Abstract

With more than X countries actively involved in the sport, Football is arguably the most widely played sport of all time. Each of these countries have several Professional Leagues for different levels of player skills. Perhaps the most anticipated Footballing event every year is the Champions League, which is an International Club League starring the top teams from Europe's best leagues. The Champions League final in 2019, by some estimates, had over 350 million viewers which is almost 3 times greater than the most watched Superbowl. This generates extremely high levels of revenue and investment into this sphere and consequently, Football (Soccer) players, for some time now, have been some of the world's highest paid athletes. With so many players across all leagues, I wondered what factors cause variances in their wages? It seems that it would depend not only on a player's innate ability, but also a host of other factors; for example, the amount of investment the players league receives or even the position they play. When players are often transferred from one club to another for a transfer fee.

1. Introduction

In this paper I will try to determine how a player's innate ability, position and league affect market value and weekly wages using OLS regression models. The models will be assessed for deviations from the core OLS assumptions to determine their reliability. I will also compare the models to each other using a variety of statistical tools and analysis to determine which model is the best to predict wage and market value. I will also compare the effects of the chosen independent variables on market value and wages and perhaps hypothesize and provide explanations for the same. To answer these questions, I will create and test several simple regression models.

2. Data set

The dataset I have chosen is scraped from the FIFA 19 video game and is publicly available. It contains 90+ player statistics such as 'Strength', 'Stamina' and 'Sprint Speed'. For the purpose of the models in this paper, only Position, Market Value, Weekly Wages, Strength, Stamina and Overall FIFA rating were selected. Overall FIFA rating is selected as a proxy variable for innate ability. The dataset was merged with another dataset to incorporate the "League" variable which categorizes each club into one of the 5 first division leagues and 'other' as shown in the table below.

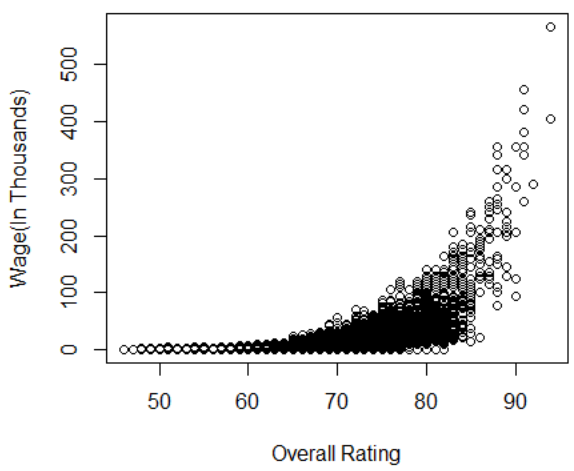
League Name	Description
La Liga	Spanish League
Bundesliga	German League
Major League Soccer (MLS)	American League
Ligue 1	French League
English premier League (EPL)	English League
Other	Miscellaneous Second and Third division teams and First division teams not from the 5 chosen countries.

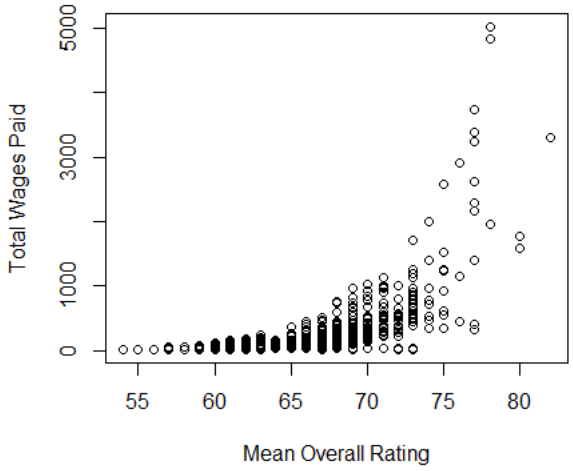
I also created and incorporated three more variables into the dataset as delineated by the table below. Concretely, all the variables used in subsequent models are as follows.

Variable	Description	Range	Units
Pace	Average of Agility, Running Speed and Acceleration.	0-99	
YAC	Years spent at current club	0-25	Years
ContractValid	Years left for current contract to expire	0-10	
League			
Spos	Position on the field played by player		
Market Value	Market Value Estimated by FIFA 19		Millions of EUR
Release Clause	Minimum amount (in EUR) specified in a player's contract needed to trigger negotiations for a transfer to another club.		Millions of EUR
Age	Age of the player		Years
Strength	FIFA 19 Strength rating	0 - 99	
Stamina	FIFA 19 Stamina rating	0 - 99	
Overall	FIFA 19 Overall Player Rating	0 - 99	
logrel	Log(Release Clause)		
logwage	Log(Wage)		
logval	Log(Market Value)		
Wage	Weekly wages of player		Thousands of EUR

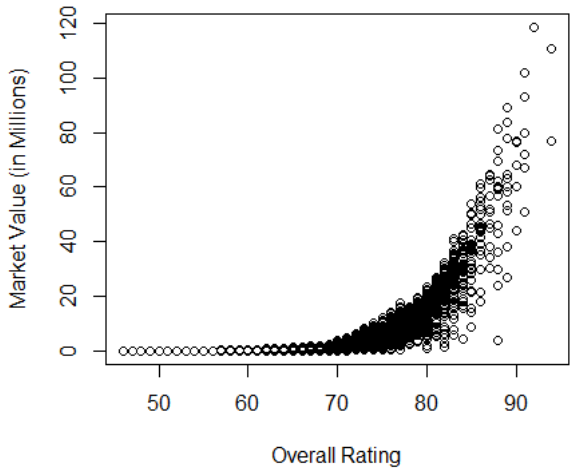
3. Preliminary Analysis and Sanity Check

One would expect players' overall ratings to have a positive relationship with their wages. This is evident in the scatter plot below. Moreover, the same trend can be observed at a higher level by plotting the total wages paid by a club against the average of all its players' ratings.



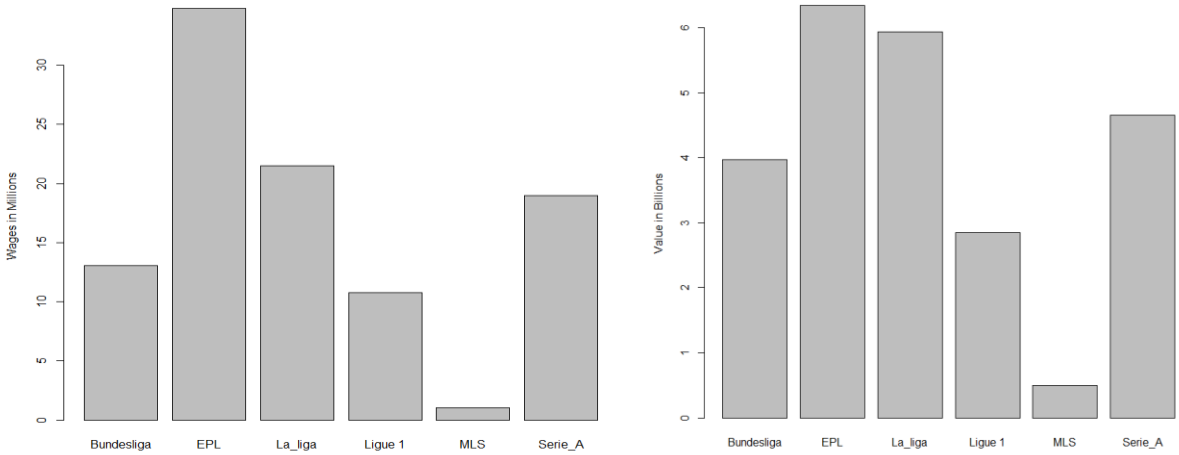


Interestingly, there seems to be a similar trends for players' market value. Unsurprisingly, players with low overall rations have low market values and players with high overall rating are mor likely to have high market values. It is important to note that the variance also is strictly increasing. I hypothesize that this is because many highly rated players are older and are approaching the end of their professional career and thus have a lower market value. Also, for I believe that for players rated between 80 and 90, the wage gap between different positions becomes increasingly significant.

