678 Project Report

The position will affect the value of soccer player, based on multilevel model analysis

Jiajun Chen

Abstract

Soccer is a popular team sport worldwide, especially in Europe and South America. Current years, the market share of soccer has become higher and higher. To fight for the championship, many clubs spend lots of money to sign players.The highest transfer value is more than 220 million EUR.Every year the transfer market will have some big deal, more than 100 million is very common now. There are so many index to evaluate the value of player, in this paper, we will use multilevel model to analysis how certain variables affect the transfer value. Meantime, we will discuss what position in a team will be sold at a higher value. And the EU five major leagues have any tendency. The result show us that clubs tend to spend their money on forward players, especially who are good at dribbling. This paper has four part: Introduction, Methodology, Result, Discussion.

Introduction

Europe is the most developed region in soccer, people who live in Europe prefer to call it football. 22 people participate as players in one match, each team will have 11 people. They have their own position on the field. Generally, they could be divided into 4 different positions, forward, midfielder,back and goalkeeper. The responsibility of the forward is attack, the midfielder should connect the whole team, the back needs to defend against attack by the opponent. The goalkeeper is special, they can use their hand to keep guard the goal. In the transfer market, the highest trending record is always created by forwards, like Cristiano Ronaldo, Bale, and Neymar. They have one similarity, they are all good at goals. They can

directly bring the lead in a match. Can we consider that clubs tend to spend a high price to sign a player who has more goals than others? However, in other positions, like midfielder and back, they are more difficult to score a goal, so their value will be lower than forward. If we do not think about these top players, do those players who in the mid rank their value will be directly affected by goals or some other points?

As a team game, a player who did well in their position should give them higher value than other players who have more goals, but not well in their own position. I consider that the value is not only affected by goals, different positions should have their own standard to estimate. In this paper, we will explore whether positions influence the value.

Methodology

Data processing

The dataset used for this paper is from Kaggle public dataset which combines two part data. One of them contains all players in five major leagues in Europe and their transfer value estimated by Transfermarket.de which is a professional organization. However, the real transfer value is affected by multiple reasons, like the fund reserves of each club, player contract length, the willingness of both clubs to trade, and personal ideas. Even though there are some differences between real value and estimated value, generally the transfer value recorded by Transfermarket.de is still a good reference. The other part is from fbref.com which is a website that records player data, such as goals, assists, passes completed and so on. There are more than 350 variables. In this paper we will focus on some direct data, like goals, assists to analysis and try to find whether the position will affect value.

