Are English football players valued the most?

Elias Irens, Danielius Michailovas and Rimgaudas Jurgaitis

ISM University of Management and Economics

ECO 105: Econometrics

PhD. Cand. Aleksandr Christenko

Oct. 31, 2024

Hypothesis

Football is one of the most popular and highest revenue-generating sports in the world and especially in the UK. This is evidenced by the most recent international tournament Euro 2024 viewership - combined peak audience of 24.2 million across the BBC and ITV watched England's matches during the tournament (Sim, 2024). While the total revenue of the top 20 football clubs, by generated revenue, during the 2022/23 was €10.5bn (Deloitte, 2024). Considering the popularity of this sport and the huge amount of money flowing through it, it is fair to assume that the value of football players is also quite high. There are many factors that determine the valuation of a footballer. One of them is nationality. Players of some nationalities are valued more than others. Well-known example of this phenomenon that is being discussed on the internet in the present day – English footballers are more expensive than others. Fans and other enthusiasts of the sport believe that English players are often overvalued in comparison to others. They claim the main reason for this is exposure. The majority of English representatives in the sport play in the English Premier league - the most watched football league worldwide and the Championship which is the second league in England. This means English media gets more exposure than other national media which leads to increased attention towards English players (Bell et al., 2023). A question arises if English footballers are actually valued higher on the market. The aim of this analysis essay is to explore whether football professionals with English nationality have higher valuation on the market than players with other nationalities using OLS regression analysis.

Data commentary

Dataset for the OLS regression analysis was taken from Kaggle. Data comes from Transfermarkt - football portal with transfers, market values, rumours and statistics. Initial dataset contained information on 32400 male professional football players registered on Transfermarkt. Removal of players with missing data and filtering led to a total of 14741 player observations born in 10 different countries. Variables used in the analysis are presented in Table 1 below.

Table 1.

Variable description.

Variable	**Explanation**
Market value (market_value_in_eur)	Dependent numeric variable used to compare footballers based on their estimated value on the market (Transfermarkt) in euros. Valuation takes into account many factors that determine how good a player is currently.
Nationality (country_of_birth)	Independent categorical variable which represents the player's country of origin. Countries selected for this analysis based on the amount of high-valued top players: England, Argentina, Belgium, Brazil, France, Germany, Italy, Netherlands, Portugal and Spain.
Position	Independent categorical variable that describes in which area of the football pitch a person plays and what roles and responsibilities he has on the field.
Age (date of birth)	Independent numeric variable which shows how old an athlete is in years.
League (current_club_domestic_competition_id)	Independent categorical variable that determines the domestic competition (league) in which plays the club the football player represents.

Describing data

The frequency distribution analysis done on independent categorical variable "Country of Birth" demonstrates that the categories for countries are not equal frequency and proportion wise, as shown in the table below:

Table 2.

Frequency of "country_of_birth" variable.

Name	Frequency	Percent, %
France	2228	15.11%
Spain	1925	13.06%
Italy	1829	12.41%
England	1715	11.63%
Germany	1596	10.83%
Brazil	1587	10.77%
Netherlands	1451	9.84%
Portugal	1021	6.93%
Belgium	772	5.24%
Argentina	617	4.19%

The fact that categories' segment sizes range from 4.19% to 15.11%, implies that some categories may be over-represented compared to others, which may lead to bias in coefficient and p-values.

Frequency distribution table on independent categorical value "Positions" results were as follows:

Table 3.

Frequency of "position" variable.

Name	Frequency	Percent, %
Attacker	3863	26.21%
Midfield	4283	29.06%
Defender	4899	33.23%
Goalkeeper	1696	11.51%

Since football is entirely team based sport, the quantities of separate positions should be proportional. With 11 total players possible in a team, we tried to account for that. Defender and midfielder percentages (33.23% and 29.06% respectively) have aligned with our expectations. Generally, a football team has 3 to 5 players for both of those positions, so, proportionally, it is reasonable for them to be between 27.(27)% - 45.(45)%. The same goes

for attackers, with a usual proportion of 9.(09)% - 27.(27)% in a team, 26.21% is reasonable. Goalkeeper, percentage of 11.51% exceeds the expected 9.(09)%, which may be caused for variable reasons, such as lesser physical requirements and retired players being considered active.