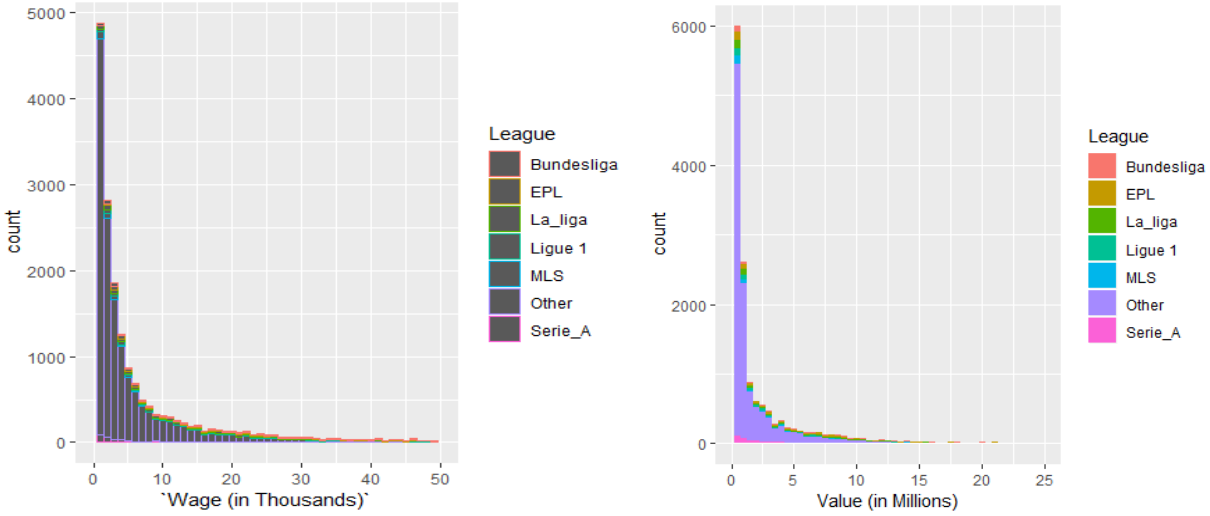


A closer look into the League variable reveals that there is a high descrepancy in the total wages paid in each league. All of these leagues have about the same number of teams (18-20), which implies that this descrepancy is caused by the amount of investment flowing into the leagues and perhaps, the level of revenue generated by the leagues through broadcasting rights and merchandise sales. Player Market Value on the other hand is influenced by a host of other factors as well, which makes the variation

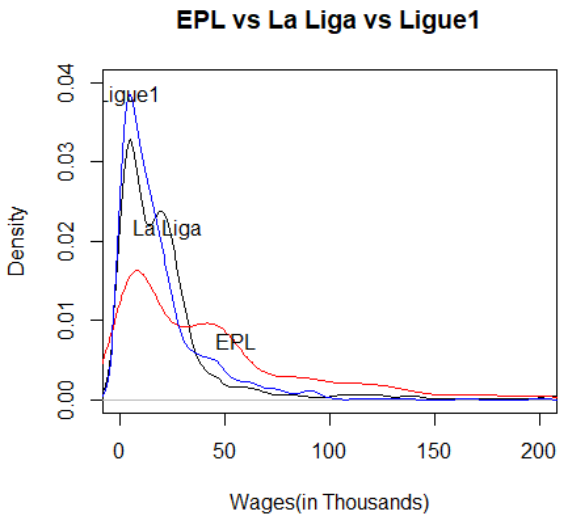
arising from difference in leagues not quite as dramatic as that of Wage. For example, the difference in total wages in the EPL and La Liga is quite large, but the difference in their total market value of players is not. The MLS, quite predictably, is the lowest in terms of wages paid as well as total market value of players.

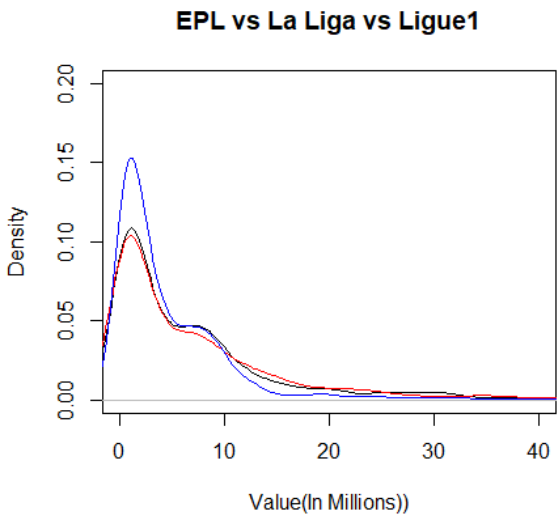


The Wage distribution is extremely left skewed due to the presence of significantly more low wage players than high wage players. This is because this dataset contains data from second and third division leagues in which wages are very low compared to first division leagues. The market value distribution is left skewed for the same reason. Applying a log transformation reduces the amount of right skew present in the distribution.

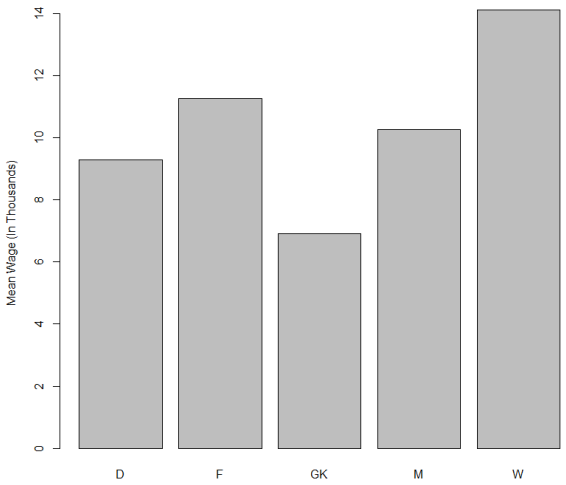


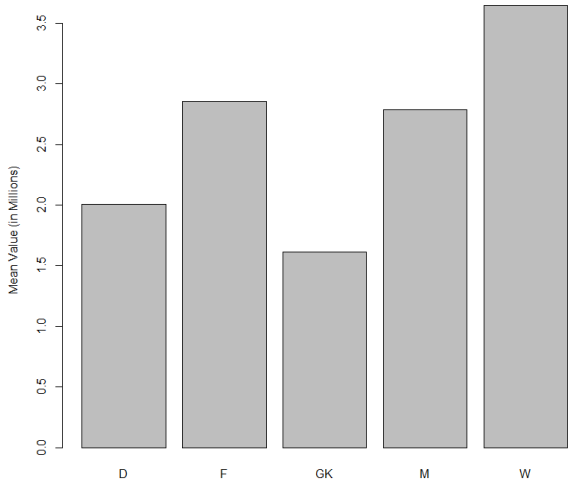
Isolating La Liga, Ligue 1 and the English Premier League, which are three of Europe's biggest leagues produces a more discernable wage distribution. The figure below indicates that the EPL tends to pay higher wages than the other two leagues. The Market Value distribution follows a very similar trend.



