Predicting Soccer Player Market Value

Dav King, Luke Thomas, Thomas Barker, Harry Liu

Introduction and Data

Introduction and Research Question

Every year, tens of thousands of soccer players are transferred on the international transfer market. Soccer clubs pay other clubs, sometimes hundreds of millions of pounds, to sign players to their team in hopes of improving club performance. These astronomical transfer prices are often determined by clubs with the help of Transfermarkt, a popular online soccer database that uses data scouts to collect data from the match sheets of soccer games. Our goal with this research is to take a deeper look into Transfermarkt's valuations of soccer players around the globe. Specifically, we are asking: how can we use a player's individual characteristics (nationality, age, position, etc.) and performance (goals, assists, minutes, yellow cards, etc.) to predict their market value? We predict that factors like being younger, playing as a forward, having more goals and assists, and playing more minutes will all contribute positively/increase a player's market value.

Data Description

There are several different datasets present in the overall data (players, player_valuations, appearances). As mentioned in the introduction, these data were collected by Transfermarkt's data scouts who look at the match sheets of soccer games and collect statistics like these. In the players dataset, each observation is a player and contains information such as their name, player id, country of citizenship, position, and date of birth. In the player valuations dataset, each observation is a market value of a player, and also includes the player id and date of the valuation. In the appearances dataset, each observation is a player's appearance in a soccer game, and it also includes the player id, number of goals scored in that appearance, number of assists in that appearance, number of minutes played in that appearance, and the number of yellow and/or red cards obtained in that appearance.

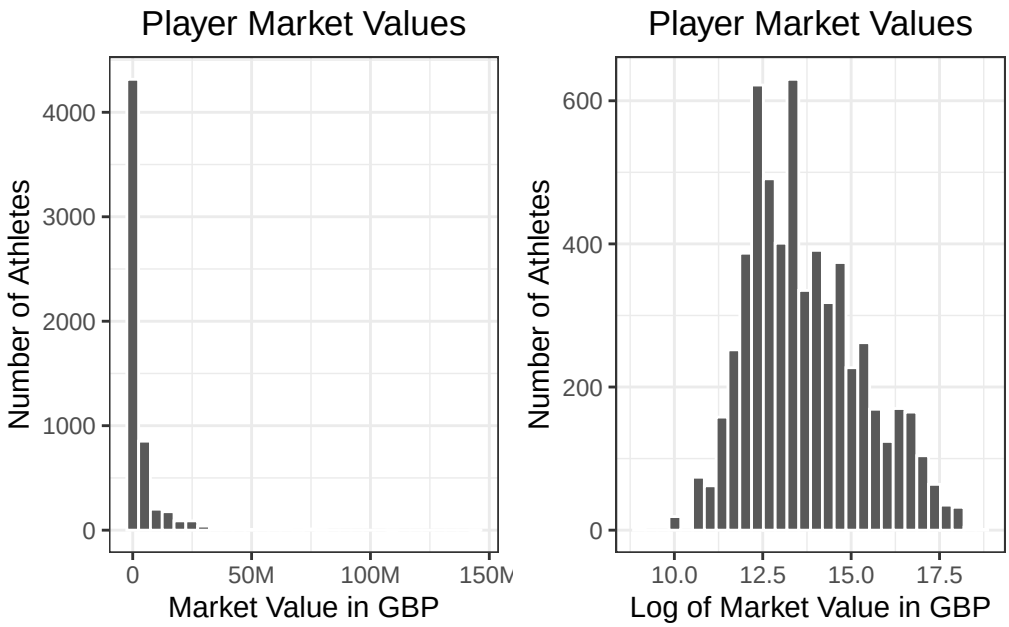
These data sets have been combined such that each observation is a player and contains their final market value. We conducted data cleaning, which includes dropping a number of

repetitive or irrelevant variables generated during the join (these includes game_id, datetime, appearance_id, etc), and filtering the dataset to only include those players who played at least one minute in the 2018-2019 season. Variables, including total_goals, total_assists, total_minutes, and total_yellow are created by grouping by player_id and summing the the goals, assists, minutes_played, and yellow_cards for each player respectively. These summed-up number are the variables we are using for building the model and they represent the total number of goals, assists contributed by each player, their total time (in minutes) played, and the total number of yellow cards received, all exclusively during the 2018-2019 season. After these numbers being summed for each player, we dropped the original variables including goals, assists, minutes_played, and yellow_cards. Additionally, we created the age_in_days variable to represent how old the player is - this is produced by subtracting their birth date from the date that their final valuation was recorded in the 2018-2019 season. Together from these transformations, after aggregating the data together, it is ensured that there is exactly one observation for each player. Additionally electing to include for analysis only for the 2018-2019 season allows us to derive more meaningful predictions from our dataset and eliminate the effects of time-related co-founding variables such as inflation and age. After the data cleaning, we are left with N = 5872 players to include in our analysis.