EDA
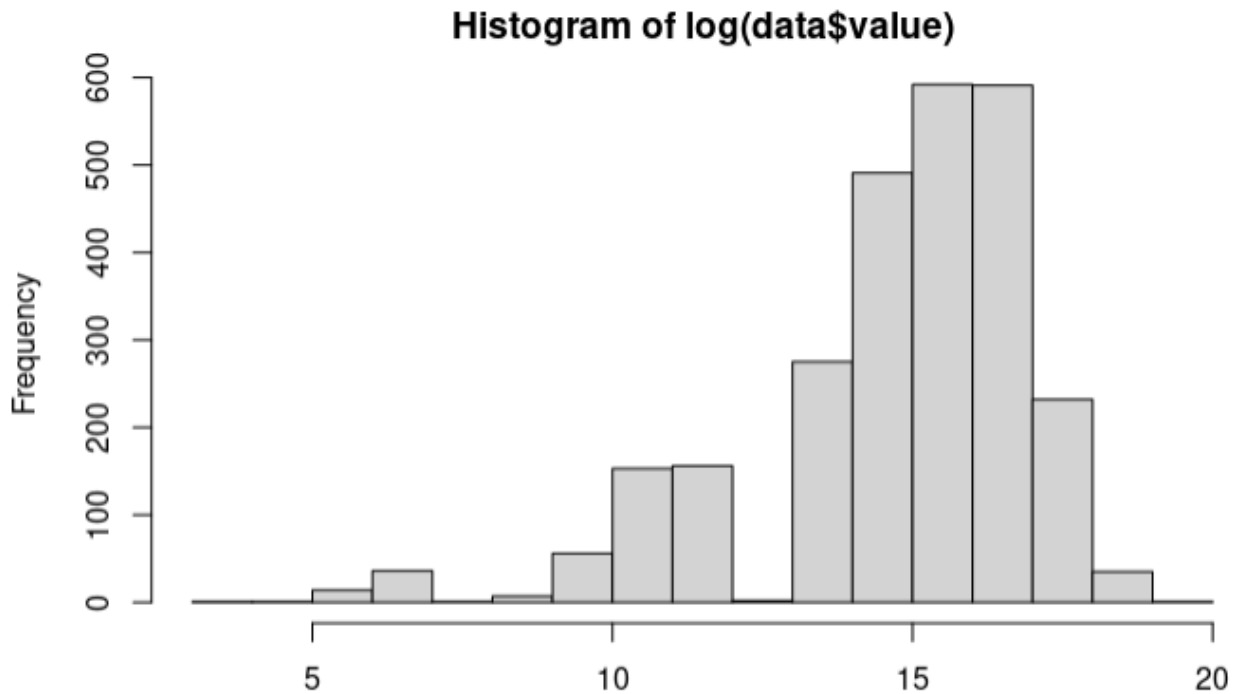

**Histogram of log(data$value)**

Figure1

In figure 1, we can see the distribution of players' value after log. It is not a standard normal distribution, the center is leaning to the right. However, our data is from the major leagues, so the lowest value players have a high probability to go to the second leagues next season. Thus we can drop some low value. And plot it again, we will get figure2, that show us the transfer value of players are basically follow normal distribution, so we can use linear regression to analyze these data.