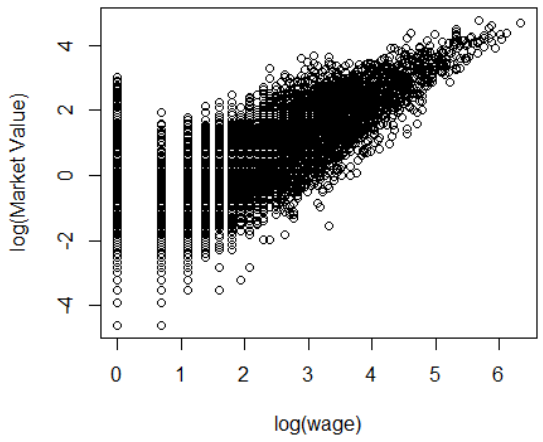


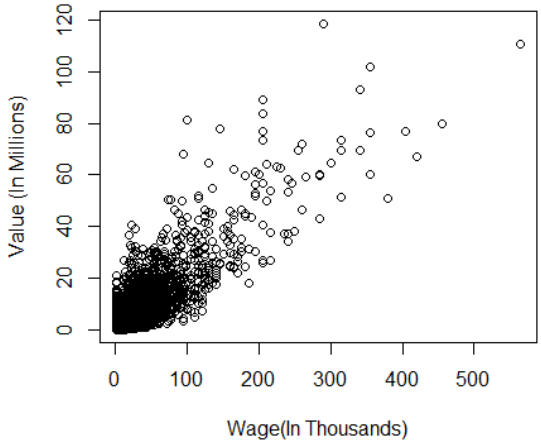
The dataset initially contained more than 20 detailed position classifications. For the purpose of this model I have broadened the classifications into just 5 categories: Defender(D), Forward(F), Midfielder(M), Winger(W) and Goalkeeper (GK). With the original classifications, there was a risk of creating fluctuations in wage that are caused by position in the model but cannot be reconciled with real life wages. For example, some players are often played interchangeably between similar position such as Center Attacking Midfielder (CAM) or Right Midfielder (RM). If these two narrow categories were left in the model, they would each have a different coefficient. This result would be meaningless because most CAM's can also be played as RM's and this trivial difference in position, in the real world, would be inconsequential. Midfielder and Winger are two separate categories because one could hypothesize that the requirements to succeed in these roles are inherently different and Clubs pay higher wages to players who meet these requirements. For example, Clubs generally tend to look for Wingers with high levels of Speed, Agility and Dribbling, whereas, for Midfielders they might be willing to pay more for Strength, Stamina and Athleticism. These are all variables available in the original dataset but excluded from the model. Separating Winger and midfielder might, to some extent, help the model capture the impacts of the excluded variables as well. The Position variable seems to impact market value as well in the same way. The "Defender" position seems to have a larger impact Wage than on Market Value. However, whether this difference is statistically significant is yet to be determined.





Based on the above revelations and supported by the plots below, one could conclude that wage and market value are positively correlated and are affected similarly by the chosen variables. This can roughly be reaffirmed from the figures below. The plot of Market Value against Wage demonstrates what appears to be a positive linear relationship between the variables. The dense cluster at the bottom left of the Wage vs Value plot indicates that there are many more instances of relatively low wages than of high wages. The lateral breaks/gaps in the log-log figure are perhaps due to the rounding scheme used for wages at the lower end of the spectrum. Although, the exact cause of this is unclear.





4. Relevant Questions and Hypotheses

From the previous section of this paper, one might vaguely draw the conclusion that Overall Rating, League and Position are important factors in predicting the Wage and Market Value of a player. Moreover, one can also construe that these variables have similar impacts on Wage and Market Value.

However, this paper will further explore the quantitative magnitude of these impacts as well as their statistical significance at the 5% level with the help of Ordinary Simple Regression models. Furthermore, additional variables deemed to be reasonably correlated with Wage and Market Value will also be incorporated into the models to reduce or perhaps entirely avoid any omitted variable bias, thereby making the estimators unbiased and trustworthy.

The Overall Rating of a player is determined by FIFA annually. Each annual update of this rating is based on every individual's performance in the most recent season prior to the update. Good luck can be a contributing factor to a player's success in a single game, but it is unlikely that it can impact their success through an entire season in a significant way. On the hand, however, bad luck can perhaps, impede a player from reaching their full potential by looming over them for an entire season. For example, Marco Reus of Dortmund spends a substantial portion of his career on the bench due to recurring injuries. As a result, he has a player rating of 86 and a wage of only 100,000€/week. Despite these noise factors, I believe, Overall FIFA rating is a robust proxy variable for innate ability and will therefore have a significant positive relationship with both Wage and Market Value. I also hypothesize that positions will have a significant impact on players wages and market value. Typically, in the real world, Forwards and Wingers tend to get paid higher wages than Defenders and Midfielders. A recent article by The Bleacher Report comments that "The old adage that strikers win matches and that defenders win leagues, seems to get lost every time a transfer window opens." regarding the 2019 transfer window. Indeed, this appears to be the case in modern football. The reason for this is that, Strikers are always in the public eye. Every aspect of a striker's or winger's actions has a large impact on the game. Strikers and Wingers score goals and scoring goals results in winning games. On the other hand, Defensive prowess is more subtle and is often very easily (and impetuously) dismissed as good or bad fortune. Most Club managers spent less than 2 seasons at a club and are therefore more focused on getting goals as opposed to building a team. This would make goalkeepers, in my opinion, the least paid among all positions. It would be interesting to compare the quantitative impact of being a Winger or a Striker on Wage and Market Value. I also conjecture that Release Clause will have a negative impact on Wages, but a positive impact on Market Value. I believe the reason for this is that Higher Release Clauses act as a de facto transfer barrier for players that makes it unlikely for other clubs to offer them higher wages than they are currently being payed. Concretely, players with high Release Clauses will have relatively lower wages because of a lack of competing offers. On the other hand, a Release Clause is also an indication of the inherent value a Club attributes to a player making it directly positively correlated with market Value.

Moreover, I think that the years left on a player's contract (ContractValid) will be positively correlated with market value because the closer a player gets to the end of their contract, the lower their transfer price will be. This idea stems from the fact that once a player's contract has ended, they can be transferred for free. Finally, I believe that Pace, Strength and Stamina should have positive impacts on both Wage and market Value because these are inherently sought-after attributes in football.