Data Citation: Cariboo, David. *Football Data from Transfermarkt.* (data file). Kaggle, 2022. Web. 03 Nov 2022. https://www.kaggle.com/datasets/davidcariboo/player-scores

Exploratory Data Analysis

Distribution of Response Variable: End-of-Season Market Value



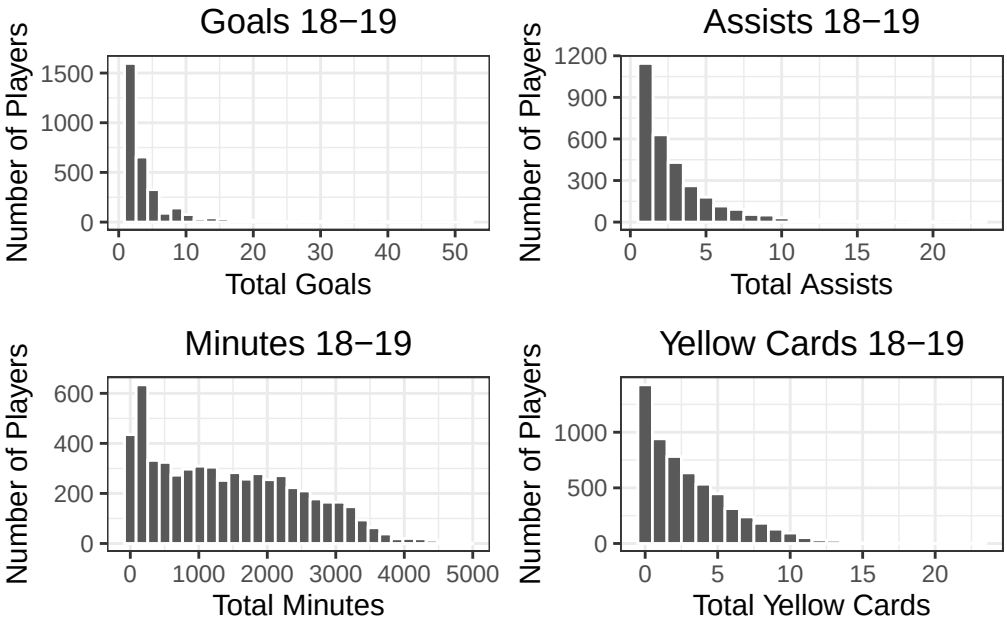
Our initial look at the univariate distribution of player market values shows a unimodal and extremely right-skewed distribution, with a range from 9,000 to 135,000,000 pounds and a median at 293000 pounds (Q1 = 180000, Q3 = 720000). This is clearly a non-normal distribution, and it may well be better considered a logarithmic distribution instead.

If we log-transform the response variable, however, it roughly follows a normal distribution. It is unimodal, with a center around 12.5 and minimal skew. It ranges from 9.1 to 18.72 and has a median at 12.59 (Q1 = 12.1, Q3 = 13.49). While we will not necessarily run our regression using the log of our response variable, this is still important to note.

Distribution of Key Predictors Variables: On-Field Performance

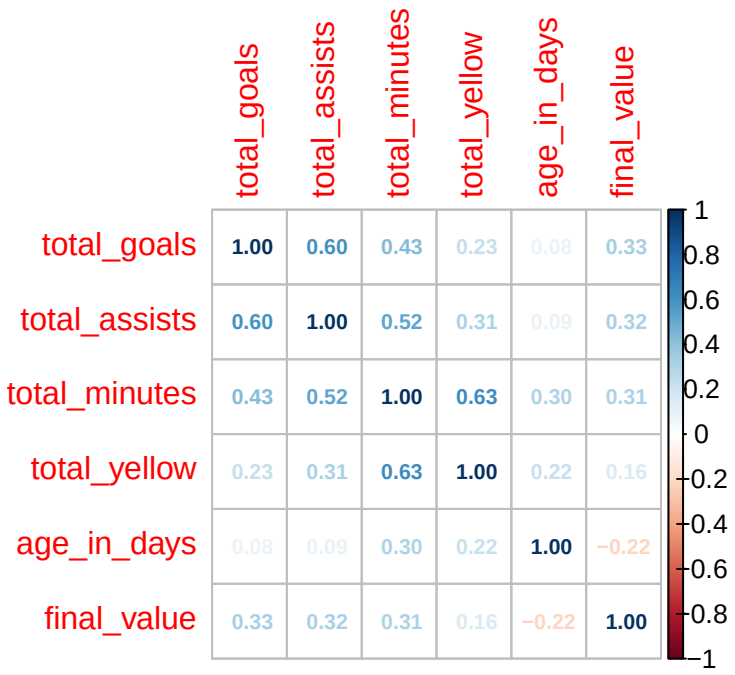
Other key predictors that will be considered are the stats for each player's on-field performance, frequently referred to in this report as their "statistics". Given that the distributions of goals, assists, and minutes played are heavily right-skewed, and any observations greater than 0 were considered outliers, we elected to present the graphs only when the values were present (i.e., greater than 0); however, we have considered the data both with and without these values. The first plot shows us the total goals scored by players with more than zero, which shows a unimodal but incredibly right-skewed distribution, centered around 2 with a range from 1 to 51. The median is 2 (Q1 = 1, Q3 = 5), with a mean at 3.83 (sd = 4.36). The second plot