Histogram of valuedata

Figure2
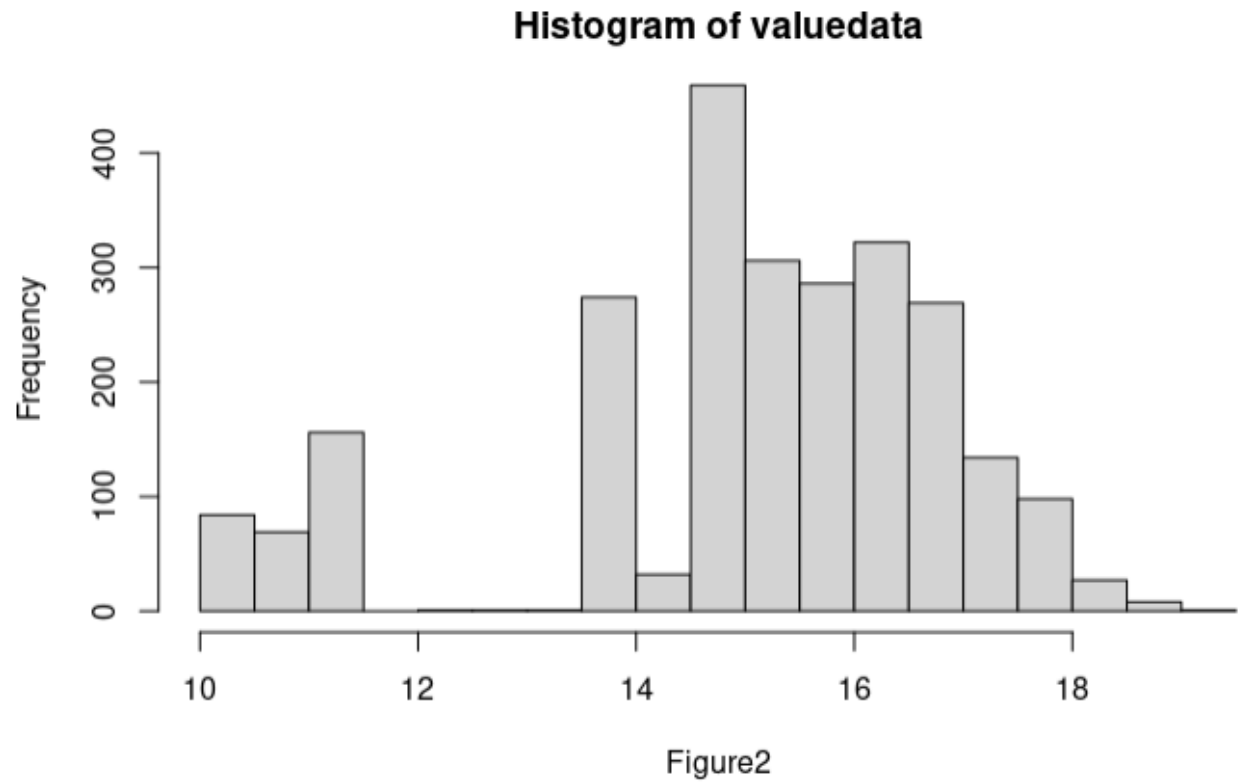
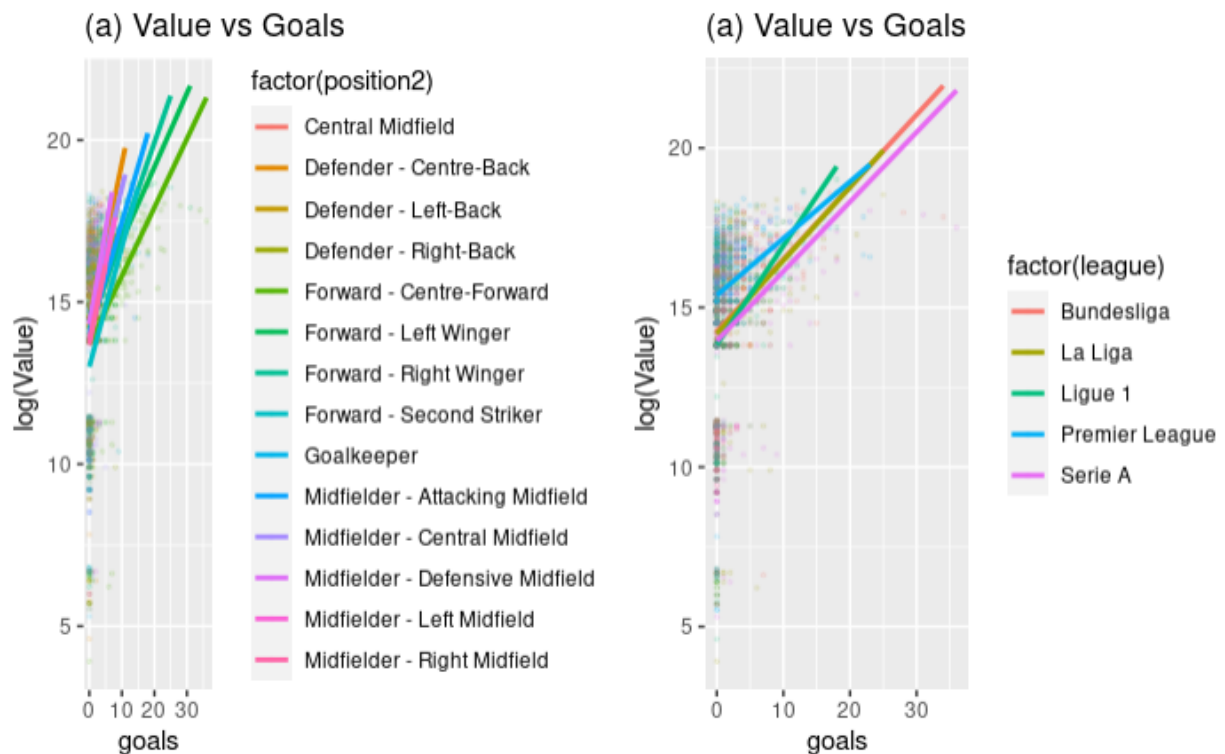As we mentioned above, we can do a linear regression for goals and value.
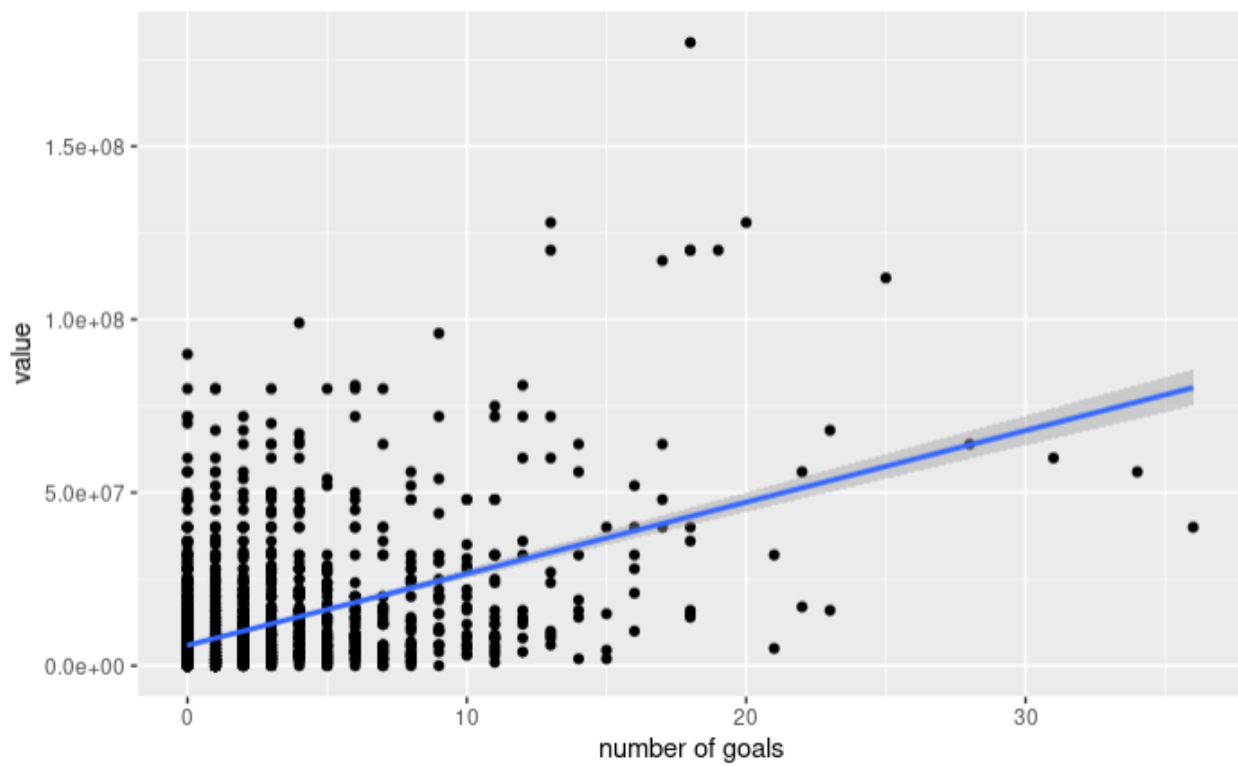
As the graphs showed that the value actually increases with the number of goals, it is not unexpected. In soccer