I will also include a variable called Years Spent at Club (YAC) because I suspect that it could have a positive player's Market Value. Staying at the same club for more years indicates that a player's performance in that club has been up to or beyond their expectations and this should in turn increase their market Value.

5. OLS Regression Models

5.1. Wage Regression Models

i. Wage Regression: Model 1

For the first regression, we have the following model:

$$\log(wage) = \beta_0 + \beta_1 Overall + \beta_2 spos + \beta_3 League + \beta_4 Age + \beta_5 \log rel + \mu$$

```
t test of coefficients:

                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -6.26752253  0.08396649 -74.6431 < 2.2e-16 ***
Overall        0.12423848  0.00125381  99.0891 < 2.2e-16 ***
sposF          0.15193542  0.01574010   9.6528 < 2.2e-16 ***
sposGK        -0.09206198  0.01676290  -5.4920 4.032e-08 ***
sposM          0.04175099  0.01221678   3.4175 0.0006335 ***
sposW          0.10920056  0.02611884   4.1809 2.918e-05 ***
LeagueEPL      0.52795068  0.02356420  22.4048 < 2.2e-16 ***
LeagueLa_liga -0.16074422  0.02787640  -5.7663 8.246e-09 ***
LeagueLigue 1 -0.11464639  0.02876390  -3.9858 6.755e-05 ***
LeagueMLS     -1.07463988  0.03223126 -33.3415 < 2.2e-16 ***
LeagueOther   -0.71447857  0.02134622 -33.4710 < 2.2e-16 ***
LeagueSerie_A -0.26114276  0.03852807  -6.7780 1.260e-11 ***
Age           -0.00058563  0.00128557  -0.4555 0.6487276
logrel        -0.01646580  0.00222602  -7.3970 1.460e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model is useful to establish a baseline for all the parameters and can be used to argue for the validity of subsequent models. The base case for the League variable used in this regression is Bundesliga and the base case for the Position(spos) variable is Defender. Moreover, it is important to note that all parameters barring Age is statistically significant at the 5% level based on the t-values and t-test for coefficients. Also, this model has an adjusted R^2 value of 72% which indicates that 72% of the variation in Wages is explained by the chosen variables.

ii. Wage Regression: Model 2

In the next model, we incorporate 5 more variable that I hypothesized should marginally impact Wage.

$$\log(wage) = \beta_0 + \beta_1 Overall + \beta_2 spos + \beta_3 League + \beta_4 Age + \beta_5 \log rel + \beta_6 YAC + \beta_7 ContractValid + \beta_8 Strength + \beta_9 Pace + \beta_{10} Stamina + \mu$$

```
t test of coefficients:

