Analysis of Players Participating in FIFA World Cup 2022 Qualification Games

Abstract

The FIFA World Cup qualification games consists of teams from countries around the world competing for the chance to hold the championship title among all football teams. Various factors may influence outcome of matches including player physiology, skill, and team dynamics. As a result, it is important to consider parameters that may increase a country's chances of winning when selecting their teams. This paper in particular seeks to analyze the linear relationship between player age and number of minutes played. Based on our linear regression test, our expected β_1 suggests a 9.363 minutes increase in playing for every 1 unit increase in age with 95% confidence. Based on our analysis our p-value for $H_a>0$ is .996. However, for an $H_a<0$ we calculate a p-value of .000367. Thus for this secondary analysis, we have enough evidence to reject $\beta_1=0$ that there is no linear relationship between age and total playing time.

Introduction

The purpose of this project is to purge data collected from the FIFA World Cup qualification games in order to determine if there is an association between age and the total number of minutes played of a given player during the 2022 World Qualification games. Ethical concerns of this analysis may arise from the potential to worsen issues that discriminate against players based on age. The study will be looking at teams specifically from the continents South America and Europe due to historical performance of teams from these regions. Did you know that over 3.5 billion people across the globe follow soccer/associated football passionately? And, the World Cup 2022 is expected to garner a total of greater than 5 billion views over the course of the entire FIFA World Cup tournament. Now, exactly what is FIFA? FIFA is a federation International de Football Association in simple words, FIFA is the organization that oversees all official matches and international competitions between various national teams for our knowledge. The FIFA World Cup, which is held every four years, is the most prestigious competition in all of the sports. According to estimates, 1.12 billion people watched the 2018 FIFA World Cup final which is 10 times more than the Super Bowl audience (99 million). FIFA World Cup Qualifiers are the pre-world cup tournaments played in different continents across the world. The 32 qualifying teams for the FIFA world cup are determined from these qualifier games where our data came from. It is so fascinating to see that this event manages to unite people from different backgrounds, ethnicity, and continents under one roof. The objective of the 90-minute game, which is divided into two 45-minute halves with 11 players on each team, is to score goals within the allotted time. A player's experience is paramount in leading their nation on the global stage while handling the pressure given what's at stake. Experience comes with the amount of games played during the entirety of a player's career, amount of years invested in the sport. Our research focuses on whether total playing time has a negative linear relationship with a player's age,, which is linked to their performance during the entire game, is of interest to us.

Data

The data was collected through FBref, an open source website that tracks data of football teams, and was extracted and compiled through an API. The dataset contains information of each player for the teams of a respective country during the qualification games including team name/affiliation, player name, player number, age, position of the player, playing time of the player, goals scored, number of assists, penalty kicks, yellow and red card violations, playing time, and other metrics. The dataset includes 567 rows and 24 columns; after wrangling the data, the final dataset includes 481 rows and 3 columns. In our study we utilized stratified sampling; we filtered for players only from teams representing countries within South America and Europe; the U.S. is included as well (*Teams: Argentina, Belgium, Brazil, Croatia, Denmark, England, France, Germany, Netherlands, Poland, Portugal, Spain, Switzerland, United States, Wales)*. We then filtered for the columns of player age, the explanatory variable, and total number of minutes played, the response variable, and conducted hypothesis testing with linear regression in R.

Hypothesis:

H_0: $\beta_1 = 0$. There is no correlation between player age and number of total number of minutes played.

H_a: $\beta_1 < 0$.There is a negative correlation between age and number of total number of minutes played.

Population Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Where

Y = Average playing time in minutes

β_0 = Average playing time in minutes where age is 0 years

B_1 = Average change in playing time for each 1 unit increase in age

ε = Errors are i.i.d ~ $N(0, \sigma^2)$

Potential confounder variables may include if a player received an injury during a match (which would cause a lower amount of playing time overall to be given), the differences in playing style (i.e. offensive vs defensive focused), and other unforeseen circumstances such as red/yellow card violations.