matches, the score is small, it is seldom goals more than 3 for one team. Generally, just one goal can bring the winning for their team. Thus players who can get the goals should directly affect the result of the game. As long as we talked about goals, assists is a big topic we should not ignore.
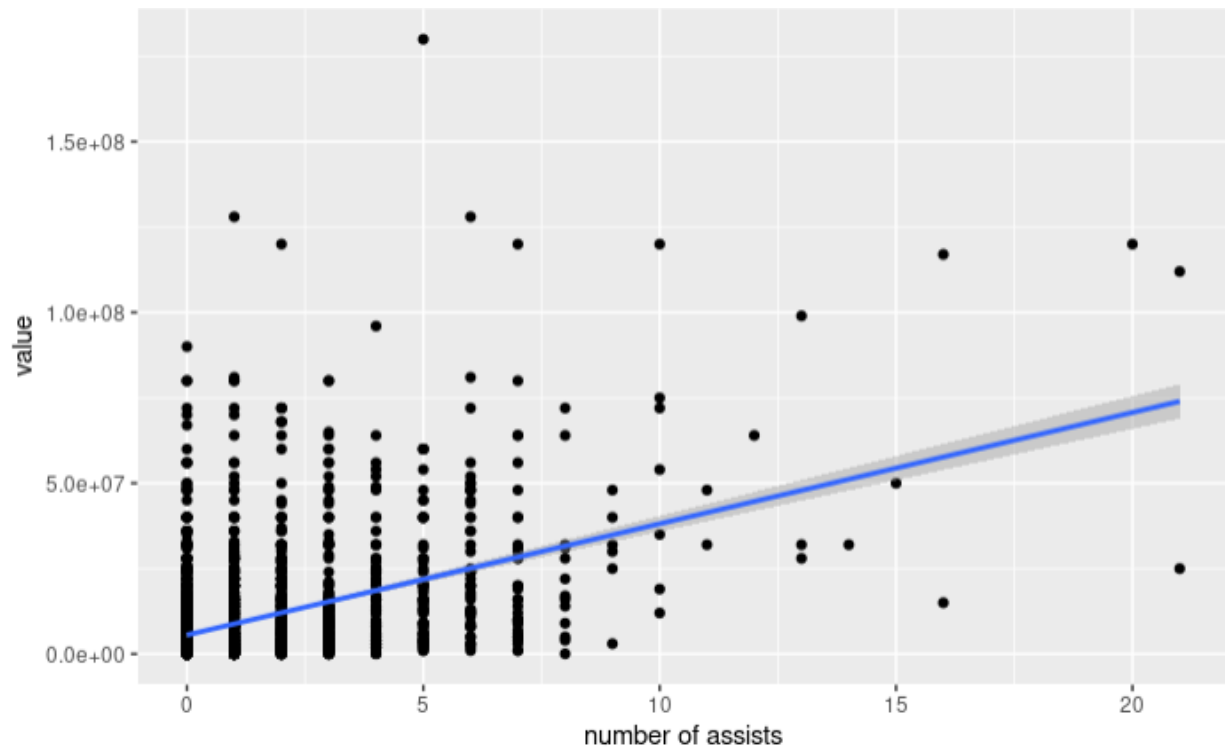