Table 4.

Frequency of "current_club_domestic-competition_id" variable.

Name	Frequency	Percent, %
BE1 - Belgium	917	6.22%
DK1 - Denmark	124	0.84%
ES1 - Spain	1737	11.78%
FR1 - France	1434	9.73%
GB1 - England	1370	9.29%
GR1 - Greece	705	4.78%
IT1 - Italy	2265	15.37%
L1 - Germany	1242	8.43%
NL1 - Netherlands	1500	10.18%
PO1 - Poland	1755	11.91%
RU1 - Russia	154	1.04%
SC1 - Scotland	626	4.25%
TR1 - Turkey	708	4.80%
UKR1 - Ukraine	204	1.38%

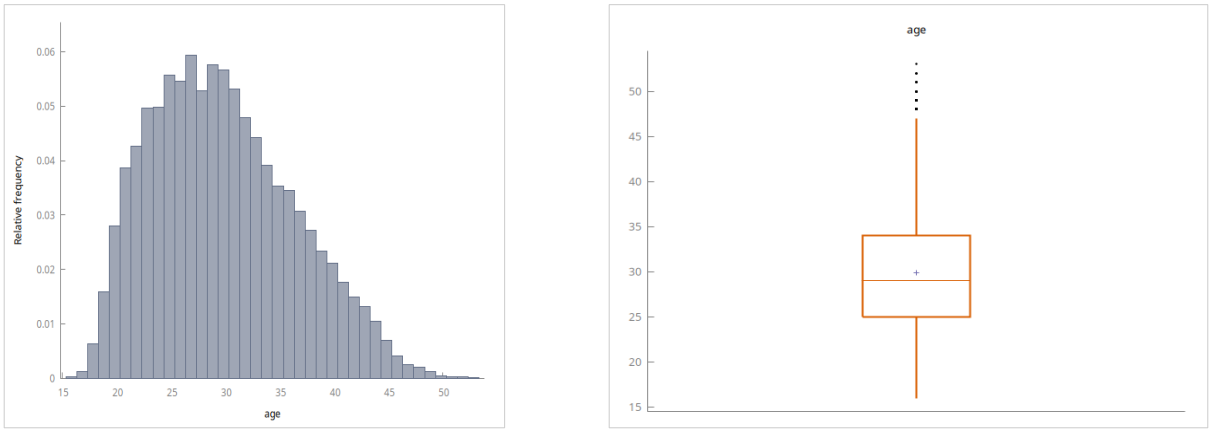
Via the frequency distribution analysis we can see that leagues which, within our datasets size limits, have most players competing in them are Italian (15.37%), Polish (11.91%) and Spanish (11.78%), whilst the leagues with the least players are Ukrainian (1.38%), Russian (1.04%) and Danish (0.84%) leagues. Similarly to our "Country of Birth" variable, "current_club_domestic-competition_id" suffers from uneven distribution across

categories, which may lead to noisy, unreliable estimates and biases in coefficients of categories, particularly, which have less than 1.5% of total players.

Age variable ranges from 16 to 53 with a interquartile range 9 of and mean of 29. As expected, 95% of the players fall in the 20-42 age range, since those years are peak performance for men. The difference between mean (29.9) and median (29) of 0.9 suggests that there is a positive skew in the data. This can also be seen in the frequency distribution graph.

Graphs 1 and 2.

Frequency distribution and a box plot of "age" variable.



In the dataset, the lowest market value of players was 10000 euros. Most valuable players are Vinicius Junior, Kylian Mbappé, Jude Bellingham, Erling Halland with estimated market value of 180 mln. euros. Since the nature of valuations, the graph is an exponent (see Graph 3), this can also be seen in the substantial difference of 1.7 mln. euros between mean (2000000) and median (300000). Only 5% of the players in the dataset have a market value that is higher than 10 mln. euros.

