, we first draw the residual scatter plot between residual and each predictor with the following result.





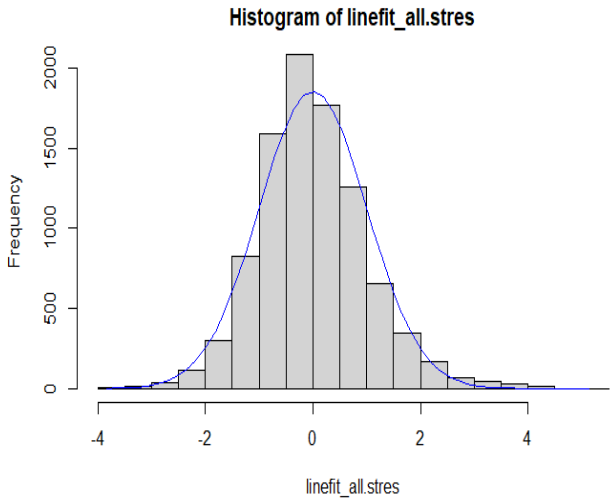


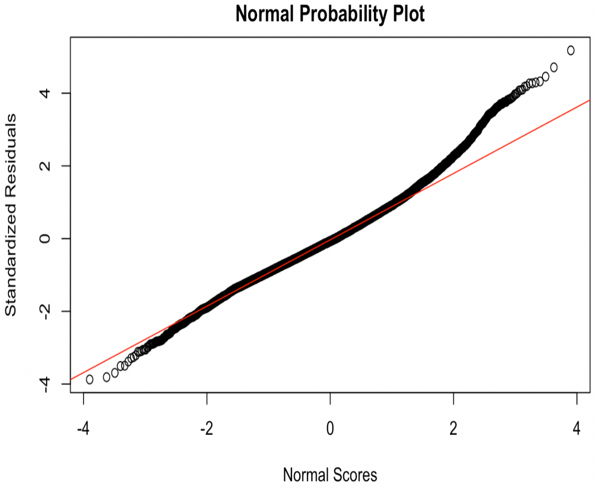




The result shows that while the plots between residual and shooting, defending and dribbling respectively show that there is no problem in the model. The scatter plot between age and residual shows that there may be some factors that are not included in the model. The scatter plot between log_wage and residual shows that there is a problem of heteroscedasticity.

As for normality checking, we plotted the histogram and qq-plot for residuals. Here are the results.





The shapiro test could not be conducted as the sample size of the dataset is higher than 5000. When sampling the residual to satisfy this number restriction, we got a significantly low p-value. Therefore, based on the above results, we can reject that the residual is normally distributed. Since our final model does not fulfill the assumptions, further analyses need to be done on this.