shows us the total assists by players with more than zero, which shows a unimodal but still right-skewed distribution, again centered around 2 with a range from 1 to 23. The median is 2 (Q1 = 1, Q3 = 4), with a mean at 2.94 (sd = 2.7). The third plot shows us the total minutes played by players with more than zero, which shows a unimodal and right-skewed (but less so) distribution, with a mode around 250 and a range from 1 to 4913. The median is 1168.5 (Q1 = 348, Q3 = 2141), with a mean at 1338.91 (sd = 1066.15). The final plot shows us the total number of yellow cards that players received, which shows a unimodal and right-skewed distribution, with a range from 0 to 23, a median of 2 (Q1 = 1, Q3 = 5), and a mean of 3.03.



Numeric Predictor and Response Relationships

The plot below shows the correlations between our key numeric predictor variables and our response variable of final market value. The most noticeable correlation coefficients are between a player's total minutes played, and their total goals, assists, and yellow cards, producing values of 0.55, 0.64, and 0.78 respectively. These each indicate a moderately strong correlation, which logically makes sense, since a player who plays more minutes would expect to accumulate more stats. Another noticeable value is the moderately strong correlation of 0.67 between a player's total goals and total assists, since these stats are both accumulated more by attacking players and less by defensive players. From this information, it may be interesting to explore the potential interaction effect between a player's position and their total goals and assists, since attackers are expected to have more goals and assists than defenders. Lastly, there does not appear to be strong correlations between age or final market value and any of the other numeric predictors.



Methodology

Our goal in this analysis was to select the best possible model for predicting a player's market value on the basis of their individual characteristics. In order to capture variability in market value, we elected to employ a multiple linear regression framework. Over the process of selecting a model, we considered seven different multiple regression models, all of which predicted a player's final value. While the first model predicted market value directly, due to extreme issues in the assumptions required to run multiple linear regression, the other six models predicted the log of market value (see Appendix for more information). The models considered a number of different predictors, including country of citizenship and birth, position, age, height, total goals, assists, yellow cards, and minutes. Additionally, the final model explored whether predictions of market value could be improved when considering potential differences in how individual player statistics predicted market value based on the position that they played.

All seven models were developed within the same framework, predicting market value (or its logarithm) with a multiple linear regression and briefly testing model assumptions and conditions. The first model (`unTransModel`) predicted market value directly from our predictors of interest (except for country of birth, which was only used in the `noStatsModel`). The second model (`logModel`) did the same, except it predicted the logarithm of market value instead. The third model (`logModel2`) was the same as `logModel`, except that it also log-transformed the predictors of minutes, goals, assists, and yellow cards, all of which were shown to have serious right-skew. The fourth model (`sigModel`) took only the predictors from `logModel`

that were statistically significant in order to predict the log of market value. The fifth model (`statsModel`) used only player on-field statistics to predict the log of market value, while the sixth (`noStatsModel`) used all other predictors. Finally, the seventh model (`interactModel`) used all predictors from `logModel`, as well as considering the interaction between position and all individual player statistics.

For all models, we mean-centered numeric predictors, dummy-coded categorical predictors and collapsed them down to meaningful levels, and removed all zero-variance predictors. In order to select the optimally performing model, ten-fold cross validation was performed on all models using the same folds, which were a random subset of 80% of the data set aside for training the models. This allowed us to be certain that we calculated more robust R^2, RMSE, AIC, and BIC values for all models, as well as identified the model that was most adept at making predictions on new data. The results from that model can be seen in the table below.

Model	RMSE	Adj. R Sq.	AIC	BIC
unTransModel	7356567.169	0.266	136645.95	136728.28
logModel	1.213	0.411	12785.91	12874.53
logModel2	1.320	0.303	13452.17	13521.94
sigModel	1.214	0.411	12786.51	12849.99
statsModel	1.406	0.208	13955.19	13986.61
noStatsModel	1.463	0.144	14016.07	14110.71
interactModel	1.213	0.414	12780.82	12938.58

Our cross validated metrics allow us to select between models. The first thing that is evident from the table is that `unTransModel` has very different statistics from the other models, as it is the only model to directly predict market value instead of its logarithmic transformation. However, as noted above and explained in appendix 1, models that did not log-transform market value saw extreme violations of the constant variance assumption required for linear regression. Thus, we will not consider that model for selection. With it eliminated, the other six models all predicted the logarithm of market value, and so their performance metrics can be compared directly without any variable transformations.