Age is quite an important variable when comparing countries as a whole. If the player base of certain countries in the dataset are on average older, this could partially explain the differences in market value of the players, based on their country of birth. As seen on the table of average age by country, there are quite a few fluctuations. The difference between the average youngest (Netherlands) and oldest (Argentina) country is 3.7 years. This could already explain minor differences of market value by country, although the majority of the countries have an average of around 30 years.

Table 5.

Player's average age by "country_of_birth".

Name	Average age
France	30.02
Spain	30.11
Italy	29.11
England	29.28
Germany	29.39
Brazil	31.45
Netherlands	29.01
Portugal	30.43
Belgium	29.6
Argentina	32.71

Discrepancies among a country's share of positions should in fact influence the average valuation of that country's players. We can observe from the table below that some nations are defending (defenders and goalkeepers) heavy, while others are leaning towards the attacking side of the game. To classify a country in this way, we should compare the value to dataset averages. Attack leaning countries are: Brazil, Argentina. More defend leaning are Portugal, Italy and Spain. Italy has an unexpectedly way bigger goalkeepers section than average (+5.3%), while Argentina is uniquely dominating both attackers and midfielders proportion share (67.4% of the players vs. 55.4% dataset average). Discussion and analysis of each position's impact on market value will be explored in subsequent sections.

Table 6.

Player's position distribution by "country_of_birth".

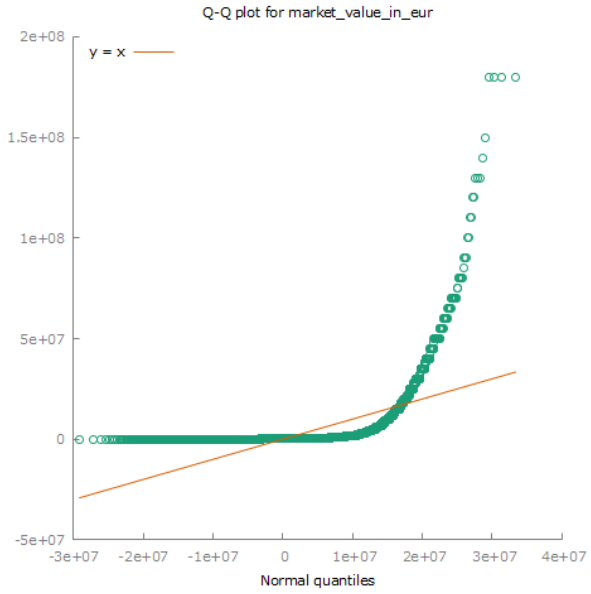
Name	Attackers, %	Midfielders, %	Defenders, %	Goalkeepers, %
France	26.12	29.31	34.87	9.69
Spain	26.96	27.43	33.97	11.64
Italy	21.6	28.43	33.13	16.84
England	27.23	28.63	34.46	9.68
Germany	24.37	31.95	30.08	13.6
Brazil	31.51	26.21	34.34	7.94
Netherlands	25.71	28.26	32.87	13.16
Portugal	23.8	29.48	35.95	10.77
Belgium	23.96	32.12	30.83	13.08
Argentina	34.04	33.39	26.58	6

Dependent variable analysis

Dependant variable "market_value_in_eur" was tested for normality via Doornik-Hansen, Shapiro-Wilk, Lilliefors, Jarque-Bera tests and Q-Q plot (see Graph 3). The results for all of the normality tests for "market_value_in_eur" show a p-value of 0 or approaching it, which implies that the tests have detected a significant difference between the data distribution and a normal distribution. The results of Doornik-Hansen test and Jarque-Bera test highlight that dependent variable is likely not normally distributed in terms of skewness and/or kurtosis.

Graph 3.

Q-Q plot for "market_value_in_eur".