                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -6.16833948  0.08780197 -70.2529 < 2.2e-16 ***
Overall        0.12518014  0.00142423  87.8932 < 2.2e-16 ***
sposF          0.15076373  0.01634903   9.2216 < 2.2e-16 ***
sposGK        -0.15769188  0.02624963  -6.0074 1.925e-09 ***
sposM          0.03769172  0.01302896   2.8929 0.0038218 **
sposW          0.10215532  0.02746336   3.7197 0.0002001 ***
LeagueEPL      0.52928949  0.02365247  22.3778 < 2.2e-16 ***
LeagueLa_liga -0.16331878  0.02795657  -5.8419 5.258e-09 ***
LeagueLigue 1 -0.11432151  0.02893563  -3.9509 7.818e-05 ***
LeagueMLS     -1.06352118  0.03262209 -32.6013 < 2.2e-16 ***
LeagueOther   -0.70612432  0.02175717 -32.4548 < 2.2e-16 ***
LeagueSerie_A -0.26400799  0.03850544  -6.8564 7.311e-12 ***
Age            0.00098284  0.00139318   0.7055 0.4805332
logrel        -0.01669089  0.00222637  -7.4969 6.864e-14 ***
YAC           -0.00206851  0.00275630  -0.7505 0.4529860
ContractValid  0.01584091  0.00431206   3.6736 0.0002399 ***
Strength      -0.00113409  0.00054306  -2.0883 0.0367820 *
Pace          -0.00099291  0.00060274  -1.6473 0.0995106 .
Stamina       -0.00090150  0.00056466  -1.5965 0.1103867
---
```

The result of adding the five variables (YAC, ContractValid, Pace, Strength and Stamina) to the model are as follows:
The Adjusted R^2 value for this model is still 72%.

Result	Interpretation/Comment
The coefficient for sposGK drops from -0.09 to -0.15.	Playing as a Goalkeeper now reduces wage by 15% (from 9%) compared to playing as a defender.
None of the other coefficients have drastic changes. Interpretations of these coefficients is done in the Model 3 section.	All the new variables have low correlation with the variables from Model 1
Strength has negative coefficient and is statistically significant at the 5% level. This result is contrary to the initial hypothesis.	This result might be coincidental because it is highly counter-intuitive
ContractValid has a positive coefficient and is significant at the 5% level.	An additional year of contract length seems to increase wage by 1.6%.
Pace and Stamina also have negative coefficients that are contrary to the initial hypothesis but are insignificant at the 5% level.	This result is also highly counter-intuitive and is possibly coincidental.

iii. Wage Regression: Model 3

In the final model we incorporate every variable used in all prior models as well as an interaction term. The inclusion of the interaction term is in response to the theory that different leagues might differently value different positions based on the dominant playing style of the

league. For example, the Italian League has been known to produce excellent defenders due its overall defensive style gameplay.

$$\log(wage) = \beta_0 + \beta_1 Overall + \beta_2 spos + \beta_3 League + \beta_4 Age + \beta_5 \log rel + \beta_6 YAC + \beta_7 ContractValid + \beta_8 Strength + \beta_9 Pace + \beta_{10} Stamina + \beta_{11} spos * League + \mu$$

t test of coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-6.17962854	0.09284031	-66.5619	< 2.2e-16	***
Overall	0.12498285	0.00142845	87.4957	< 2.2e-16	***
sposF	0.18203040	0.05441361	3.3453	0.0008237	***
sposGK	-0.23846109	0.06369488	-3.7438	0.0001819	***
sposM	0.08542629	0.04288447	1.9920	0.0463863	*
sposW	0.11934751	0.09555047	1.2491	0.2116637	
LeagueEPL	0.56363348	0.04397959	12.8158	< 2.2e-16	***
LeagueLa_liga	-0.13817899	0.05156002	-2.6800	0.0073703	**
LeagueLigue 1	-0.06797946	0.05286070	-1.2860	0.1984571	
LeagueMLS	-1.04303371	0.05331831	-19.5624	< 2.2e-16	***
LeagueOther	-0.69746474	0.03750981	-18.5942	< 2.2e-16	***
LeagueSerie_A	-0.24726088	0.06984228	-3.5403	0.0004008	***
Age	0.00128536	0.00139689	0.9202	0.3575036	
logrel	-0.01718824	0.00223127	-7.7034	1.4e-14	***
YAC	-0.00260387	0.00276304	-0.9424	0.3460046	
ContractValid	0.01583010	0.00431691	3.6670	0.0002462	***
Strength	-0.00107545	0.00054443	-1.9754	0.0482429	*
Pace	-0.00085497	0.00060519	-1.4127	0.1577518	
Stamina	-0.00096677	0.00056674	-1.7059	0.0880547	.
sposF:LeagueEPL	-0.08066569	0.07477874	-1.0787	0.2807260	
sposGK:LeagueEPL	-0.11994636	0.08316203	-1.4423	0.1492307	
sposM:LeagueEPL	-0.05499928	0.05679882	-0.9683	0.3329000	
sposW:LeagueEPL	0.21311455	0.11954530	1.7827	0.0746518	.
sposF:LeagueLa_liga	-0.03337217	0.09040734	-0.3691	0.7120346	
sposGK:LeagueLa_liga	-0.10403662	0.10035281	-1.0367	0.2998868	
sposM:LeagueLa_liga	-0.06743123	0.06574488	-1.0256	0.3050715	
sposW:LeagueLa_liga	0.36394485	0.15308476	2.3774	0.0174461	*
sposF:LeagueLigue 1	-0.05130118	0.08819211	-0.5817	0.5607780	
sposGK:LeagueLigue 1	-0.09114810	0.09603402	-0.9491	0.3425718	
sposM:LeagueLigue 1	-0.06584493	0.07011941	-0.9390	0.3477239	
sposW:LeagueLigue 1	-0.10925356	0.15709082	-0.6955	0.4867641	
sposF:LeagueMLS	-0.04719504	0.08945306	-0.5276	0.5977871	
sposGK:LeagueMLS	0.00890282	0.11168174	0.0797	0.9364641	
sposM:LeagueMLS	0.02291018	0.07246537	0.3162	0.7518900	
sposW:LeagueMLS	-0.45582155	0.18862020	-2.4166	0.0156765	*
sposF:LeagueOther	-0.02424441	0.05724655	-0.4235	0.6719297	
sposGK:LeagueOther	0.11754643	0.06326794	1.8579	0.0631988	.
sposM:LeagueOther	-0.04982782	0.04486127	-1.1107	0.2667097	
sposW:LeagueOther	-0.05654908	0.09943795	-0.5687	0.5695762	
sposF:LeagueSerie_A	-0.16223053	0.11723256	-1.3838	0.1664275	
sposGK:LeagueSerie_A	-0.01030148	0.12962866	-0.0795	0.9366604	
sposM:LeagueSerie_A	-0.02983794	0.09490921	-0.3144	0.7532333	
sposW:LeagueSerie_A	0.25553509	0.17598818	1.4520	0.1465201	

The Results of adding the interaction term as follows:

Result	Interpretation
The coefficient of sposGK drops even further from -0.16 to -0.24.	Playing as a Goalkeeper now reduces wage by 24% (from 15%) compared to playing as a defender.

The coefficient of sposF rises from 0.15 to 0.18.	Playing as a Forward now increases wage by 18% (from 15%) compared to playing as a defender.
The coefficient of sposM rises from 0.037 to 0.085	Playing as a Midfielder now increases wage by 8.5% (from 3.7%) compared to playing as a defender.
None of other coefficients have dramatic fluctuations.	All the new variables have low correlation with the variables from Model 2.
SposW:MLS and sposW:LaLiga are statistically significant negative and positive coefficients respectively.	Playing as a Winger might indicate a change in wage depending on the league one is playing at. For MLS there is a decrease in wage of about 46% from the base case, whereas we can see an increase in wage of 36% from the base case for Wingers playing in La Liga, the Spanish first division league.
The coefficient for Ligue1 increases from -0.11 to -0.07. It is also no longer statistically significant at the 5% level.	Playing in League1 results in a drop of 11% in wage compared to the Bundesliga.
Overall rating has consistently had a coefficient of 0.125 which is significant at the 5% level for all three models.	This shows that the interaction term as well as the variables added in model 2 are not closely correlated with Overall rating. Also, an additional unit of overall rating results in about a 12.5% increase in wage.

The Adjusted R^2 value for this model is still 72%

Model Comparison:

While comparing models, I notice that the R^2 value is about 72% for all of them. This indicates that none of the variable the were added were better at explaining the variance in wage. However, based on the model results, highlighted in the previous section, it is evident that some of the coefficients change when new variables are added. This can because of three one of reasons or some combination of these reasons.

- It can be simply due to random fluctuations in the relevant dependent variable subsets due to additions of new variables.
- It can be due to the correlation between independent variables (low levels of multicollinearity). Concretely, it can be because some of the added variables need to be controlled for in order to observe the true amount of impact of one of the variables on wage.

From my understanding of the real world of football, I strongly hypothesized that the impact of position on wage, to some extent, depend on the League of the player. From model 3 we observe that for some combinations of league and position (SposW:MLS and sposW:LaLiga) this hypothesis is statistically significant at the 5% level.

Moreover, using the “Analysis of variance tables method” with the “Anova()” function on R, I have compared the models with each other. The results are as follows.

```
Analysis of Variance Table

Model 1: logwage ~ Overall + spos + League + Age + logrel + YAC + ContractValid +
    Strength + Pace + Stamina
Model 2: logwage ~ Overall + spos + League + Age + logrel + YAC + ContractValid +
    spos * League + Strength + Pace + Stamina
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  16624 7016.7
2  16600 6992.1 24    24.607 2.4341 0.0001086 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Model 1: logwage ~ Overall + spos + League + Age + logrel
Model 2: logwage ~ Overall + spos + League + Age + logrel + YAC + ContractValid +
    spos * League + Strength + Pace + Stamina
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  16629 7028.1
2  16600 6992.1 29    36.024 2.9491 1.843e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Model 1: logwage ~ Overall + spos + League + Age + logrel
Model 2: logwage ~ Overall + spos + League + Age + logrel + YAC + ContractValid +
    Strength + Pace + Stamina
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1  16629 7028.1
2  16624 7016.7  5    11.417 5.41 5.626e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since in each case the p value resulting from the F statistic is less than 0.05, we can reject the null hypothesis that the simpler model is better at the 5% level.

Therefore, I have decided to conclude that Model 3(the most complex model), is the best model among the three to describe the impacts of the independent variables on wage.

5.2 Market Value Regression Models

To model the Market Value of players, we use the same variables as the model 2 under Wage regression. Now, we have two models: one with the interaction term and one without.