Because our goal of this process is to develop the optimal model for *predicting* a player's market value, we will focus primarily on the metrics of RMSE and BIC. RMSE, which is a measure of the average error within a model (and thus a value that should be minimized) and is the most important value to consider for prediction, identified two models that were best for selection: `logModel` and `interactModel`, which both had the same RMSE scores down to at least the third decimal place. Between these, `logModel` had the better BIC score of 12874.53. Among the six models, it also had the second-best AIC score and was tied for the best adjusted R^2 score, meaning that it also performed well in terms of explanation. Thus, this is the model that we will use to predict a player's market value going forward.

Before settling on the model, we needed to evaluate its conditions for linear regression. The first condition is independence. This dataset is full of players who are different from one another. There is no reason to suspect that any of our predictors except for possibly minutes have anything to do with one another, as they do not have maximum values and there is no reason to suspect that one player can tell us anything about another player. Additionally, market values are also independent from one another, as there is no salary cap in this soccer league and teams have proven very willing to pay extensively for players. Thus, the independence condition is met. The second condition, linearity, is also met. As seen in the plots shown in appendix 2.1, none of the predictors appear to have relationships with the residuals that are non-linear. The third condition is normality, which can be seen in appendix 2.2. The residuals seem to be distributed approximately normally; further, as we have many more than 30 observations, the CLT allows us to say that the normality condition is met regardless. The final condition is constant variance, which can be seen in the residuals vs. fitted plot in appendix 1. Unfortunately, there do seem to be some issues with constant variance in this plot. In order to check whether our conclusions were statistically sound in light of this, we simulated this model on 2,000 bootstrap samples of our data. This output can be found in appendix 3. Because all of our model's values were consistent with the findings of the bootstrapped model, we are confident in our model's validity despite the violation of constant variance. Additionally, VIF values for all of our variables were well below 10, which gives us confidence that we do not have any issues with multicollinearity in our model; these can be seen in Appendix 2.3.

Results

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	13.72449	0.01826	751.58388	0.00000	13.68869	13.76029
age_in_days	-0.00049	0.00001	-36.45455	0.00000	-0.00051	-0.00046
total_goals	0.06020	0.00689	8.73898	0.00000	0.04670	0.07371
total_assists	0.06710	0.01012	6.62931	0.00000	0.04726	0.08695
total_minutes	0.00062	0.00003	21.94142	0.00000	0.00056	0.00067
total_yellow	-0.00787	0.00833	-0.94541	0.34450	-0.02419	0.00845
height_in_cm	0.00988	0.00134	7.34644	0.00000	0.00724	0.01251
country_of_citizenship_France	0.35308	0.10759	3.28187	0.00104	0.14216	0.56400
country_of_citizenship_Spain	0.45405	0.10250	4.42998	0.00001	0.25311	0.65500
country_of_citizenship_other	-0.21465	0.07754	-2.76832	0.00566	-0.36666	-0.06264
position_Defender	-0.02893	0.05151	-0.56164	0.57439	-0.12991	0.07205
position_Goalkeeper	0.00330	0.08308	0.03971	0.96832	-0.15958	0.16618
position_Midfield	0.11777	0.05404	2.17930	0.02936	0.01182	0.22372

$\hat{final\_value} = e^{13.7} \times e^{0.00049 \times age\_in\_days} \times e^{0.0602 \times total\_goals} \times e^{0.0671 \times total\_assists} \times e^{0.00062 \times total\_minutes} \times e^{-0.00787 \times total\_yellow} \times e^{0.00988 \times height\_in\_cm} \times e^{0.35308 \times citizenship\_France} \times e^{0.45405 \times citizenship\_Spain} \times e^{-0.21465 \times citizenship\_Other} \times e^{-0.02893 \times position\_Defender} \times e^{0.00323 \times position\_Goalkeeper} \times e^{0.11777 \times position\_Midfield}$