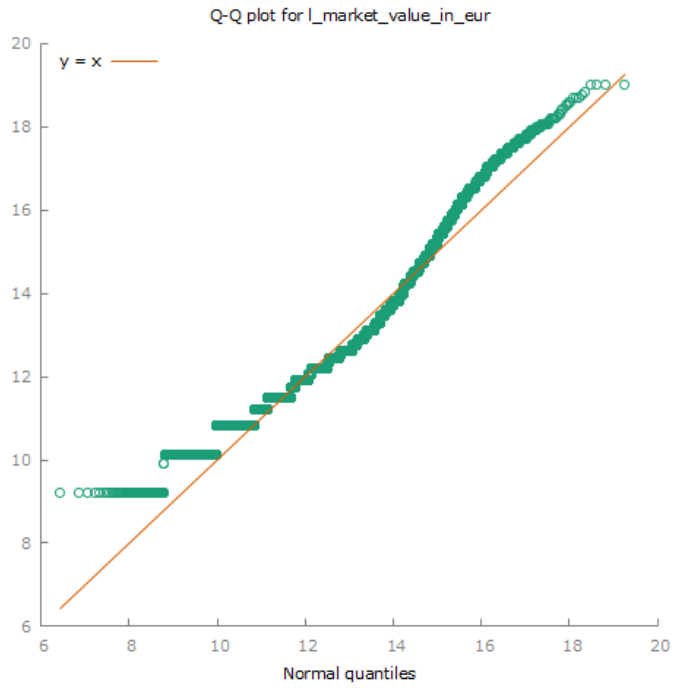


The results of the normality tests are further supported by the Q-Q plot, which depicts that quantiles of "market_value_in_eur" do not align with quantiles of normal distribution. The majority of dependent variable observations are relatively near 0, however the outliers make the curve positively skewed.

Overall, the test and plot results indicate that transformatory actions to address skewness were needed. To fix non-normality of our dependent variable, we used logarithmic transformation (see Graph 4). This method was chosen based on its ability to fix positive skewness and its general interpretability.

Graph 4.

Q-Q plot for "l_market_value_in_eur".



Model

Heteroscedasticity

Breusch-Pagan test indicates that error terms do not have a constant variance. White's test also indicates the violation of the OLS assumption. This issue will be fixed via accounting for the heteroscedasticity with using robust standard errors, although current model can no longer be BLUE (best linear unbiased estimator).

Multicollinearity

Results of collinearity analysis (see Table 5) suggest that only 3 variables might have some issues. Italy's Serie A league has the highest VIF value of 6.021, which suggests that it could cause some issues, but that is to be assumed, since national leagues contain mostly players from the same country. It was decided to keep the variables that are a problematic, since their VIF value suggests that the issue is quite small and should only make a minimal impact on the accuracy of coefficients.

Table 7.

VIF values for model variables

Variable	VIF Value
Age	1.044
Leagues	
German Bundesliga	3.806
Italy's Serie A	6.061
Turkey's Süper Lig	2.103
Spain's La Liga	4.72
French Ligue 1	3.636
Scottish Premiership	1.403
Belgian Pro League	2.807
Netherlands Eredivisie	4.219
Ukrainian Premier League	1.382
Russian Premier League	1.275
Polish Ekstraklasa	4.242
Super League Greece	2.151
Danish Superliga	1.175
Positions	
Goalkeeper	1.302
Defender	1.526
Midfield	1.504
Nationalities	
Germany	4.315
Brazil	3.629
Portugal	3.172
Argentina	2.077
France	4.531
Belgium	2.63
Netherlands	4.167
Italy	5.731
Spain	5.016

Misspecification

Ramsey's RESET test shows more concerning results than previous tests. The p value of 9.94347e-52 indicates that the current model weakly explains the market value of a player. Since only the age variable could be transformed, the most likely issue is omitted-variable bias. Variables such as proneness to injury, if the player is currently injured or current contract length could better help in explaining the market value of a player, although they were not included in the dataset that have been chosen.

Normality of residuals

By the nature of the market value variable, the data was largely positively skewed. Logarithmic transformation did vanish the majority of the skewness, but it is worth checking again with a different approach. The upper part of the adjusted market value still contributes to a minor positive skew. This can also be seen in the test for normality of residuals, which suggests that the residuals are not normally distributed and, thus, violating one of the Gauss-Markov OLS assumptions.

Age transformation testing

As found in the *Describing data* section, age variable is not normally distributed and, thus, worth comparing the model accuracy when using transformed age variable. To test, logarithm will be used, since it would still allow for fairly understandable interpretations and quick comparison. As a key indicator of whether the model improves, Ramsey's RESET test and basic change of p and coefficient values will be used.