Based on the graphs, we can find that the assists also have the influence for value, and their slopes are similar. Thus, they are equally important for people to estimate the value of players.

Result

(a) Value vs Goals — factor(position2): Central Midfield, Defender - Centre-Back, Defender - Left-Back, Defender - Right-Back, Forward - Centre-Forward, Forward - Left Winger, Forward - Right Winger, Forward - Second Striker, Goalkeeper, Midfielder - Attacking Midfield, Midfielder - Central Midfield, Midfielder - Defensive Midfield, Midfielder - Left Midfield, Midfielder - Right Midfield

(a) Value vs Goals — factor(league): Bundesliga, La Liga, Ligue 1, Premier League, Serie A

After grouped these data and doing the multilevel analysis, we can find that different positions have an effect on value, even their goal numbers are the same. We can focus on the slope from positions, even if the forward has higher value and more goals, their value increases with goals slower than other positions. The biggest slope is from midfielder and back.

If grouped by leagues, we can find that ligue 1 players have the lowest value, but they have greater slope. It is in line with the real. Ligue 1 is from France, they have no competitive team except PSG paris, so their whole value is lower than other leagues. However some young players tend to go to Ligue 1 because they have more chances to show themselves. Their potential will be displayed to people, so their value increases quickly if they get some goals.

(b) Value vs Assist
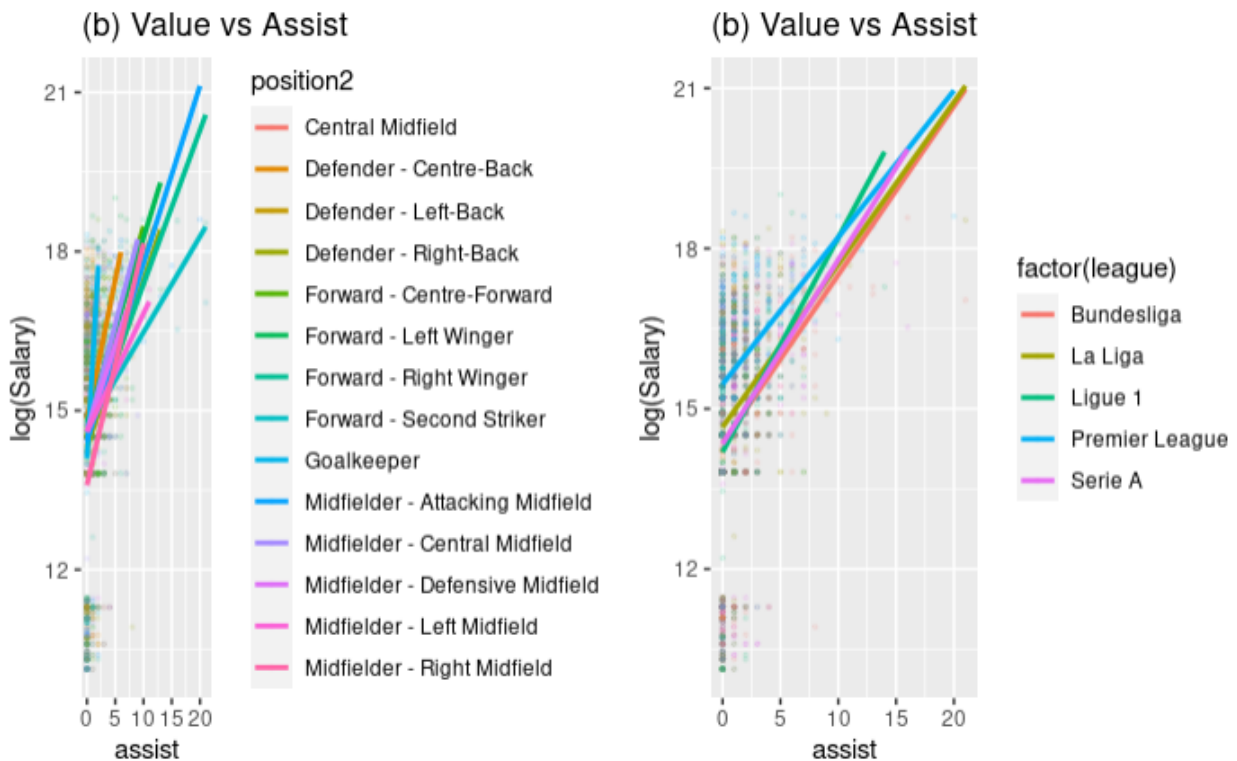
For assists, it is similar with the result we discussed previously like goals. The forward, especially winger, generally have higher value, but the slope is lower than midfielder except attacking midfield. The reason I consider is that the attacking midfield's responsibility is similar to forward winger, but in different areas on field. Their main responsibility is to join the attack and support forward, so they are easier to get assists than other midfielders who prefer to organize.