Each of these slopes/effects will help us assess the hypothesis we made about our research questions, which was that being younger, playing as a forward, having more goals and assists, and playing more minutes will all contribute positively/increase a player's expected market value. We found, interestingly, that playing a position besides an attacker (midfielder, goalkeeper) actually contributed more positively to a player's expected market value than being an attacker, holding all else constant. However, upon further thought this makes sense, as a midfielder is likely more valuable than an attacker with the exact same number of goals and assists. The rest of the attributes in our hypothesis (more goals, being younger, more assists, more minutes), did indeed contribute positively to a player's expected market value. Specifically, a player's market value was impacted by these key statistics/attributes in the following ways. Significant relationships were found for age, total goals, total assists, and minutes played, in which for each additional one year increase in age, total goal scored, total assist made, and minute played the median market value of a player is expected to multiply by a factor of 0.83705, 1.0621, 1.0694, and 1.000615, respectively, on average, all else held constant. Among the effects of positions, only the effect of playing as a midfielder was significant, by which the market value of a midfielder is expected to be 1.1250 times that of an attacker, on average, all else constant. The effect of the positions of defender and goalkeeper after controlling for our other variables, however, were not significant, as their p-values were 0.574 and 0.968, respectively.

In order to evaluate our model, we then tested it on the remaining 20% of our data set aside for testing. The R^2 value when the model is applied to testing data is 0.391, which is only 0.02 less than the R^2 of the model when applied to the testing data, which is 0.411. Additionally, the rmse of the model when applied to the testing data is 1.264, which is not much higher than the rmse of the model when applied to the training data, which is 1.213. These small differences indicate that our model performs nearly just as well as it did on the data it was trained on. In light of this, we believe that our model is well-fit, but not over-fit, for the prediction of soccer player market value.

Discussion and Conclusion

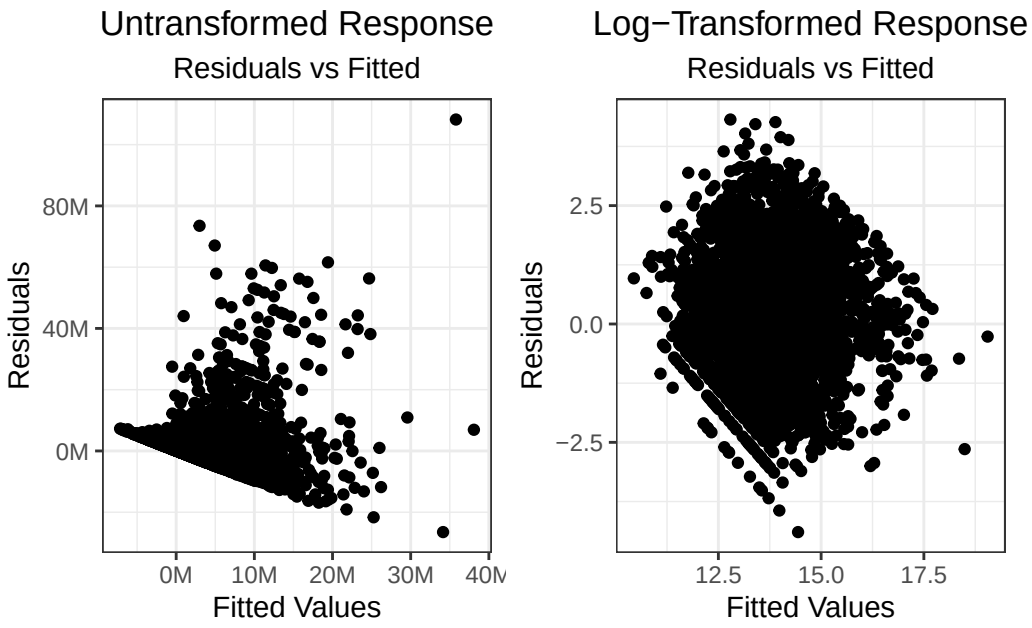
In the international soccer transfer market, soccer clubs often pay millions of pounds to sign promising players to their team. However, tens of thousands of players are transferred each year, making it impossible for coaches or scouts to evaluate them all. It is desirable for a club to identify players that they could sign for less than their true market value or sell for more than their true market value, so a successful predictive model for a player's market value would hold great power and practicality among the professional soccer community. In this project, we identified potentially significant and logical predictors, such as a player's age, position, nationality, height, total goals, total assists, total yellow cards, and total minutes played in the 2018-19 season, and used them to predict a player's market value at the end of the season.

In our final selected model, we found that total goals, assists, and minutes played all contributed to a positive increase in player market value, all backed by a p-value of less than 0.05 to justify significance. This supports our hypothesis and makes logical sense in the context of the data, since players with more statistical contributions would expect to demand a higher value. Additionally, we found that increased age significantly contributes to lower market value, since younger players are generally valued more highly. Interestingly, we found that greater height significantly contributes to greater player market value, suggesting that tall players are more desirable. We also found that players from popular nations (Brazil, France, and Spain) valued significantly higher, on average, than players from other nations, potentially suggesting that those nations produce more good soccer players. Lastly, we found that midfielders are valued significantly higher than forwards, on average, while the other positions, as well as number of yellow cards, were not significant predictors in our model.