Original model has a Ramsey's RESET test p-value of 9.94347e-52, while the same model only with transformed age variable reaches 9.99702e-36 p-value, which is closer to having adequate specification. In fact, the improvement is relatively quite big, but the final result still suggests that the model weakly describes the relationship and, thus, in absolute terms shows no significant improvement. Change of p-values and coefficients had a small change in terms of relative and absolute measures (+/-1%). Overall, transforming the age variable would bring minor improvements, but become harder to interpret the coefficient, so it was decided to keep the original model.

Model interpretations

Quite large dataset led to almost all variables being statistically significant. Only the midfield and Germany dummy variables did not reach the threshold of significance (p-value of 0.05). Leagues have a large influence on the market value of a player, the biggest being Ukrainian league, which has a -88% (adj.) impact on player's value compared to a player from Premier League. The model clearly showcases that older players are valued less, to be exact, around a 6% drop in value per year. The coefficient of positions suggests that attackers and midfielders are valued way higher than defenders and goalkeepers and being a goalkeeper has the biggest negative impact on market value. All country dummy variables fall into 3 categories. First being German players, that does not have any differences on market value compared to being an English player. Then, there are Italian players, whose impact on market value is more negative compared to English players. Being a Brazilian, Portuguese, Argentinian, French, Belgian, Dutch or Spanish player has a positive impact on market value compared to English players. The impact fluctuates from 40 to 95% (without adjustment) of all selected nationalities, except for German, which was not significant. Only about 18.5% (Adjusted R-squared result) observation variance can be explained with selected variables, this shows that there are opportunities to explore the same relationship with additional variables.

Table 8.
Model results

Variable	Coefficient	p-value
const	15.6348	0
Germany	0.0682984	0.475
Brazil	0.952601	5.49E-28
Portugal	0.710111	1.98E-13
Argentina	0.618766	9.80E-22
France	0.590967	3.16E-12
Belgium	0.646499	2.13E-11
Netherlands	0.513408	3.87E-08
Italy	-0.484806	1.93E-05
Spain	0.412078	0.0156
Goalkeeper	-0.65355	1.73E-55
Defender	-0.405416	0.0008
Midfield	0.0680537	0.6655
German_Bundesliga	-0.698987	1.64E-11
Italian_Serie_A	-1.02023	3.24E-20
Turkeys_Super_Lig	-1.59271	3.86E-28
Spanish_La_Liga	-1.81191	4.36E-36
French_Ligue_1	-1.19538	4.83E-36
Scottish_Premiership	-1.51242	1.16E-125
Belgian_Pro_League	-1.69499	8.27E-82
Netherlands_Eredivisie	-1.77519	6.26E-76
Ukrainian_Premier_League	-2.12462	2.56E-64
Russian_Premier_League	-0.68413	3.64E-17
Polish_Ekstraklasa	-1.29151	3.50E-48
Super_League_Greece	-1.69228	3.32E-79
Danish_Superliga	-1.48368	3.56E-32
age	-0.0579901	4.48E-252

Note. Statistically significant coefficients bigger than 0.15 should be adjusted using Euler's number. Interpretation is measured in percentage terms, since dependent variable was transformed using logarithm.

Conclusions

Discussions in the media suggested that Englishmen in football are more expensive than others. This mini-research essay aimed to determine the possible higher valuation of English association football players on the market in comparison to those of other nationalities.

The OLS regression analysis results indicate that England, on average, has lower value players than other countries selected for this analysis. Only exceptions to this are Germans that the OLS deemed not significant and Italians which were valued lower in line with the hypothesis. Italy has a relatively high number of goalkeepers, hence as our model would suggest, a country that has proportionally more goalkeepers, would have on average lower player market value. Thus, it is highly expected that Italian players would have average lower market value. This analysis presents a conclusion opposite to the hypothesis – English football professionals are worth less on the market than representants from other nations.

The dataset that was used was far from perfect. Some players have their market value taken from different years, some might be already retired. Maybe there can be a few errors in the value of age or nationality, although not likely. Trying to make the dataset more accurate, could lead to a better model overall, despite that, improvements would be minor and should not lead to vastly different results.