i. Market Value Regression: Model 1

$$\log(mktval) = \beta_0 + \beta_1 Overall + \beta_2 spos + \beta_3 League + \beta_4 Age + \beta_5 \log rel + \beta_6 YAC + \beta_7 ContractValid + \beta_8 Strength + \beta_9 Pace + \beta_{10} Stamina + \beta_{11} spos * League + \mu$$

```
t test of coefficients:

                      Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)         -1.2367e+01  3.6656e-02 -337.3933 < 2.2e-16 ***
Overall              2.1813e-01  5.5970e-04  389.7283 < 2.2e-16 ***
sposF                2.0465e-01  3.0911e-02    6.6208 3.682e-11 ***
sposGK              -6.4761e-02  3.5478e-02   -1.8254 0.0679598 .
sposM                1.0316e-01  2.3992e-02    4.3000 1.718e-05 ***
sposW                7.9294e-02  4.0074e-02    1.9787 0.0478657 *
LeagueEPL            7.4488e-03  2.2483e-02    0.3313 0.7404123
LeagueLa_liga       -2.8011e-02  2.4162e-02   -1.1593 0.2463497
LeagueLigue 1       -1.5204e-02  2.4936e-02   -0.6097 0.5420672
LeagueMLS           -1.0571e-01  2.8363e-02   -3.7272 0.0001943 ***
LeagueOther         -6.1074e-02  1.8541e-02   -3.2941 0.0009894 ***
LeagueSerie_A       -3.8189e-02  2.6058e-02   -1.4656 0.1427873
Age                 -9.4941e-02  8.2298e-04 -115.3616 < 2.2e-16 ***
YAC                 -6.3977e-03  1.3187e-03   -4.8516 1.236e-06 ***
ContractValid        1.2438e-03  1.6299e-03    0.7631 0.4454182
logrel               3.9092e-03  7.6711e-04    5.0960 3.508e-07 ***
Strength            -1.3037e-04  2.0913e-04   -0.6234 0.5330406
Pace                 3.6514e-04  2.5764e-04    1.4173 0.1564261
Stamina              1.0982e-03  2.3581e-04    4.6569 3.235e-06 ***
sposF:LeagueEPL     -2.5358e-02  4.0670e-02   -0.6235 0.5329598
sposGK:LeagueEPL    -7.1964e-02  5.7867e-02   -1.2436 0.2136626
sposM:LeagueEPL      1.3523e-02  3.0902e-02    0.4376 0.6616716
sposW:LeagueEPL     -1.2840e-02  6.2209e-02   -0.2064 0.8364814
sposF:LeagueLa_liga -4.0420e-02  4.1184e-02   -0.9814 0.3263909
sposGK:LeagueLa_liga -4.5215e-02 4.8197e-02   -0.9381 0.3481986
sposM:LeagueLa_liga  4.3438e-02  3.2351e-02    1.3427 0.1793871
sposW:LeagueLa_liga  9.5876e-03  5.5223e-02    0.1736 0.8621700
sposF:LeagueLigue 1 -3.5009e-02  4.5501e-02   -0.7694 0.4416577
sposGK:LeagueLigue 1 -4.9297e-02 5.0741e-02   -0.9715 0.3312987
sposM:LeagueLigue 1  6.7266e-03  3.4305e-02    0.1961 0.8445492
sposW:LeagueLigue 1 -3.6571e-02  6.4855e-02   -0.5639 0.5728347
sposF:LeagueMLS      6.2725e-02  4.9024e-02    1.2795 0.2007463
sposGK:LeagueMLS     7.7437e-02  6.0354e-02    1.2830 0.1994992
sposM:LeagueMLS      1.5842e-01  3.8906e-02    4.0718 4.686e-05 ***
sposW:LeagueMLS      1.2396e-01  6.5131e-02    1.9032 0.0570288 .
sposF:LeagueOther    2.2195e-02  3.1426e-02    0.7063 0.4800294
sposGK:LeagueOther   1.2330e-04  3.5730e-02    0.0035 0.9972467
sposM:LeagueOther    4.5923e-02  2.4403e-02    1.8818 0.0598752 .
sposW:LeagueOther    7.9866e-02  4.1086e-02    1.9439 0.0519281 .
sposF:LeagueSerie_A  3.3556e-02  4.5520e-02    0.7372 0.4610305
sposGK:LeagueSerie_A -8.2333e-02 6.7026e-02   -1.2284 0.2193240
sposM:LeagueSerie_A  4.9104e-02  3.3718e-02    1.4563 0.1453207
sposW:LeagueSerie_A  8.3161e-02  5.4846e-02    1.5163 0.1294704
---
```

This model yields some interesting results.

The adjusted R^2 for this model is 97%

Result	Interpretation
Overall rating has a positive coefficient that is significant at the 5% level.	An incremental unit of Overall Rating results in a 21.8% increase in market value. This is a reasonable result and in line with the initial hypothesis.
The coefficient for sposM:MLS is also significant at the 5% level and has a positive coefficient. This was not initially anticipated.	This result suggests that being a midfielder in the MLS adds an additional 15.8% in Market Value as opposed to the base case.
The coefficient for Age is also significant at the 5% level and has a deeply negative impact on Market value.	An additional year of Age results in a lower market value by 9.5%. Barring a few outliers, this result is consistent with the real world because older players tend to have fewer years of their professional careers remaining

	and therefore have lower market values.
The coefficients for MLS and "Other" are also significant at the 5% level and are negative values.	Playing in the MLS or in Other Smaller leagues results in a lower market value by 6.5% and 10.5% respectively compared to the Bundesliga
The coefficients for sposM and SposF are significant at the 5% level and are positive values.	Being a midfielder or a forwards increases market value by about 20% and 10% respectively. This is in line with the initial hypothesis.

ii. Market Value Regression: Model 2

$$\log(mktval) = \beta_0 + \beta_1 Overall + \beta_2 spos + \beta_3 League + \beta_4 Age + \beta_5 \log rel + \beta_6 YAC + \beta_7 ContractValid + \beta_8 Strength + \beta_9 Pace + \beta_{10} Stamina + \mu$$

```
t test of coefficients:

                 Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)   -1.2384e+01  3.3653e-02 -367.9952 < 2.2e-16 ***
Overall        2.1811e-01  5.5877e-04  390.3389 < 2.2e-16 ***
sposF          2.2144e-01  5.6192e-03   39.4080 < 2.2e-16 ***
sposGK        -7.3151e-02  1.2956e-02   -5.6461 1.668e-08 ***
sposM          1.4715e-01  4.6358e-03   31.7432 < 2.2e-16 ***
sposW          1.4778e-01  8.9677e-03   16.4795 < 2.2e-16 ***
LeagueEPL      6.9594e-04  1.3517e-02    0.0515 0.9589373
LeagueLa_liga -2.2134e-02  1.3448e-02   -1.6458 0.0998153 .
LeagueLigue 1 -2.5731e-02  1.4497e-02   -1.7749 0.0759347 .
LeagueMLS     -2.7322e-02  1.6845e-02   -1.6219 0.1048347
LeagueOther   -3.7952e-02  1.0772e-02   -3.5232 0.0004275 ***
LeagueSerie_A -2.1441e-02  1.4860e-02   -1.4429 0.1490665
Age           -9.4979e-02  8.2276e-04 -115.4389 < 2.2e-16 ***
YAC           -6.2898e-03  1.3163e-03   -4.7783 1.783e-06 ***
ContractValid  1.2729e-03  1.6286e-03    0.7816 0.4344572
logrel         3.7814e-03  7.6726e-04    4.9284 8.372e-07 ***
Strength      -1.1644e-04  2.0883e-04   -0.5576 0.5771449
Pace           3.3075e-04  2.5673e-04    1.2883 0.1976546
Stamina        1.1025e-03  2.3521e-04    4.6873 2.790e-06 ***
```

Excluding the interaction term yields the following results.

The adjusted R^2 for this value is still 97%.

- sposF, sposGK, sposM and sposW are now statistically significant at the 5% level. This shows that adding the interaction term induces a lot of variance for each position within a certain league.
- Age has a negative impact on market value and is still significant at the 5% level.