Using this information, soccer clubs can have a better idea of what contributes positively and negatively to a player's market value in the international transfer market. While the TransferMarkt data for this analysis is quite reliable, there are some limitations that should be addressed. All statistics for on-field performance (goals, assists, minutes, etc.) are very right-skewed, with most observations having 0, and many attempts were made to make these variables suitable for a regression model, such as log-transforming. Having 0 on-field statistics does not necessarily indicate that someone should have a market value of 0, however it can be difficult to predict their market value from other categorical predictors alone. Acknowledging that variables are skewed, we conducted testing of seven different models and found that a multiple linear regression with log-transformed response was the most appropriate. Another major drawback is that this model failed to meet some of the required statistical assumptions, which limits the validity of our findings, although it was consistent with the results of running the model on bootstrapped samples. One idea for future work include conducting analysis and prediction across multiple seasons, rather than just one season, to get a better understanding of the player's whole career, and avoid potentially inaccurate predictions due to injury or other effects during the 2018-19 season. Additionally, more predictors could be included in the model, such as which league a player plays in or which club they play for.

Appendix

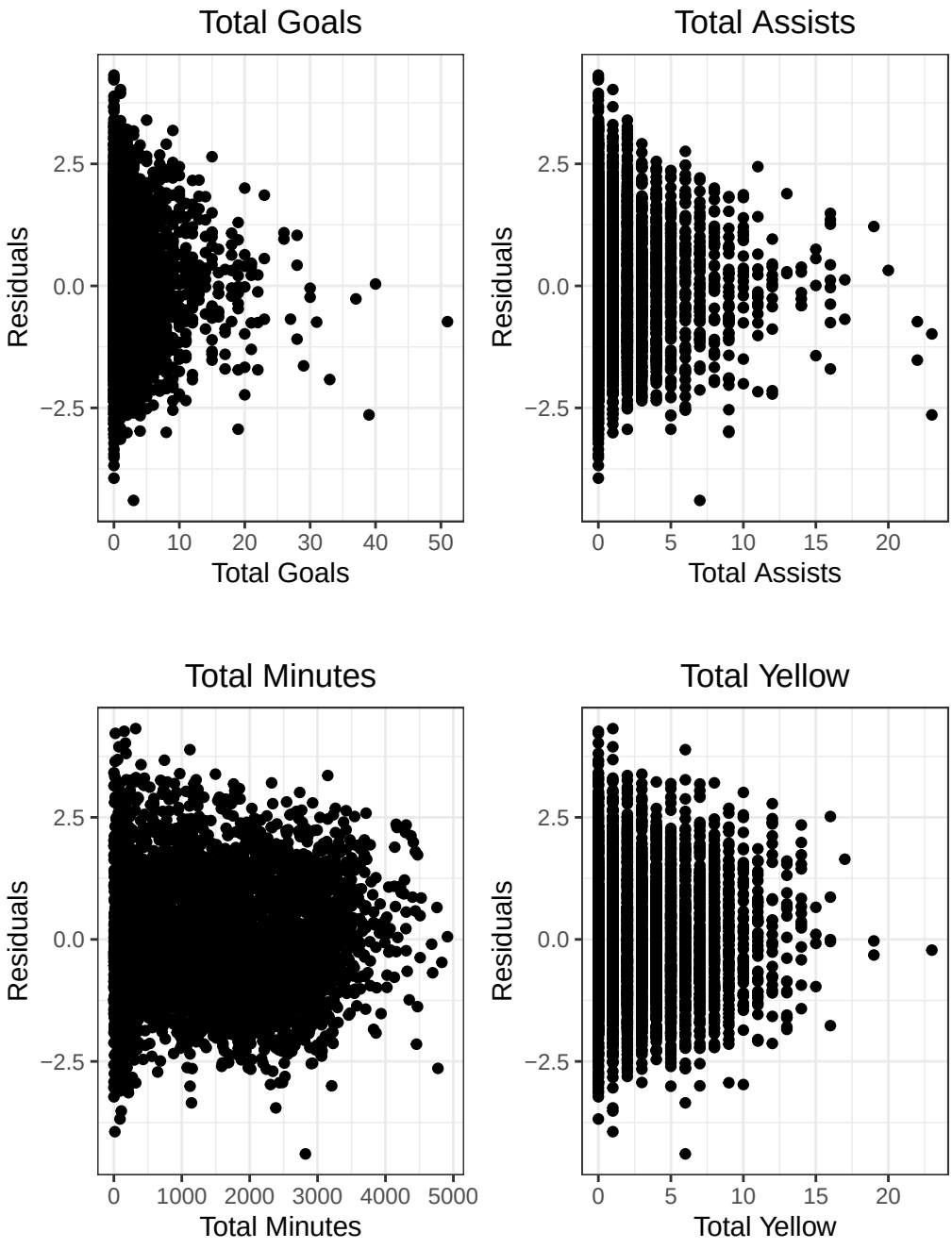
Appendix 1: Reasoning for Log-Transformed Response Variable