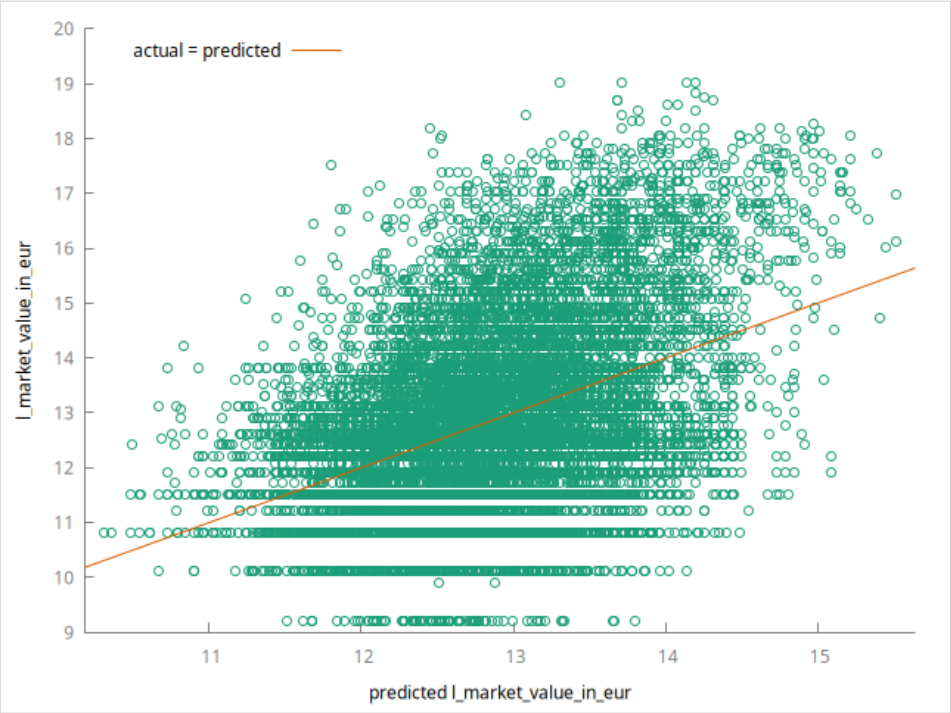
Major improvements in predicting market value could probably only appear from additional variables that would describe current physical form or consistency and, therefore, willingness to make a longer term contract. The signs of omitted variables are observable, because Ramsey’s reset test is showing weak predicting power of the model and most coefficients of countries are unexpected. The inclusion of aforementioned variables could drastically change values or even signs of the current variables. Overall, predicting a market value based on a player's external data is really hard to do accurately, since it mostly depends

on character traits and dedication. For this reason, the current model's predicting power (see Graph 5) of market value based on country of birth is acceptable to be low. Although, some variables such as position or league might also include some traits on an individual player.

Improvements on the predicting power of a model could also come from using a different model than Ordinary Least Squares or using different market value normalization techniques.

Graph 5.

Predicted vs. actual values.



References

Sim, J. (2024, July 18). Euro 2024 in numbers: Attendance records, viewership peaks and a boost in alcohol-free beer sales. SportsPro. https://www.sportspromedia.com/insights/analysis/euro-2024-stats-tv-ratings-viewership-attendance-social-media-sales/

Deloitte. (2024, January 25). Deloitte Football Money League 2024. www.deloitte.com. https://www.deloitte.com/uk/en/services/financial-advisory/analysis/deloitte-football-money-league.html

Bell, A. R., Brooks, C., & Brooks, R. (2023). Are English football players overvalued? Applied Economics, 1–17. https://doi.org/10.1080/00036846.2023.2192032

Football Data from Transfermarkt. (n.d.). www.kaggle.com. https://www.kaggle.com/datasets/davidcariboo/player-scores

Transfermarkt. (2024). Football Transfers, Rumours, Market Values, News and Statistics. Transfermarkt.com. https://www.transfermarkt.com/

Some concepts clarified by ChatGPT; OpenAI. https://chatgpt.com/