Model Comparison:

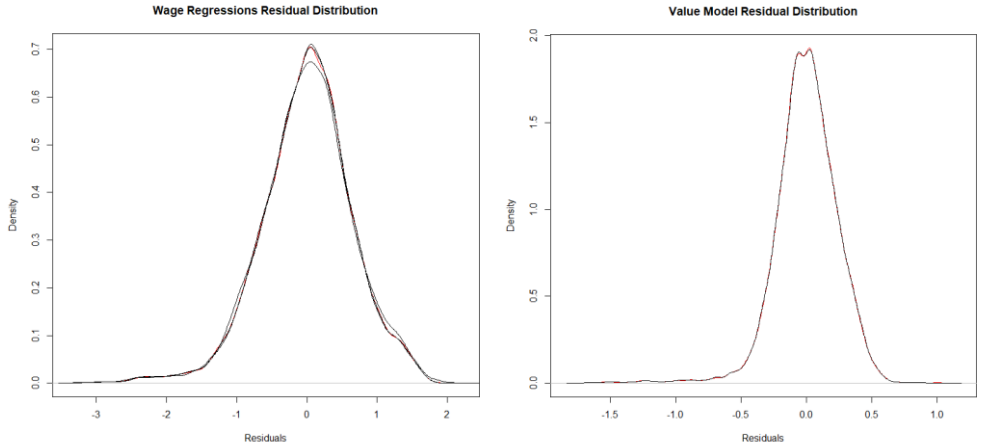
Once again, I compare the models using the ANOVA function.

```
Analysis of Variance Table

Model 1: logwage ~ Overall + spos + League + Age + logrel
Model 2: logwage ~ Overall + spos + League + Age + logrel + YAC + ContractValid +
    Strength + Pace + Stamina
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1  16629 7028.1
2  16624 7016.7  5    11.417 5.41 5.626e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

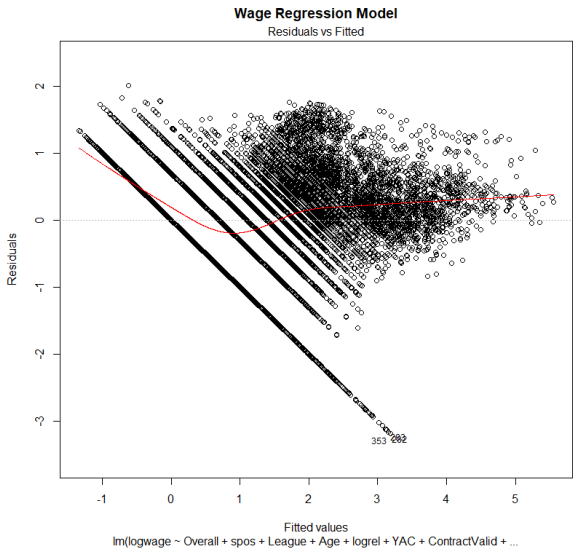
At the 5% significance level we can reject the null hypothesis that the simpler model (Model 2) is better. However, at the 1% level we fail to reject it.

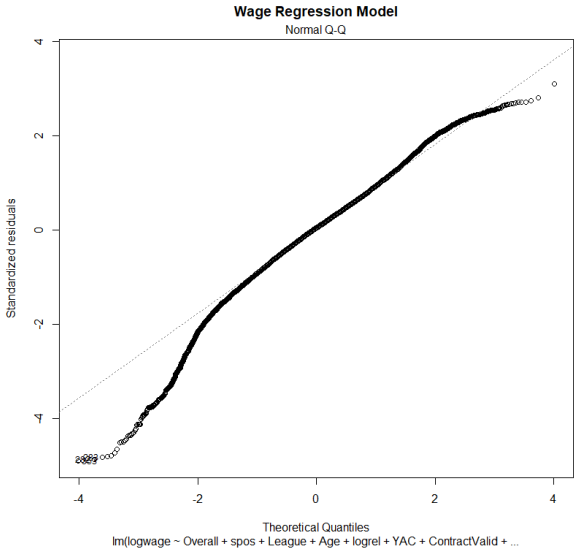
Robustness and assumption testing



The figures above show the distrubution of the residuals from all 3 wage regression models as well as both Market value Regression models. It is difficult the discern the different curves because they very precisely overlap one another. In any case, the figures indicate that the distribution of residuals is centered almost symmetrically around 0. This confirms, to a reasonable extent, that the mean of the residuals in all the aforementioned models is 0 and therefore a critical assumption in ordinary linear regression models is not violated.

Wage Regression

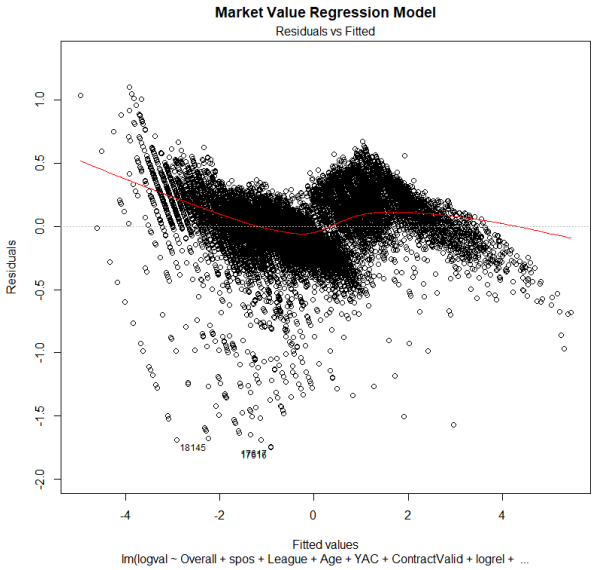


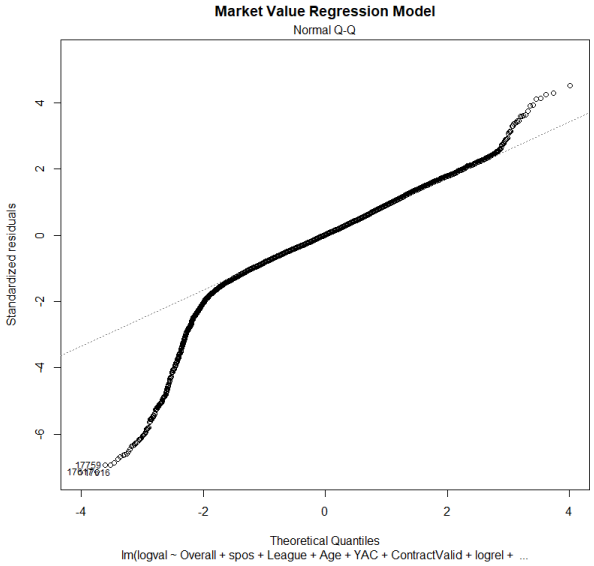


The Residual vs Fitted values plot above indicates that although the residuals follow a distribution centered at 0, it is not, on average, 0 for all levels of fitted values. Concretely, for lower levels of predicted wage, the residuals, on average, are greater than 0 and eventually converge to 0 towards higher levels of predicted wage. The reason for this behavior is quite likely to be due to the strange 'parallel lines' trend of the residuals for lower levels of predicted wage. This could be due to abnormalities in the data or an oversight in the model. In my understanding, it could also be a consequence of my dependent variable not being purely continuous. Perhaps, for lower values of wage, the wage variable in the data set is rounded in a way that leaves large gaps between successive values. It could also be a consequence of having an omitted variable that only significantly impacts the model at low levels of predicted wage. Correcting this would, however, result in a non-linear model and is therefore, out of scope for this paper. Clearing this abnormality, in my opinion, can greatly improve the value of this analysis.

The Q-Q plot above indicates the presence of slight heteroskedasticity since the quantiles of the residual distribution do not quite completely line up with the theoretical quantiles. This is specifically true for the lower quantiles. Once again, this is due to the aforementioned 'parallel lines' trend for lower levels of wage visible in the residuals vs fitted plot. This theoretically violates one of the core assumptions of OLS regressions. However, since this dataset contains 18,000+ instances, one could dismiss this issue by invoking the central limit theorem, which states that for a large enough data set, the predicted coefficients should be normally distributed.

Market Value Regression





In this regression model as well, it is evident that there is a strange "parrelel lines" trend in residuals for lower levels of predicted market value (although to a lower extent compared to the Wage regression model). The red line in the residual vs fitted is not horizontal at any level of predicted wage, indication that although the distribution of residuals is centered around 0, it is not, on average, 0 for any level of predicted wage.

The Q-Q plot for this model appears to be a straight line for higher quantiles but not quite so for lower quantiles. Once again, I hypothesize that this is due to the 'parallel lines' trend of residuals seen in the residual vs fitted plot. The reasons for this, are the same as those hypothesized above, in the wage regression analysis section.

Note on Robust coefficient testing:

Despite the heteroskedasticity in the models uncovered in the previous section, I believe that the coefficient t-tests are sufficiently reliable because they have been tested using standardized errors derived from a variance-covariance matrix. In R, this can be achieved using the 'Sandwich' package and its "coeftest" function. This achieves the same result as using the 'robust' command while running regressions in STATA.

8. Further exploration

- In wage regression model 3, I incorporated an interaction term between position and League that yielded some results that I did not anticipate. In order to further explore this phenomenon, I applied wage regression model 2 to three of Europe's top leagues and thereby removed the League variable from the equation. The summary of the results is as follows:

EPL

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-7.7292457	0.1725818	-44.7860	< 2.2e-16	***
Overall	0.1493220	0.0036548	40.8568	< 2.2e-16	***
sposF	0.1055943	0.0467295	2.2597	0.0241388	*
sposGK	-0.3975137	0.0885891	-4.4872	8.402e-06	***
sposM	0.0313500	0.0359696	0.8716	0.3837344	
sposW	0.2355067	0.0717760	3.2811	0.0010838	**
Age	0.0017354	0.0053618	0.3237	0.7462903	
logrel	0.0302703	0.0090570	3.3422	0.0008742	***
YAC	0.0200114	0.0065463	3.0569	0.0023194	**
ContractValid	-0.0109970	0.0145036	-0.7582	0.4485632	
Strength	0.0012517	0.0015960	0.7843	0.4331265	
Pace	0.0016894	0.0017286	0.9773	0.3287353	
Stamina	-0.0036022	0.0018457	-1.9517	0.0513672	.

Ligue 1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-9.16709774	0.24846882	-36.8944	< 2.2e-16	***
Overall	0.15982719	0.00516935	30.9182	< 2.2e-16	***
sposF	0.10958597	0.06322738	1.7332	0.083678	.
sposGK	-0.32750747	0.11334730	-2.8894	0.004028	**
sposM	-0.00738661	0.04823690	-0.1531	0.878356	
sposW	0.06445780	0.11763576	0.5479	0.583976	
Age	-0.00091376	0.00582254	-0.1569	0.875359	
logrel	0.03061026	0.01275194	2.4004	0.016741	*
YAC	0.02910087	0.01076381	2.7036	0.007094	**
ContractValid	0.02880528	0.01979062	1.4555	0.146161	
Strength	0.00246095	0.00245537	1.0023	0.316699	
Pace	0.00280328	0.00246659	1.1365	0.256293	
Stamina	-0.00371507	0.00214428	-1.7325	0.083796	.

La Liga

t test of coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-8.70948076	0.23086870	-37.7248	< 2.2e-16	***
Overall	0.16355829	0.00476472	34.3270	< 2.2e-16	***
sposF	0.12445365	0.07065430	1.7614	0.0786466	.
sposGK	-0.28302959	0.09903025	-2.8580	0.0044034	**
sposM	-0.02550346	0.04680633	-0.5449	0.5860336	
sposW	0.41045250	0.11574792	3.5461	0.0004199	***
Age	-0.01646619	0.00605957	-2.7174	0.0067603	**
logrel	-0.01005304	0.01198754	-0.8386	0.4019976	
YAC	0.00886647	0.00811027	1.0932	0.2747052	
ContractValid	-0.01406055	0.01726723	-0.8143	0.4157850	
Strength	-0.00250999	0.00228704	-1.0975	0.2728472	
Pace	0.00027785	0.00216498	0.1283	0.8979235	
Stamina	-0.00040038	0.00229232	-0.1747	0.8614009	

We note that the signs of several coefficients (including position i.e. spos) switch when the League changes. It is interesting to think about whether these changes are statistically significant. Wage regression Model 3 already tests the interaction between Position and League and deems some of this fluctuation to be statistically insignificant. This can be tested by

incorporation more interaction terms into the model. Further, the same exercise can be done with the Market Value regression Model 2 as well.

- The robustness of the model can also be tested by using only half of the available dataset to build the model and the other half of the dataset to test to accuracy of its wage/market value predictions.
- Wages and Market Value are likely also impacted by the incremental revenue generated by a player for his/her club. This incremental revenue ultimately hinges on the level of popularity a player has in the footballing world. Incorporating a variable like "Instagram followers" or "Jersey Sales" could help determine this effect.

9. Model 1 vs Model 2 comparison

- Overall rating seems to have a much higher impact on market value than on wage.
- Age does not have a significant impact on wage but does have a significant negative impact on market value.
- Playing in the English Premier League seems to have a large positive impact on wage but not on market value. The wage effect is probably because of the high levels of foreign investment, especially from the middle east, attracted by top English clubs.
- Playing in a certain position has similar impacts on wages and market value except for playing goalkeeper. Being a goalkeeper has a much larger negative impact on wage than on market value. (about 7% vs about 20%)

10. Conclusion

On the whole, it is evident that the chosen independent variables are better at explaining market value than wage. Moreover, it is also evident that some of these variables have different impacts on market value and wage. While this paper does have a few technical shortcomings, I believe it answers the research question, "What are the impacts of Overall Rating, League and Position on Market Value and Wage". Working through this exercise was very engaging and I learnt a lot from the process.

11. Sources

Long, Jacob. Plotting Interactions among Categorical Variables in Regression Models, 4 July 2019, cran.r-project.org/web/packages/interactions/vignettes/categorical.html.

Phillips, Nathaniel D. "YaRrr! The Pirate's Guide to R." *Home*, 22 Jan. 2018, bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html.

X, Mr. "Are Strikers Really Worth More Than Defenders in the Transfer Market?" *Bleacher Report*, Bleacher Report, 2 Oct. 2017, bleacherreport.com/articles/1719218-are-strikers-really-worth-more-than-defenders-in-the-transfer-market.