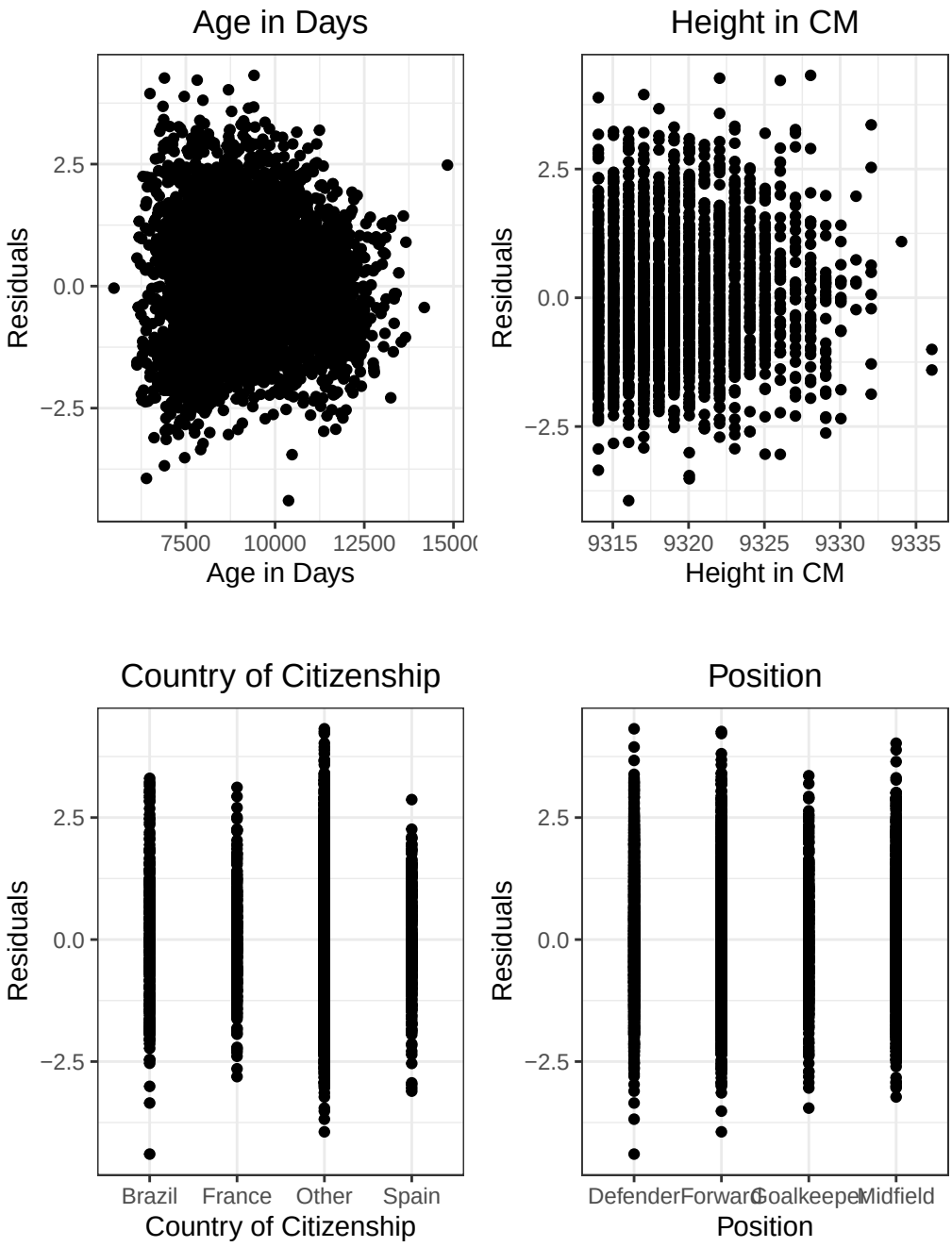


These two plots show the residuals vs fitted for `unTransModel` (left) and `logModel` (right). The first plot very clearly shows a trend in the residuals that violates constant variance for an untransformed response of market value. The second plot shows the same plot for `logModel`, which is the exact same model except that market value was log-transformed. Although there are still clear trends in this plot that make this model less than ideal, as mentioned above, it still comes so much closer to meeting the assumption of constant variance that we can remove from consideration all models which do not predict the log of market value.