The league is almost the same with goals, I think the reason is also similar. Therefore we will not discuss it again.

Fit a multilevel model and try to explain the relationship among them.

According to the result of code:

```
stan_lmer(value ~ goals + dribbles + assists + (1 + goals + dribbles
+ assists | position2) + (1 + goals + dribbles + assists | league))
```

```
Fixed effects:
              Estimate Std. Error         df t value Pr(>|t|)
(Intercept) 1.340e+01  2.389e-01 7.838e+00   56.100 1.71e-11 ***
goals       4.091e-01  5.982e-02 2.332e+03    6.839 1.01e-11 ***
dribbles    5.096e-01  3.170e-02 2.184e+03   16.077  < 2e-16 ***
assists     2.941e-01  6.949e-02 2.520e+03    4.232 2.40e-05 ***
---
```

```
$position2
                                (Intercept)     goals   dribbles   assists
Central Midfield                   13.45328 0.4352380 0.4960250 0.2910719
Defender - Centre-Back             14.16407 0.4774109 0.5105139 0.2746459
Defender - Left-Back               13.34652 0.4178691 0.5258782 0.3051390
Defender - Right-Back              13.20737 0.4325815 0.4987951 0.3069437
Forward - Centre-Forward           13.35941 0.4493422 0.4457448 0.2474792
Forward - Left Winger              13.40792 0.4471239 0.4554458 0.2771296
Forward - Right Winger             13.39177 0.4425110 0.4382780 0.2889255
Forward - Second Striker           13.20010 0.4264440 0.4732047 0.2954629
Goalkeeper                         14.06359 0.4681578 0.4397152 0.2934421
Midfielder - Attacking Midfield    13.28741 0.4124540 0.5303802 0.3061514
Midfielder - Central Midfield      13.35859 0.3968783 0.5644712 0.3502935
Midfielder - Defensive Midfield    13.47741 0.4418553 0.5483676 0.2741490
Midfielder - Left Midfield         13.26504 0.4261925 0.4869055 0.2779719
Midfielder - Right Midfield        13.12236 0.4164413 0.4926733 0.2980108

$league
                (Intercept)     goals   dribbles   assists
Bundesliga         13.13826 0.4526829 0.5158732 0.3161553
La Liga            13.52321 0.4310108 0.4748498 0.2825381
Ligue 1            13.12441 0.4442823 0.4896348 0.3129592
Premier League     14.23580 0.4114654 0.4746765 0.2519522
Serie A            13.14099 0.4340977 0.5131150 0.2935643
```

We can fit a model with fixed effect: goals, assists, dribbles, and random effect : position and league.

log(Value) = 13.4 + 0.4091 goals + 0.5096 * dribbles + 0.2941 * assists*

And then depends their position, for example, if a player is forward - left winger,

log(Value) = 13.4 + 0.4493 goals + 0.4554 * dribbles + 0.2771 * assists*

Discussion

It is not difficult to get results, the value of players will be affected by their positions. Geneally, the forward and attacking midfielder has higher value, but at the same time they have more goals and assists. The midfielders, who prefer to organize or defend, have lower value than forwards, but they have a bigger slope, that means if they get goals and assists, their value increases faster than forward. For back and goal keepers, their main responsibility is defending, so the goal and assists cannot represent their level. To estimate their value, we have better pick another data to analyze.

Soccer is a complex sport, we cannot only use goals, assists to estimate a player, especially when they are midfielders, backs or goalkeepers. These dataset have a lot of variables, we can do more deep analysis about that to find the correlation between different variables and positions in the future.