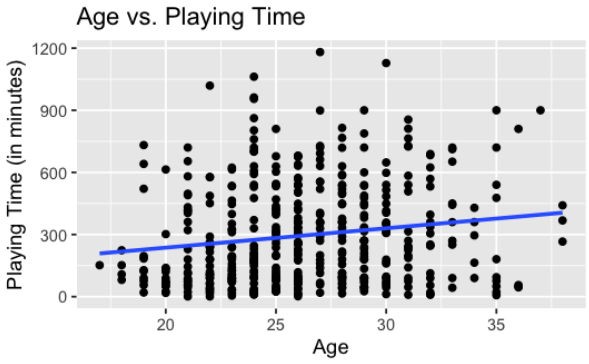
Testing Conditions

Conditions for linearity, normality, and constant variance of the residuals are met. It is safe to proceed to a linear regression test with caution with regards to the abundance of zeroes in the data. [1, 2, 3]

Summary Statistics

Mean Age	**Standard Deviation Age**
26.376	*4.098*
Mean Total Playing Time	**Standard Deviation Total Playing Time**
296.100	*241.432*

Regression Test



Modeling

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	38.953	70.861	0.550	0.582777
Age	9.749	2.655	3.672	0.000267

Conclusion

We hypothesized in the experiment that age and playing minutes have a negative linear relationship. As a result of our hypothesis, we cannot reject the null hypothesis. In our regression plot, it is clearly evident that age and playing minutes have a positive linear relationship. Based on our original alternative hypothesis where $H_a<0$, P-value= 0.999867, and thus we cannot reject the null hypothesis; however, in a secondary analysis where we hypothesize that $H_a>0$ (a positive linear relationship between age and total playing time), then our p-value of 0.000134 indicates that we can reject the null hypothesis that states there is no linear relationship between age and total playing time. We can additionally conclude with a 95% confidence that the expected values for β_1, the expected playing time for each 1 unit increase in age is between 4.533 and 14.966; our calculated value for the average increase in total playing time for each 1 unit increase in age was 9.749.

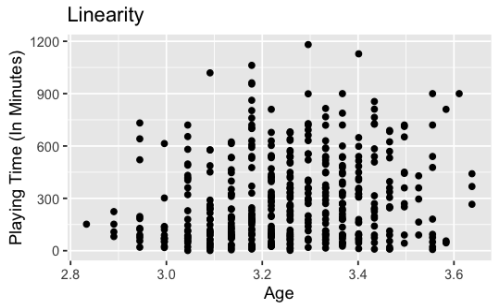
There are limitations to our data since it includes data of players who are not playing or who are given fewer minutes than regular starters. This results in many values of zero within our data.

References

1. https://www.kaggle.com/datasets/mateusdesousamartins/world-cup-2022-national-teams-data-set

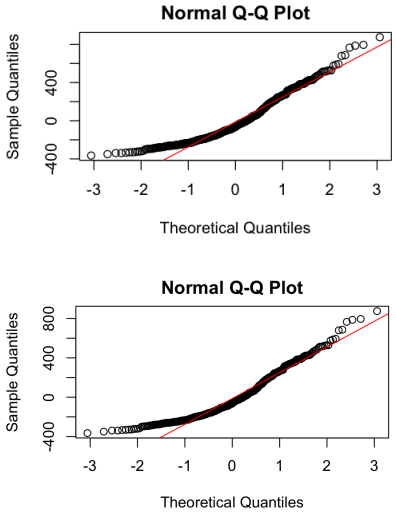
Appendix

Figure 1. Linearity



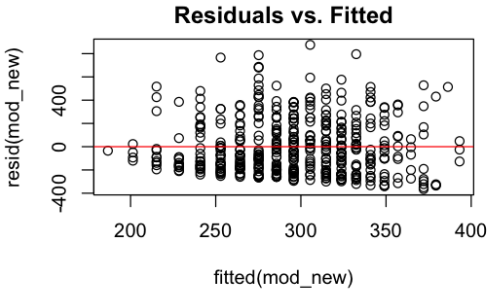
Linearity of data is observed upon visual inspection.

Figure 2. Normality



A log transformation was applied in an attempt to correct for normality of the original data (top figure) but upon application of the log transformation (bottom figure) no significant change was observed; therefore, the data is left in its original form.

Figure 3.
Constant Variance of Residuals



Constant variance of the residuals is observed.