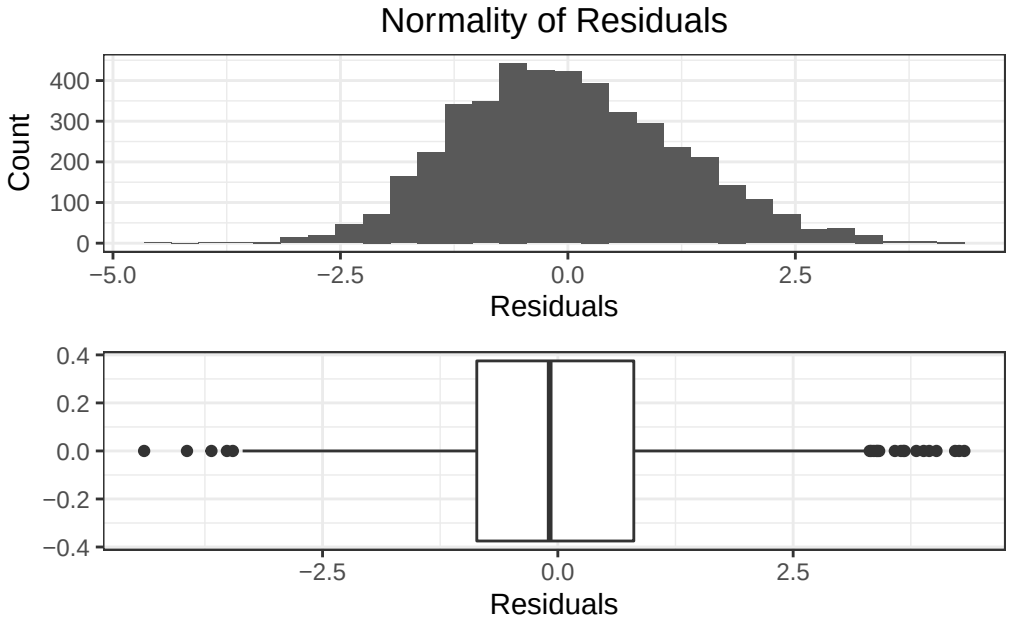
Appendix 2: Assumptions

Appendix 2.1: Linearity




Age in Days
Residuals
2.5
0.0
−2.5
7500
10000
12500
Age in Days
Height in CM
9315
9320
9325
9330
9335
Height in CM
Country of Citizenship
Brazil
France
Other
Spain
Country of Citizenship
Position
Defender
Forward
Goalkeeper
Midfield
Position

Appendix 2.2: Normality



Appendix 2.3: VIF Values

age_in_days	total_goals
1.131424	1.997554
total_assists	total_minutes
1.899186	2.650630
total_yellow	height_in_cm
1.939660	1.055685
country_of_citizenship_France	country_of_citizenship_Spain
1.851580	1.981347
country_of_citizenship_other	position_Defender
2.708793	1.790871
position_Goalkeeper	position_Midfield
1.545923	1.499627

Appendix 3: Bootstrapped Model

term	.lower	.estimate	.upper	.alpha	.method
(Intercept)	13.67717	13.72427	13.77305	0.05	percentile
age_in_days	-0.00051	-0.00049	-0.00046	0.05	percentile
country_of_citizenship_France	0.14556	0.35452	0.57640	0.05	percentile
country_of_citizenship_Germany	-0.13075	0.05716	0.24997	0.05	percentile
country_of_citizenship_Greece	-0.91097	-0.83242	-0.75386	0.05	percentile
country_of_citizenship_Italy	0.06704	0.29774	0.51330	0.05	percentile
country_of_citizenship_Netherlands	-0.63290	-0.39392	-0.16305	0.05	percentile
country_of_citizenship_other	-0.38348	-0.21339	-0.04014	0.05	percentile
country_of_citizenship_Portugal	-0.59304	-0.30818	-0.08688	0.05	percentile
country_of_citizenship_Russia	-0.72043	-0.50188	-0.29772	0.05	percentile
country_of_citizenship_Spain	0.24480	0.45585	0.66808	0.05	percentile
country_of_citizenship_Ukraine	-0.73947	-0.61993	-0.49619	0.05	percentile
height_in_cm	0.00815	0.00989	0.01203	0.05	percentile
position_Defender	-0.12273	-0.02804	0.07427	0.05	percentile
position_Goalkeeper	-0.14911	0.00371	0.17137	0.05	percentile
position_Midfield	0.01403	0.11826	0.22903	0.05	percentile
total_assists	0.04931	0.06762	0.08629	0.05	percentile
total_goals	0.04789	0.06016	0.07252	0.05	percentile
total_minutes	0.00056	0.00062	0.00067	0.05	percentile
total_yellow	-0.02394	-0.00826	0.00792	0.05	percentile